JASA ARTICLE

Reconsidering commonly used stimuli in speech perception experiments

Matthew B. Winn[1,a)] and Richard A. Wright[2]
[1]Department of Speech-Language-Hearing Sciences, University of Minnesota, Minneapolis, Minnesota 55455, USA
[2]Department of Linguistics, University of Washington, Seattle, Washington 98195, USA

ABSTRACT:
This paper examines some commonly used stimuli in speech perception experiments and raises questions about their use, or about the interpretations of previous results. The takeaway messages are: 1) the Hillenbrand vowels represent a particular dialect rather than a gold standard, and English vowels contain spectral dynamics that have been largely underappreciated, 2) the /ɑ/ context is very common but not clearly superior as a context for testing consonant perception, 3) /ɑ/ is particularly problematic when testing voice-onset-time perception because it introduces strong confounds in the formant transitions, 4) /dɑ/ is grossly overrepresented in neurophysiological studies and yet is insufficient as a generalized proxy for "speech perception," and 5) digit tests and matrix sentences including the coordinate response measure are systematically insensitive to important patterns in speech perception. Each of these stimulus sets and concepts is described with careful attention to their unique value and also cases where they might be misunderstood or over-interpreted. https://doi.org/10.1121/10.0013415
(Received 24 March 2022; revised 12 July 2022; accepted 18 July 2022; published online 2 September 2022)
[Editor: James F. Lynch] Pages: 1394–1403

I. INTRODUCTION

The goal of this paper is to provide critical reflection on commonly used stimuli in speech perception experiments. This critique will first try to highlight the positive value of each stimulus type, and then point out the ways in which it is tempting to over-interpret or misappropriate those stimuli in the pursuit of some research goals. Although there are innumerable types of stimuli for a wide range of research questions, we focus here on those that are commonly used and which would be appealing to specialists and non-specialists, including perception of vowels, consonants, stop-consonant voicing, fricatives, and sentences. Our main points will be applicable mostly for stimuli that are used to test perception of North American English, partly because of familiarity to the authors, but also because the power matrices that support and sustain research and researchers have historically given privilege to those who use and write about English. Research involving English has disproportionately large representation in the literature and thus disproportionately large opportunity for impact as well as reflection and criticism.

Since methods and stimuli are passed between labs and sometimes used for the sake of convenience, a mission of this paper is to provide insight or alternatives, as well as historical context for why researchers gravitate toward habitual practices. In the sections that follow, we present critical perspective on five commonly used sets of stimuli that are used as complementary tasks for a wide range of studies with other main goals (understanding auditory perception, dyslexia, hearing impairment, etc.). The stimuli therefore have wide impact far beyond the field of phonetics, despite the appearance of relating to phonetic perception specifically. In this review, we attempt to learn from the consistencies or inconsistencies that are observed across the literature with the aim of creating guidelines for high-quality future research.

II. VOWELS–THE HILLENBRAND VOWEL STIMULI

Both speech production and speech perception experiments commonly use "hVd" words (e.g., hid, heed, hood, head) for understandable reasons. This syllable structure is convenient for English since the onset /h/ does not cause tongue- or lip-based coarticulation (freeing the vowel formant onsets from consonant effects), because the /d/ offset is compatible with all vowels (some of which are not permissible without a closing consonant, such as the vowels in *had* and *head*), and because all of the words are plausibly real words in English, even if some stretch credulity (e.g., *hod* and *hayed*). Vowel stimuli in particular are often used in evaluating auditory perception of spectral (frequency) properties of sound because their spectral properties convey a great deal of information but the amplitude envelopes on their own are not uniquely contrastive.

The vowel acoustic data published by Hillenbrand *et al.* (1995) has been adopted as a *de facto* gold standard even though the paper was intended as a description of methods using a snapshot of a single dialect at a particular point in time. In a series of papers by the research group led by Hillenbrand, the citation of the 1995 paper ("The acoustics

[a)]Present address: 164 Pillsbury Dr. SE, Minneapolis, MN 55455, USA. Electronic mail: mwinn@umn.edu

of American English vowels") far surpasses the citations of the other papers in this series. As such, "The Hillenbrand paper" can be understood to refer to the former with no risk of confusion (although there are valuable relevant insights gained from the other papers as well). Hillenbrand *et al.* (1995) adopted the methods of an earlier vowel study by Peterson and Barney (1952) but with sophisticated new controls and the benefit of more than 40 years of technological improvements.

Despite the strengths and popularity of the Hillenbrand *et al.* (1995) paper and its core approach, some of the main messages of that paper have often been overlooked, including the importance of the dynamics of vowel spectra (i.e., formant movement within the vowel), and the notion that the paper was a description of a *method* of collecting acoustic data rather than a standard against which other formant measurements or audio stimuli should be compared. The discussion in the paper warned against the tempting idea that Table V in the article be considered a definitive and exhaustive description of English vowels, since the dialect of the talkers has some unique properties that are in flux, and because vowel systems undergo change in any dialect as time goes on. Furthermore, the table of steady-state formant values betrays the notion that the dynamics (illustrated in Fig. 9 of their article) carry significant information and the value can be substantiated in behavioral results (Jenkins *et al.*, 1983; Hillenbrand and Nearey, 1999) and discriminant analysis (Hillenbrand *et al.*, 1995). The authors demonstrate that the dynamic aspect of formants is more important than other properties such as vowel duration. F1 dynamics have also been observed to aid in classifying inter-talker variability, because of idiosyncrasies in jaw movements (He *et al.*, 2019). Considering the framework that perceiving the talker is an inextricable part of perceiving the speech (Tripp and Munson, 2021), we recommend maintaining these details when possible.

The importance of formant dynamics was observed decades earlier by Fairbanks and Grubb (1961) who noted that recognition of so-called "steady-state" portions of vowels was much poorer than expected, with only roughly 75% performance even for very carefully articulated vowels (but see Friedrichs *et al.*, 2017 for better performance on extracted clips of prolonged steady-state vowels in German). Specifically, the vowels /eɪ/, /æ/, /oʊ/, /ʊ/, and /u/ are most heavily affected by flattening formant contours (Hillenbrand and Nearey, 1999). It is notable that the last three of these vowels are rather similar in spectral shape when ignoring formant dynamics. Along the same lines, flattening formant contours results in a significant decline in /ɪ/ recognition, while the vowel /i/ is identified virtually perfectly (Assmann and Katz, 2005), consistent with the relatively dynamic nature of /ɪ/ and static nature of /i/. Despite these observations, the steady-state formant values are occasionally interpreted as if they were stationary idealized peaks, including a study conspicuously co-authored by both of the authors of the current paper (DiNino *et al.*, 2016). That is, the lure of expedient presentation of vowel data and the lure of simple classifications is tempting even to those who are critical enough to write about the limitations.

A. Solutions for over-reliance on the Hillenbrand vowel stimuli

We encourage awareness of the actual strengths and limitations of the Hillenbrand *et al.* (1995) paper on vowel acoustics, and whether the strengths align with the research goals of a particular study. If the goal is to seek advice on how to set up and execute high-quality vowel measurements, then the Hillenbrand *et al.* (1995) paper is incredibly valuable, supplemented by more-recent contributions in technology, such as formant tracking in software, such as Python and Praat (see Barreda, 2021). Conversely, direct comparison of new measurements against the Hillenbrand *et al.* (1995) formant table is not advisable unless the goal is literally to compare vowel production of talkers to a sample of talkers from a specific dialect in Southwest Michigan in the mid- 1990s. If the goal is more about *perception* of vowels, recognize that the listener's familiarity with the talker's dialect plays a role in perception (e.g., Wright and Souza, 2012). The dialect of the vowels measured by Hillenbrand *et al.* (1995) is a distinct one, with some notable features including raising of the low-front vowel in "had" (which to some ears might sound more like "hee-ad"). Incidentally, vowel dynamics are a key feature in defining the differences between dialects (Fox and Jacewicz 2009), reinforcing main messaging by Hillenbrand *et al.* (1995). The convenience of using pre-existing recordings from the 1995 study is likely outweighed by the potential improvements in stimulus quality if experimenters make new recordings of the hVd words that cover a wider range of dialects that match the dialects used by the intended study participants, and if the new recordings were made using a higher sampling rate to enable a wider range of questions pertaining to sound localization, perceiving speech in noise, and hearing impairment (Trine and Monson, 2020; Flaherty *et al.*, 2021).

Finally, if the goal is to gain understanding of the acoustics of vowels, there is no substitute for direct experience with making measurements by hand [as described directly in the Hillenbrand *et al.* (1995) paper], where one discovers that formants are not static across time, and are not sharp peaks at a single frequency but rather bands of energy that span multiple harmonics, whose spacing might make it challenging to determine the exact location of the peak (Chen *et al.*, 2019). Seasoned experimentalists will view Table V in the paper by Hillenbrand *et al.* and recognize it as a shorthand proxy for vowel measurements that are much more complicated in their full form (Shadle *et al.*, 2016). For example, Fig. 1 shows the vowel data from the women in the Hillenbrand *et al.* (1995) study, including the formant dynamics and duration characteristics that give a richer illustration of the vowel space.

Appreciation of vowel dynamics and variability is not new (cf. Fairbanks and Grubb, 1961), but is in need of continual revival as simpler interpretations unfortunately push the original value of the Hillenbrand *et al.* (1995) paper out

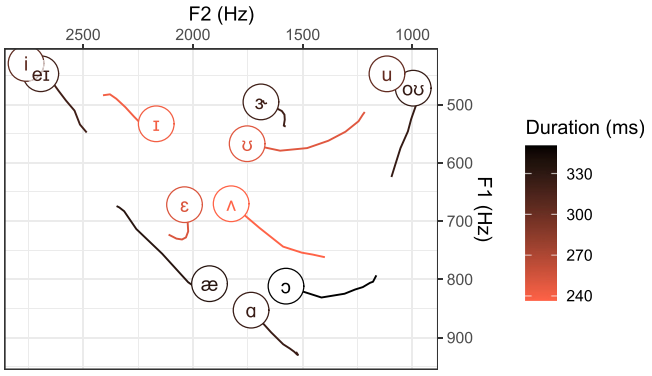

FIG. 1. (Color online) Vowel space of women measured by Hillenbrand *et al.* (1995), showing extensive dynamics of formant frequencies as well as systematic differences in duration among vowels.

of the spotlight. Appreciation of the scope and value of the original work will likely lead to better measurement of perception, better generalization of results to wider populations, and more accurate description of healthy and impaired hearing as well as the variety of vowel productions across languages and dialects.

III. CONSONANTS WITH /ɑ/ AS A CONTEXT (ta, na, ra, ka, ETC.)

The use of the /ɑ/ context is pervasive, to the point of being a "default" syllable environment for testing perception of consonants in stimuli like ba, ta, ga, la, pa, and so on. For decades, this type of stimulus has been common in studies of consonant perception in various conditions such as background noise (Phatak *et al.*, 2008), hearing loss (Wang *et al.*, 1978), aging (Gordon-Salant, 1987), high- and low-pass filtering (Miller and Nicely, 1955), cochlear implants (Fu *et al.*, 1998; Devries *et al.*, 2016), spectral shifting (Başkent and Shannon, 2003), reverberation (Reinhart *et al.*, 2016), vocoding (Shannon *et al.*, 1995), and phonetic categorization (Winn and Litovsky, 2015). We reviewed 50 papers since 1955 that tested consonant recognition in nonsense syllables, finding that 36 (72%) used only the /ɑ/ consonant environment; in the rare cases where other vowels are used, experimenters typically used a combination of /i/, /a/, and /u/ (see supplementary material for information about this literature scan).[1]

There are many attractions of "Ca" (Consonant-/ɑ/) or aCa stimuli. First of all, the stimuli are already in common use after being shared across many laboratories, so there are accessible recordings and much available data. In English, putting a consonant before /ɑ/ usually does not make a real word, which can be very helpful if an experimenter wishes to avoid complicated effects of lexical activation. Most languages have the vowel /ɑ/ or something close to it (e.g., /a/), meaning the same stimuli could in theory be comparable across languages. Other vowels can sometimes be problematic, if there are phonotactic restrictions (e.g., lax vowels cannot end a syllable in English) or neutralization patterns (e.g., /s/ and /ʃ/ merge to become the same sound when before /i/ in many languages such as Thai and Korean, but maintain contrast before other vowels).

Despite the attraction of using /ɑ/ as a standard context, there are some downsides. The vowel /ɑ/ is far from being a "neutral" vowel, with both the articulation and acoustics having extreme patterns relative to other vowels. The tongue movements for /ɑ/ are specific and can be exaggerated, resulting in the most extreme open-jaw position. Consequently, the first formant is higher than for most other vowels and undergoes a more dramatic transition from consonant constriction to vowel midpoint–which perhaps explains the relatively higher success rate for recognizing consonants with /ɑ/ compared to other vowels (Dubno and Levitt, 1981). The first two formant frequencies for /ɑ/ are very close together, unlike most other vowels, leading to an abnormally high intensity peak in the spectrum where the formants might merge into a single prominence. Normally, /ɑ/ is the loudest vowel, which makes it both atypical and uniquely unfortunate in the case of testing listeners who have hearing impairment, since preserving a target intensity level in stimuli would result in the attenuation of the consonant energy (the difficult part) to compensate for the high intensity of /ɑ/ (the easy repetitive part).

For those focusing on spoken English, /ɑ/ is not a particularly common sound. The Carnegie Mellon University (CMU) pronouncing dictionary was cross-referenced with the SUBTLEX database (Brysbaert and New, 2009) to analyze 48 352 words that were common to both databases, containing 121 435 vowels. Among 106 646 vowels following a consonant, /ɑ/ was found to be only the 7th or 8th most common vowel (Fig. 2; uncertain arising from the lack of distinction between /ʌ/ and schwa in the CMU guide). The vowel /ɑ/ appears only once in the top 25 most prevalent consonant-vowel sequences, only 3 times in the top 50, only 5 times in the top 100, and only 15 times in the top 200.

A. Solutions for the peculiarity of consonant + /ɑ/

Vowel environments other than /ɑ/ can and have been used successfully in some influential papers (cf. Wang and Bilger, 1973; Bilger and Wang, 1976; Dubno *et al.*, 1982) and continue to add richness to recent literature (Miller *et al.*, 2017; Rødvik *et al.*, 2019). In such studies, vowel environments typically include /ɑ/, /i/, and /u/, although there are no reasons to specifically constrain to those vowels at the corners of the vowel space. Another solution is to abandon the uniformity of consonant+vowel syllable structure and instead opt for highly controlled sets of word choices that are designed to target specific phonemes (cf. the Iowa Test of Consonant Perception, Geller *et al.*, 2021).

IV. /dɑ/ IN NEUROPHYSIOLOGICAL STUDIES

There is an abnormally large number of studies whose conclusions rest on neural responses to a single syllable /dɑ/ that is band limited and generated by a synthesizer. The attraction is clear: obtaining a quick, easy, and reliable physiological measure that can stand as a proxy for

https://doi.org/10.1121/10.0013415

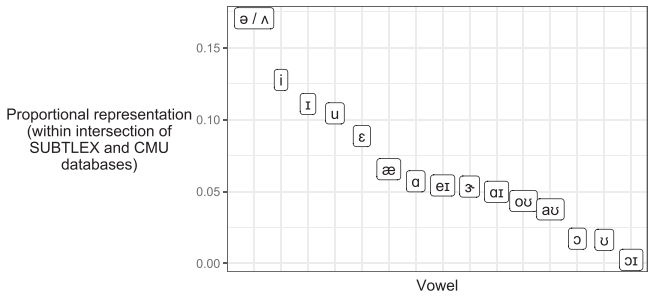

FIG. 2. The prevalence of different vowels per million words of spoken American English as reported by Brysbaert and New (2009). These data represent only the words that were also available in the Carnegie Mellon Pronouncing Dictionary. The /ʌ/ vowel is collapsed with /ə/ in this dataset.

time-intensive behavioral testing could transform clinical practice. Curiously, a sizable number of these studies make claims about issues that generate wide popular interest that expands beyond the scope of typical phonetic perception studies. For example, the /dɑ/ stimulus has been used in studies that claim that speech encoding is affected by musicianship (Musacchia *et al.*, 2008), bilingualism and executive function (Krizman *et al.*, 2012), aging (Anderson and Kraus, 2011), reading ability (Hornickel *et al.*, 2012), the impact of high school music education (Tierney *et al.*, 2013), the impact of maternal education (Skoe *et al.*, 2013), and developmental dyslexia (Chandrasekaran *et al.*, 2009). Response to /dɑ/ has also been used to claim that the enhancement from bilingual experience is robust to differences in maternal education level (Krizman *et al.*, 2016), and that the impact of musicianship is robust to differences in aging (Parbery-Clark *et al.*, 2012; Bidelman and Alain, 2015). It has been used to further claim that the impact of musicianship on /dɑ/ encoding can reflect the influence of a critical period of development (Skoe *et al.*, 2013). The /dɑ/ stimulus has also been used in a study that sought to discover hemispheric asymmetries in processing (Abrams *et al.*, 2006). Neural responses to /dɑ/ have been claimed to have a relationship with perception of sentences in noise (Anderson *et al.*, 2011; Song *et al.*, 2011), suggesting that responses to this simple stimulus can be interpreted to generalize to perception of longer utterances. That a single speech syllable could be an index for all of these issues that span far beyond peripheral auditory encoding is bewildering.

Despite the apparently remarkable success of using a single synthetic syllable to track a wide range of auditory, developmental, and social issues, there are two fundamental critiques of this constellation of findings, which are rooted both in the data and the philosophy of testing. These points stand regardless of which single syllable (e.g., /bi/, /ku/, etc.) would have been used–they are not specific to /dɑ/. First, the method of analysis is inconsistent and unclear based on a review of the literature. Differences between comparison groups in these studies are sometimes smaller than the difference between control groups across studies. For example, the difference of 0.02 μV between younger and older groups tested by Anderson *et al.* (2012) (see Fig. 5 in the article) is an entire order of magnitude smaller than the range of scores for younger listeners tested by Musacchia *et al.* (2008) (Fig. 3), which ranged from 0.1 to more than 0.3 μV. Skoe *et al.* (2013) [Fig. 2(B)] showed a difference of less than 0.005 μV between children with either higher or lower levels of maternal education. The corresponding difference in a follow-up study by Krizman was simply not reported (the data were expressed as correlations rather than magnitudes), but the size of the effect is conspicuous in light of the range between different control groups across studies far exceeding 0.005 μV. In some cases, there are alterations between expressing data in terms of amplitude (Krizman *et al.*, 2012), latency (Tierney *et al.*, 2013), phase difference (Kraus *et al.*, 2014), correlation (Skoe *et al.*, 2013), and alterations between presenting comparisons to coding of F0 versus other stimulus components (Anderson and Kraus, 2010; Song *et al.*, 2011). Each of these inconsistencies impedes critical meta-analysis and opens the door to type I errors (e.g., choosing the metric that most strongly reflects a difference between test groups, but overlooking other equally valid metrics that do not produce a difference). Alternatively, the variety in reported analyses might reflect the experimenter's choice to fit the approach to the research problem at hand (Skoe and Kraus, 2010), yet the uniformity of the stimulus across these many studies calls that explanation into question.

The more general critique of these studies is that the single syllable /dɑ/ reflects "speech" no more than a single key on the piano reflects "music." Incidentally, Musacchia *et al.* (2007) indeed did test neural encoding of a single syllable /dɑ/ and a single musical note G, and described these stimuli as ecologically valid representations of speech and music, respectively. Perspectives on ecological validity are complex (Beechey, 2022), and opinions vary; the current paper will not reproduce these extensive discussions. Although it is not controversial that /dɑ/ and the note G are found in speech and music, the successful perception of the F0 of these short stimuli is not sufficient for successful communication or the appreciation of music, respectively. Even at a mechanistic level, the neural encoding of the fundamental frequency is notoriously unnecessary since the auditory system can recover F0 despite a missing fundamental (Schouten *et al.*, 1962). The characterization of a single syllable as "speech" invites generalization that is entirely unwarranted, and deserving of methodological improvement.

A. Solutions for electrophysiological studies that use /dɑ/

In recent years, neuroscientific studies of speech perception have expanded to include continuous running speech, which offers tremendous advantage over studies that use a single syllable (Brodbeck and Simon, 2020), both for basic research and also clinical translation. For example, Polonenko and Maddox (2021) have used phase-realigned running speech ("peaky speech") to elicit subcortical signatures of auditory processing commonly used in audiological assessment. Brodbeck *et al.* (2018) presented continuous

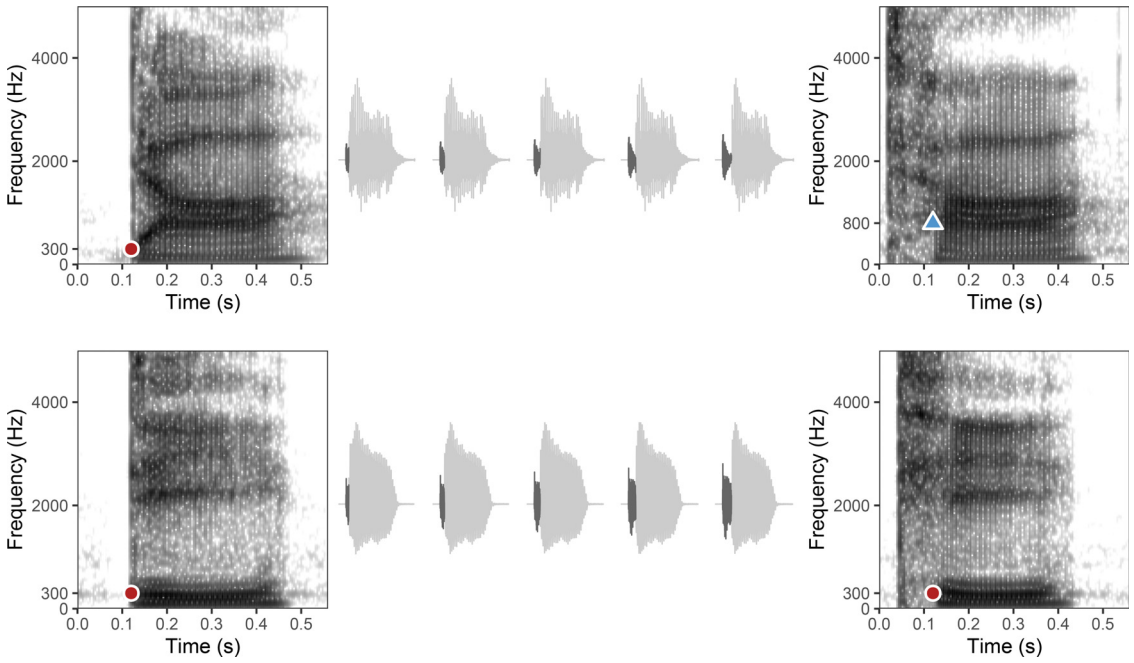

FIG. 3. (Color online) Top row: Continuum between /dɑ/ and /tɑ/ showing spectrograms at the left and right end points, with waveforms of intervening steps. Spectrograms show a confounding difference of the first formant frequency at the onset of voicing. Bottom row: Corresponding spectrograms and waveforms for a /di/ and /ti/ continuum, showing no confound with F1 frequency.

speech and recovered neural signatures of acoustic landmarks and lexical processing. Etard and Reichenbach (2019) distinguished signatures of acoustic perception versus language comprehension. Ding *et al.* (2016) found that neural activity could be broken into different time scales that correspond to words, phrases, and sentences in running speech, and these responses are modulated by attention (Teoh *et al.*, 2022). Gillis *et al.* (2021) tracked even more fine-grained information, including neural signatures of phoneme surprisal, cohort entropy, word surprisal, and word frequency in running speech. Although disagreements might exist over how to characterize concepts, such as semantic perception (e.g., whether machine models are a sufficient proxy for human lexical activation) and the units of perception (e.g., does it make sense to think of perceiving a linear sequence of phonemes?), the advancement of recent methods beyond subcortical envelope tracking holds immense promise for discovering the neural representation of spoken input.

V. VOICE-ONSET-TIME STIMULI WITH THE /ɑ/ VOWEL

The /ɑ/ environment is especially problematic for testing perception of voice-onset-time (VOT). The crux of the issue is that the onset of the vowel following a voiced stop includes a rising F1 transition, and the cutback of devoicing into the vowel (which is how VOT is systematically altered) will not only change VOT, but *also* increase the frequency of F1 at the onset of voicing [see Fig. 1 in Winn (2020) and Fig. 1 of Jiang *et al.* (2006)]. This formant transition can span multiple octaves for a low vowel like /ɑ/ since it has a high F1; the cutback of VOT occurs exactly where the transition resides and therefore presents a stimulus confound, rendering the experimenter unsure if perception was driven by VOT or the formant transition. This problem has long been recognized in the literature within speech acoustics, having been examined parametrically by Liberman *et al.* (1958), Stevens and Klatt (1974), and Lisker (1975), among others. However, as the decades have worn on, the literature has become crowded with studies that tested perception of VOT continua using the /ɑ/ context. Curiously, many of these studies reside outside the typical speech acoustics literature and extend into the neurosciences or auditory sciences. A review of 100 studies that tested VOT perception yielded 82 that used /ɑ/ or another low vowel (æ or ɑɪ) as the vowel context, with stimuli like ba-pa, da-ta, and ga-ka. A small number of these studies also used other vowels as well (see supplementary material for information about this literature scan).

Although there is nothing inherently wrong with exploring perception of VOT+/ɑ/, the issue to reconsider is that many studies frame these stimuli as explicitly testing perception of VOT, when there is a significant undesirable confound in the stimuli. Furthermore, some studies use VOT+/ɑ/ toward the goal of explicitly examining auditory *temporal* processing, sometimes in pursuit of understanding reading difficulties in children (Breier *et al.*, 2001), temporal processing in people who use cochlear implants (Caldwell and Nittrouer, 2013), children who have auditory neuropathy spectrum disorder (McFayden *et al.*, 2020), and in search of auditory-perceptual signatures of schizophrenia (Haigh *et al.*, 2019). These examples suggest that the inattention to the VOT-formant (i.e., temporal-spectral) stimulus confound is not merely quarantined within scholarly domains, but potentially has far-reaching consequences for how we understand, diagnose, and manage vulnerable populations. We offer these

https://doi.org/10.1121/10.0013415

examples not to chastise or discourage further exploration, but to encourage more refined exploration of this perceptual phenomenon *especially* when it is used in sensitive situations that involve medical diagnoses.

Physiological responses thought to reflect VOT could also reflect encoding of F1; this possibility was identified and examined by Sinex and McDonald (1989) and Sinex *et al.* (1991) in a series of papers, ultimately resulting in the conclusion "Neural responses to the onset of voicing are unrelated to other measures of temporal resolution" [the title of a paper by Sinex and Chen (2000)], perhaps because VOT is not a stimulus that taps purely into temporal processing when it is paired with the /ɑ/ vowel. The precise nature of temporal versus spectral coding is sufficiently important in cortical brain mapping that Fox *et al.* (2020) devoted a study to the concept of converting a temporal code in speech to a cortical spatial map. However, Fox *et al.* used VOT+/ɑ/ stimuli as the "temporal" speech contrast, despite the considerable spectral confound, ironically criticizing *other* studies for including non-temporal stimulus attributes in similar pursuits. Perhaps the mapping of VOT to a spatial code in auditory cortex by Fox *et al.* (2020) might reflect cortical encoding of F1 frequency, consistent with the well-established idea that the cortex encodes frequency in a spatial map (Humphries *et al.*, 2010). Thanks to Fox *et al.* openly providing their study materials, it can be observed that their continuum end points vary in onset F1 considerably, with a frequency of roughly 530 Hz at the /b/ end and roughly 725 Hz at the /p/ end, corresponding to 2 mm (about 6%) of cochlear space.

The F1 cue for stop-consonant voicing is not merely an oddity of laboratory stimuli; it is ecologically useful, as normal amounts of background noise might mask the relatively weaker aspiration. Accordingly, F1 is observed to be the *dominant* perceptual cue in such listening conditions (Jiang *et al.*, 2006). Therefore, the confound between VOT and F1 cannot be dismissed simply because F1 is relatively under-represented in the literature and because of cross-disciplinary inertia in the practice of using VOT+/ɑ/ as a stimulus paradigm.

A. Solutions for better testing of VOT perception

The generic solution is for each experimenter to critically ask of their own experiment, "How do I know that the listener is making judgments based on the VOT?" For those who are primarily interested in perception of the temporal aspects of VOT specifically, using a high vowel such as /i/ or /u/ as the context is advisable, since the F1 is low enough that there is ostensibly no perceptible spectral cue that would covary with VOT. In many varieties of English, the /u/ vowel would still be somewhat tainted by F2 transition (i.e., /u/-fronting), but would still be less problematic than /ɑ/. For those who are interested in perception of "voicing" without any regard to which acoustic cue is driving perception, then there is likely no risk for any particular stimulus, except the temptation by readers to misinterpret the results as pertaining to auditory temporal processing or VOT perception specifically. It is possible to hold the formants constant as VOT changes, by lengthening a pre-appended aspiration segment (or, when using the method described by Winn (2020), having a 0% ratio of VOT to vowel cutback). However, this method results in a stimulus that is implausible for a vocal tract to produce, and might result in data driven by the listener's management of perceptual uncertainty rather than their perception of VOT. Using the /i/ environment accomplishes the basic goal of minimizing the F1 transition while maintaining a plausibly natural signal (especially for /d/-/t/ continua, where the F2 is similarly stationary). Using multiple VOT continua (cf. McMurray *et al.*, 2008; Toscano and McMurray, 2012) is also an attractive option because it would allow the experimenter to determine whether observed effects are restricted to specific vowel environments.

VI. MATRIX SENTENCES AND DIGIT TESTS

Matrix sentences are sentence-length stimuli of consistent syntactic structure assembled by concatenating single words together in sequence. Common forms include the Oldenburg sentence test (Wagener *et al.*, 1999) and English translations with structure such as name-verb-number-adjective noun (e.g., "Bob found two red hats") or the Coordinate Response Measure (CRM) (Bolia *et al.*, 2000) stimuli (e.g., "Ready Charlie go to blue four now"). The listener usually indicates their response among options presented on the screen in a matrix. Matrix sentences are used increasingly often; PubMed reports that the original CRM study by Bolia *et al.* (2000) received 15 citations in its first 10 years of publication, but 82 citations in the past 5 years.

There are understandable attractions to testing with matrix sentences. First, they can be automatically scored by a computer, which reduces experimenter testing time as well as any potential subjectivity of experimenter bias in interpreting an unclear spoken response. It also allows a large number of potential stimuli, as the possible permutations of words in the matrix are far more numerous than most existing speech stimulus sets. It also ensures that all of the stimuli are equal in syntactic and semantic complexity. In situations where researchers need to control timing or simultaneity of specific words, matrix texts can be the only feasible option (even if such perfectly timed events would be exceedingly unlikely in everyday listening). Matrix tests have been developed in multiple languages (Kollmeier *et al.*, 2015). Digit tests (digit triplet test, digits-in-noise test; cf. Smits *et al.*, 2013) share some of the same attractions as matrix sentences (closed set, apparently limitless stimuli, automatic scoring), and are also subject to the same critiques.

The primary shortcoming of matrix sentence tests is that they are *especially* insensitive to various types of errors in speech perception, which might counteract and overcome the benefits stated in the previous two paragraphs. The listener will choose the option that most closely matches what they heard, which can cover up misperceptions. Suppose the listener thought they heard "see" when the stimulus was "three." If the

options are "two," "three," and "four," then "three" will be chosen and the listener is wrongfully regarded as having heard the stimulus correctly. This is especially problematic among the digits 1–9 in English, which all have unique vowels[2] and therefore can be recognized with 100% accuracy *even if 100% of the consonants are misperceived.* The data therefore loses useful information about the difference between the stimulus and what the listener thought they heard, which is the primary value of intelligibility scoring. Mertes (2021) found that very large differences in CRM performance (e.g., roughly 22% change) were necessary to establish statistical differences across conditions, further underscoring the test's limited sensitivity. In a detailed evaluation of the CRM, Brungart (2001) pointed out that the test is especially useful for "extremely difficult listening environments" to discover the amount of distortion that would "render a communications channel inoperative." Thus, the CRM is an ideal corpus to test the absolute limits of when speech perception is utterly destroyed, but it might be less ideal for typical situations where people regularly communicate.

The use of matrix stimuli and digits prevents a fair comparison of results across studies. As opposed to open-set stimuli where 50% recognition is observed with signal-to-noise ratios (SNRs) around −9 or −5 dB for modulated and unmodulated noises, respectively (Festen and Plomp, 1990), obtaining 50% performance in the Coordinate Response Measure requires an SNR of −15 or −9 dB, respectively, which is an arguably unrealistic signal-to-noise ratio (SNR) for any typical conversational situation. In a study by Brungart *et al.* (2001), there were conditions where even at −12 dB SNR, CRM performance was up around 60% and 80% when the masker was modulated or produced by another talker. For stimuli processed with a four-channel vocoder (a severe degradation), Milvae *et al.* (2021) reported roughly 90% intelligibility for digits (presented to a single ear), while Goupell *et al.* (2021) reported roughly 80% intelligibility for matrix sentences and Friesen *et al.* (2001) reported performance below 45% for more natural Hearing in Noise Test (HINT) sentences, all in the same level of stimulus degradation. In the studies by Milvae *et al.* (2021) and Goupell *et al.* (2021), the choice of digits and matrix sentences was strategic and justified; they needed to control the synchrony of words presented to the left and right audio channels to examine binaural interference using stimuli with the complexity of speech. However, for studies that prioritize speech intelligibility for its own sake, matrix materials could potentially be problematic and insensitive to perceptual errors.

Further corroborating that closed-set tests disproportionately depend on pure audibility, Polspoel *et al.* (2021) found that inclusion of extended high frequencies (above 8 kHz) improved recognition scores of digit triplets by 75%, but only improved words and sentences by about 22% and 24%, respectively. Therefore, closed-set tests like digit triplets (and likely the CRM and other matrix-style tests) might be less sensitive to the things that are important for regular sentence perception, while being overly sensitive to factors that are specific to the test stimuli in the moment of testing.

In addition to the insensitivity to phonetic perceptual errors, matrix sentence tests are also distinct in that they show sensitivity to *different* kinds of perceptual influences that might not generalize to other materials. Füllgrabe and Rosen (2016) found elevated influence of working memory for matrix sentences compared to other materials, possibly because repeated exposure to the same words might allow some keen listeners to cue into particular aspects of the speech signal that they remember on subsequent trials. Although working memory is a worthwhile topic for scientific inquiry, we argue that matrix tests are *unduly* influenced by working memory, since the idiosyncratic memorable stimulus attributes reflect a peculiarity of the test itself, rather than a property of speech perception that generalizes beyond the laboratory. Conversely, matrix sentences systematically disallow semantically incoherent sentences (i.e., sentences that are not merely unusual, but which do not have a sensible meaning, and which invite the listener to question their own perception), which were shown by Winn and Teece (2021) to influence listener effort more strongly than other factors, such as phonetic similarity.

A. Solutions for the peculiarity of matrix tests and digit tests

Just as for the other examples of speech stimuli described in this paper, there is nothing inherently wrong with using closed-set matrix speech tests. Problems arise only when the stimuli are used in situations where the research question requires a different type of data or different type of task. By continually gravitating toward simpler and more expedient testing materials like matrix sentences, CRM, and digits, we risk overemphasis of the auditory abilities and factors that specifically affect those materials at the neglect of the factors that affect everyday speech communication. The solution is to be mindful of the potential losses in translation and to avoid overinterpretation. For the bold, the solution is to design new stimulus materials that are well suited to the hypotheses at hand. A compromised position was offered by Uslar *et al.* (2013), who developed a set of matrix sentences that vary in linguistic complexity.

There are two situations worth avoiding. The first is the likely futile attempt to recover meaningful information about the listener's phonetic perception. Those who are interested in learning about phonetic misperceptions ought to use a stimulus where such misperceptions are not likely to be covered up by the listener's tendency to give a response constrained by choices that gravitate toward being correct. The second situation to avoid is the overinterpretation of closed-set matrix tasks as reflecting the difficulties of everyday conversation. If the researcher constrains responses to never have any variability in terms of syntactic structure or prosody, or to never have any potential semantic ambiguity (both of which would be constraints imposed by the design of matrix / CRM / digit-triplet tests), then the difficulties that arise because of those factors will be systematically ignored by the task. If the research question is "at what noise level can the listener no longer even

use learned patterns of speech sounds to identify at least part of a correct word?" –an admittedly extreme but worthwhile question–then matrix tests appear to be appropriate. But caution should be used before accepting matrix tests as an index of speech perception in a comprehensive way.

VII. CONCLUSION

Speech perception is a term that refers to a wide range of abilities that can be examined at the microscopic scale–like phonetic cue weighting–and also the macroscopic scale–like the effort needed to comprehend continuous speech. The generic term "speech perception" might accidentally contribute to some of the problems described here and elsewhere, as it improperly connotes similarity between tasks that are qualitatively different (cf. Strand *et al.*, 2021). We must collectively guard against the temptation to take results for a very narrow slice of stimuli and characterize them as reflecting speech communication writ large. More specifically though, there is much value to be gained by appreciating the original content of seminal works, and by critiquing the pattern of results as they emerge across studies. Just as any single experiment will have variance in an outcome measure that is later discovered to be richly structured, each paper can be considered a data point in the journey to understanding speech perception. The variance of results across papers could be traced to peculiarities of each paper's method that, when viewed through a wide lens, can shed light on large-scale patterns that are even more informative than any study alone.

[1]See supplementary material at https://www.scitation.org/doi/suppl/10.1121/10.0013415 for information about the literature scan; a table of studies that have tested consonant perception and the vowel environments that were used; and a table of studies that tested perception of voice-onset-time and the vowel environments that were used.

[2]The vowels in "five" and "nine" are arguably phonologically equal, but they are phonetically and acoustically contrastive because of the nasalization in "nine."

Abrams, D. A., Nicol, T., Zecker, S. G., and Kraus, N. (**2006**). "Auditory brainstem timing predicts cerebral asymmetry for speech," J. Neurosci. **26**(43), 11131–11137.
Anderson, S., and Kraus, N. (**2010**). "Objective neural indices of speech-in-noise perception," Trends Amplif. **14**(2), 73–83.
Anderson, S., and Kraus, N. (**2011**). "Neural encoding of speech and music: Implications for hearing speech in noise," Semin. Hear. **32**(2), 129–141.
Anderson, S., Parbery-Clark, A., White-Schwoch, T., and Kraus, N. (**2012**). "Aging affects neural precision of speech encoding," J. Neurosci. **32**(41), 14156–14164.
Anderson, S., Parbery-Clark, A., Yi, H.-G., and Kraus, N. (**2011**). "A neural basis of speech-in-noise perception in older adults," Ear Hear. **32**(6), 750–757.
Assmann, P. F., and Katz, W. F. (**2005**). "Synthesis fidelity and time-varying spectral change in vowels," J. Acoust. Soc. Am. **117**(2), 886–895.
Barreda, S. (**2021**). "Fast Track: Fast, (nearly) automatic formant-tracking using Praat," Linguist. Vanguard **7**(1), 20200051.
Başkent, D., and Shannon, R. V. (**2003**). "Speech recognition under conditions of frequency-place compression and expansion," J. Acoust. Soc. Am. **113**(4), 2064–2076.
Beechey, T. (**2022**). "Ecological validity, external validity, and mundane realism in hearing science," Ear Hear. (published online).
Bidelman, G. M., and Alain, C. (**2015**). "Musical training orchestrates coordinated neuroplasticity in auditory brainstem and cortex to counteract age-related declines in categorical vowel perception," J. Neurosci. **35**(3), 1240–1249.
Bilger, R. C., and Wang, M. D. (**1976**). "Consonant confusions in patients with sensorineural hearing loss," J. Speech Hear. Res. **19**(4), 718–748.
Bolia, R. S., Nelson, W. T., Ericson, M. A., and Simpson, B. D. (**2000**). "A speech corpus for multitalker communications research," J. Acoust. Soc. Am. **107**(2), 1065–1066.
Breier, J. I., Gray, L., Fletcher, J. M., Diehl, R. L., Klaas, P., Foorman, B. R., and Molis, M. R. (**2001**). "Perception of voice and tone onset time continua in children with dyslexia with and without attention deficit/hyperactivity disorder," J. Exp. Child Psychol. **80**(3), 245–270.
Brodbeck, C., Presacco, A., Anderson, S., and Simon, J. Z. (**2018**). "Over-representation of speech in older adults originates from early response in higher order auditory cortex," Acta Acust. united Ac **104**(5), 774–777.
Brodbeck, C., and Simon, J. Z. (**2020**). "Continuous speech processing," Curr. Opin. Physiol. **18**, 25–31.
Brungart, D. S. (**2001**). "Evaluation of speech intelligibility with the coordinate response measure," J. Acoust. Soc. Am. **109**(5 Pt. 1), 2276–2279.
Brungart, D. S., Simpson, B. D., Ericson, M. A., and Scott, K. R. (**2001**). "Informational and energetic masking effects in the perception of multiple simultaneous talkers," J. Acoust. Soc. Am. **110**(5 Pt. 1), 2527–2538.
Brysbaert, M., and New, B. (**2009**). "Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English," Behav. Res. Meth. **41**(4), 977–990.
Caldwell, A., and Nittrouer, S. (**2013**). "Speech perception in noise by children with cochlear implants," J. Speech. Lang. Hear. Res. **56**(1), 13–30.
Chandrasekaran, B., Hornickel, J., Skoe, E., Nicol, T., and Kraus, N. (**2009**). "Context-dependent encoding in the human auditory brainstem relates to hearing speech in noise: Implications for developmental dyslexia," Neuron **64**(3), 311–319.
Chen, W.-R., Whalen, D. H., and Shadle, C. H. (**2019**). "*F*0-induced formant measurement errors result in biased variabilities," J. Acoust. Soc. Am. **145**(5), EL360–EL366.
DeVries, L., Scheperle, R., and Bierer, J. A. (**2016**). "Assessing the electrode-neuron interface with the electrically evoked compound action potential, electrode position, and behavioral thresholds," J. Assoc. Res. Otolaryngol. **17**(3), 237–252.
Ding, N., Melloni, L., Zhang, H., Tian, X., and Poeppel, D. (**2016**). "Cortical tracking of hierarchical linguistic structures in connected speech," Nat. Neurosci. **19**(1), 158–164.
DiNino, M., Wright, R. A., Winn, M. B., and Bierer, J. A. (**2016**). "Vowel and consonant confusions from spectrally manipulated stimuli designed to simulate poor cochlear implant electrode-neuron interfaces," J. Acoust. Soc. Am. **140**(6), 4404–4418.
Dubno, J. R., Dirks, D. D., and Langhofer, L. R. (**1982**). "Evaluation of hearing-impaired listeners using a Nonsense-syllable Test II. Syllable recognition and consonant confusion patterns," J. Speech. Lang. Hear. Res. **25**(1), 141–148.
Dubno, J. R., and Levitt, H. (**1981**). "Predicting consonant confusions from acoustic analysis," J. Acoust. Soc. Am. **69**(1), 249–261.
Etard, O., and Reichenbach, T. (**2019**). "Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise," J. Neurosci. **39**(29), 5750–5759.
Fairbanks, G., and Grubb, P. (**1961**). "A psychophysical investigation of vowel formants," J. Speech Hear. Res. **4**, 203–219.
Festen, J. M., and Plomp, R. (**1990**). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. **88**(4), 1725–1736.
Flaherty, M., Libert, K., and Monson, B. B. (**2021**). "Extended high-frequency hearing and head orientation cues benefit children during speech-in-speech recognition," Hearing Res. **406**, 108230.
Fox, N. P., Leonard, M., Sjerps, M. J., and Chang, E. F. (**2020**). "Transformation of a temporal speech cue to a spatial neural code in human auditory cortex," eLife **9**, e53051.
Fox, R. A., and Jacewicz, E. (**2009**). "Cross-dialectal variation in formant dynamics of American English vowels," J. Acoust. Soc. Am. **126**(5), 2603–2618.
Friedrichs, D., Maurer, D., Rosen, S., and Dellwo, V. (**2017**). "Vowel recognition at fundamental frequencies up to 1 kHz reveals point vowels as acoustic landmarks," J. Acoust. Soc. Am. **142**(2), 1025–1033.

Friesen, L. M., Shannon, R. V., Başkent, D., and Wang, X. (**2001**). "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants," J. Acoust. Soc. Am. **110**(2), 1150–1163.
Fu, Q.-J., Shannon, R. V., and Wang, X. (**1998**). "Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing," J. Acoust. Soc. Am. **104**(6), 3586–3596.
Füllgrabe, C., and Rosen, S. (**2016**). "On the (un)importance of working memory in speech-in-noise processing for listeners with normal hearing thresholds," Front. Psychol. **07**, 1268.
Geller, J., Holmes, A., Schwalje, A., Berger, J. I., Gander, P. E., Choi, I., and McMurray, B. (**2021**). "Validation of the Iowa test of consonant perception," J. Acoust. Soc. Am. **150**(3), 2131–2153.
Gillis, M., Vanthornhout, J., Simon, J. Z., Francart, T., and Brodbeck, C. (**2021**). "Neural markers of speech comprehension: Measuring EEG tracking of linguistic speech representations, controlling the speech acoustics," J. Neurosci. **41**(50), 10316–10329.
Gordon-Salant, S. (**1987**). "Consonant recognition and confusion patterns among elderly hearing-impaired subjects," Ear Hear. **8**(5), 270–276.
Goupell, M. J., Eisenberg, D., and Milvae, K. D. (**2021**). "Dichotic listening performance with cochlear-implant simulations of ear asymmetry is consistent with difficulty ignoring clearer speech," Atten. Percept. Psychophys. **83**(5), 2083–2101.
Haigh, S. M., Laher, R. M., Murphy, T. K., Coffman, B. A., Ward, K. L., Leiter-McBeth, J. R., Holt, L. L., and Salisbury, D. F. (**2019**). "Normal categorical perception to syllable-like stimuli in long term and in first episode schizophrenia," Schizophrenia Res. **208**, 124–132.
He, L., Zhang, Y., and Dellwo, V. (**2019**). "Between-speaker variability and temporal organization of the first formant," J. Acoust. Soc. Am. **145**(3), EL209–EL214.
Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**(5 Pt. 1), 3099–3111.
Hillenbrand, J. M., and Nearey, T. M. (**1999**). "Identification of resynthesized /hVd/ utterances: Effects of formant contour," J. Acoust. Soc. Am. **105**(6), 3509–3523.
Hornickel, J., Anderson, S., Skoe, E., Yi, H.-G., and Kraus, N. (**2012**). "Subcortical representation of speech fine structure relates to reading ability," Neuroreport **23**(1), 6–9.
Humphries, C., Liebenthal, E., and Binder, J. R. (**2010**). "Tonotopic organization of human auditory cortex," NeuroImage **50**(3), 1202–1211.
Jenkins, J. J., Strange, W., and Edman, T. R. (**1983**). "Identification of vowels in 'vowelless' syllables," Percept. Psychophys. **34**(5), 441–450.
Jiang, J., Chen, M., and Alwan, A. (**2006**). "On the perception of voicing in syllable-initial plosives in noise," J. Acoust. Soc. Am. **119**(2), 1092–1105.
Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M. A., Uslar, V., Brand, T., and Wagener, K. C. (**2015**). "The multilingual matrix test: Principles, applications, and comparison across languages: A review," Int. J. Audiol. **54**(Suppl 2), 3–16.
Kraus, N., Slater, J., Thompson, E. C., Hornickel, J., Strait, D. L., Nicol, T., and White-Schwoch, T. (**2014**). "Music enrichment programs improve the neural encoding of speech in at-risk children," J. Neurosci. **34**(36), 11913–11918.
Krizman, J., Marian, V., Shook, A., Skoe, E., and Kraus, N. (**2012**). "Subcortical encoding of sound is enhanced in bilinguals and relates to executive function advantages," Proc. Natl. Acad. Sci. U.S.A. **109**(20), 7877–7881.
Krizman, J., Skoe, E., and Kraus, N. (**2016**). "Bilingual enhancements have no socioeconomic boundaries," Dev. Sci. **19**(6), 881–891.
Liberman, A. M., Delattre, P. C., and Cooper, F. S. (**1958**). "Some cues for the distinction between voiced and voiceless stops in initial position," Lang. Speech **1**(3), 153–167.
Lisker, L. (**1975**). "Letter: Is it VOT or a first-formant transition detector?," J. Acoust. Soc. Am. **57**(6), 1547–1551.
McFayden, T. C., Baskin, P., Stephens, J., and He, S. (**2020**). "Cortical auditory event-related potentials and categorical perception of voice onset time in children with an auditory neuropathy spectrum disorder," Front. Hum. Neurosci. **14**, 184.
McMurray, B., Aslin, R. N., Tanenhaus, M. K., Spivey, M. J., and Subik, D. (**2008**). "Gradient sensitivity to within-category variation in words and syllables," J. Exp. Psychol. Hum. Percept. Perform. **34**(6), 1609–1631.
Mertes, I. (**2021**). "Reliability and critical differences for an implementation of the coordinate response measure in speech-shaped noise," JASA Express Lett. **1**, 015202.
Miller, G. A., and Nicely, P. E. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**(2), 338–352.
Miller, J. D., Watson, C. S., Leek, M. R., Dubno, J. R., Wark, D. J., Souza, P. E., Gordon-Salant, S., and Ahlstrom, J. B. (**2017**). "Syllable-constituent perception by hearing-aid users: Common factors in quiet and noise," J. Acoust. Soc. Am. **141**(4), 2933–2946.
Milvae, K. D., Kuchinsky, S. E., Stakhovskaya, O. A., and Goupell, M. J. (**2021**). "Dichotic listening performance and effort as a function of spectral resolution and interaural symmetry," J. Acoust. Soc. Am. **150**(2), 920–935.
Musacchia, G., Sams, M., Skoe, E., and Kraus, N. (**2007**). "Musicians have enhanced subcortical auditory and audiovisual processing of speech and music," Proc. Nat. Acad. Sci. U.S.A. **104**(40), 15894–15898.
Musacchia, G., Strait, D., and Kraus, N. (**2008**). "Relationships between behavior, brainstem and cortical encoding of seen and heard speech in musicians and non-musicians," Hear. Res. **241**(1—2), 34–42.
Parbery-Clark, A., Anderson, S., Hittner, E., and Kraus, N. (**2012**). "Musical experience strengthens the neural representation of sounds important for communication in middle-aged adults," Front. Ag. Neurosci. **4**, 30.
Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of vowels," J. Acoust. Soc. Am. **24**, 175–184.
Phatak, S. A., Lovitt, A., and Allen, J. B. (**2008**). "Consonant confusions in white noise," J. Acoust. Soc. Am. **124**(2), 1220–1233.
Polonenko, M. J., and Maddox, R. K. (**2021**). "Exposing distinct subcortical components of the auditory brainstem response evoked by continuous naturalistic speech," eLife **10**, e62329.
Polspoel, S., Kramer, S. E., van Dijk, B., and Smits, C. (**2021**). "The importance of extended high-frequency speech information in the recognition of digits, words, and sentences in quiet and noise," Ear Hear. **43**, 913–920.
Reinhart, P. N., Souza, P. E., Srinivasan, N. K., and Gallun, F. J. (**2016**). "Effects of reverberation and compression on consonant identification in individuals with hearing impairment," Ear Hear. **37**(2), 144–152.
Rødvik, A. K., Tvete, O., Torkildsen, J., Wie, O. B., Skaug, I., and Silvola, J. T. (**2019**). "Consonant and vowel confusions in well-performing children and adolescents with cochlear implants, measured by a nonsense syllable repetition test," Front. Psychol. **10**, 1813.
Schouten, J. F., Ritsma, R. J., and Lopes Cardozo, B. (**1962**). "Pitch of the residue," J. Acoust. Soc. Am. **34**, 1418–1424.
Shadle, C. H., Nam, H., and Whalen, D. H. (**2016**). "Comparing measurement errors for formants in synthetic and natural vowels," J. Acoust. Soc. Am. **139**(2), 713–727.
Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**(5234), 303–304.
Sinex, D. G., and Chen, G. D. (**2000**). "Neural responses to the onset of voicing are unrelated to other measures of temporal resolution," J. Acoust. Soc. Am. **107**(1), 486–495.
Sinex, D. G., and McDonald, L. P. (**1989**). "Synchronized discharge rate representation of voice-onset time in the chinchilla auditory nerve," J. Acoust. Soc. Am. **85**(5), 1995–2004.
Sinex, D. G., McDonald, L. P., and Mott, J. B. (**1991**). "Neural correlates of nonmonotonic temporal acuity for voice onset time," J. Acoust. Soc. Am. **90**(5), 2441–2449.
Skoe, E., and Kraus, N. (**2010**). "Auditory brain stem response to complex sounds: A tutorial," Ear Hear. **31**(3), 302–324.
Skoe, E., Krizman, J., and Kraus, N. (**2013**). "The impoverished brain: Disparities in maternal education affect the neural response to sound," J. Neurosci. **33**(44), 17221–17231.
Smits, C., Theo Goverts, S., and Festen, J. M. (**2013**). "The digits-in-noise test: Assessing auditory speech recognition abilities in noise," J. Acoust Soc. Am. **133**(3), 1693–1706.
Song, J. H., Skoe, E., Banai, K., and Kraus, N. (**2011**). "Perception of speech in noise: Neural correlates," J. Cogn. Neurosci. **23**(9), 2268–2279.

Stevens, K. N., and Klatt, D. H. (**1974**). "Role of formant transitions in the voiced-voiceless distinction for stops," J. Acoust. Soc. Am. **55**(3), 653–659.
Strand, J. F., Ray, L., Dillman-Hasso, N. H., Villanueva, J., and Brown, V. A. (**2021**). "Understanding speech amid the jingle and jangle: Recommendations for improving measurement practices in listening effort research," Aud. Percept. Cognit. **3**(4), 169–188.
Teoh, E. S., Ahmed, F., and Lalor, E. C. (**2022**). "Attention differentially affects acoustic and phonetic feature encoding in a multispeaker environment," J. Neurosci. **42**(4), 682–691.
Tierney, A., Krizman, J., Skoe, E., Johnston, K., and Kraus, N. (**2013**). "High school music classes enhance the neural processing of speech," Front. Psychol. **4**, 855.
Toscano, J. C., and McMurray, B. (**2012**). "Cue-integration and context effects in speech: Evidence against speaking-rate normalization," Atten. Percept. Psychophys. **74**(6), 1284–1301.
Trine, A., and Monson, B. B. (**2020**). "Extended high frequencies provide both spectral and temporal information to improve speech-in-speech recognition," Trends Hear. **24**, 233121652098029.
Tripp, A., and Munson, B. (**2021**). "Perceiving gender while perceiving language: Integrating psycholinguistics and gender theory," Wiley Interdiscip. Rev. Cogn. Sci. **13**, e1583.
Uslar, V. N., Carroll, R., Hanke, M., Hamann, C., Ruigendijk, E., Brand, T., and Kollmeier, B. (**2013**). "Development and evaluation of a linguistically and audiologically controlled sentence intelligibility test," J. Acoust. Soc. Am. **134**(4), 3039–3056.
Wagener, K., Kühnel, V., and Kollmeier, B. (**1999**). "Entwicklung und evaluation eines satztests für die deutsche sprache I: Design des Oldenburger satztests" ("Development and evaluation of a speech intelligibility test for German I: Design of the Oldenburg sentence test")," Z. Audiologie **38**, 4–15.
Wang, M. D., and Bilger, R. C. (**1973**). "Consonant confusions in noise: A study of perceptual features," J. Acoust. Soc. Am. **54**(5), 1248–1266.
Wang, M. D., Reed, C. M., and Bilger, R. C. (**1978**). "A comparison of the effects of filtering and sensorineural hearing loss on patients of consonant confusions," J. Speech Hear. Res. **21**(1), 5–36.
Winn, M. B. (**2020**). "Manipulation of voice onset time in speech stimuli: A tutorial and flexible Praat script," J. Acoust. Soc. Am. **147**(2), 852–866.
Winn, M. B., and Litovsky, R. Y. (**2015**). "Using speech sounds to test functional spectral resolution in listeners with cochlear implants," J. Acoust. Soc. Am. **137**(3), 1430–1442.
Winn, M. B., and Teece, K. H. (**2021**). "Listening effort is not the same as speech intelligibility score," Trends Hear. **25**, 233121652110276.
Wright, R., and Souza, P. (**2012**). "Comparing identification of standardized and regionally valid vowels," J. Speech, Lang., Hear. Res. **55**(1), 182–193.