


## Nonspeech sounds are not all equally good at being nonspeech<sup>a)</sup>

Christian E. Stilp,<sup>b)</sup>  Anya E. Shorey, and Caleb J. King

Department of Psychological and Brain Sciences, University of Louisville, Louisville, Kentucky 40292, USA

### ABSTRACT:

Perception of speech sounds has a long history of being compared to perception of nonspeech sounds, with rich and enduring debates regarding how closely they share similar underlying processes. In many instances, perception of nonspeech sounds is directly compared to that of speech sounds without a clear explanation of how related these sounds are to the speech they are selected to mirror (or not mirror). While the extreme acoustic variability of speech sounds is well documented, this variability is bounded by the common source of a human vocal tract. Nonspeech sounds do not share a common source, and as such, exhibit even greater acoustic variability than that observed for speech. This increased variability raises important questions about how well perception of a given nonspeech sound might resemble or model perception of speech sounds. Here, we offer a brief review of extremely diverse nonspeech stimuli that have been used in the efforts to better understand perception of speech sounds. The review is organized according to increasing spectrotemporal complexity: random noise, pure tones, multitone complexes, environmental sounds, music, speech excerpts that are not recognized as speech, and sinewave speech. Considerations are offered for stimulus selection in nonspeech perception experiments moving forward.

© 2022 Acoustical Society of America. <https://doi.org/10.1121/10.0014174>

(Received 11 March 2022; revised 26 August 2022; accepted 30 August 2022; published online 22 September 2022)

[Editor: Benjamin V Tucker]

Pages: 1842–1849

### I. SPEECH PERCEPTION VERSUS NONSPEECH PERCEPTION

The human vocal tract is unrivaled in its sound-producing capabilities relative to other organisms. The speech signal is widely noted for its seemingly paradoxical nature. On the one hand, speech displays extreme acoustic variability within and across sounds as well as within and across talkers. On the other hand, the speech signal also displays considerable acoustic redundancy, with listeners displaying outstanding facility in understanding speech even in adverse listening conditions. These facts conspired to spark a spirited debate: is the perception of speech sounds a fundamentally unique accomplishment by human listeners? The initial appeal of this notion was complicated by subsequent research. While unanimous opinion may be lacking, the balance of evidence points towards the answer being “no.”

One way this “specialness” of speech has been assessed is through comparisons between the perception of speech sounds and the perception of nonspeech sounds. Speech and nonspeech stimuli that produced different patterns of perception could be taken to support the “specialness” of speech, whereas similar patterns of performance across sound classes challenged that claim. Such comparisons were famously made in investigations of categorical perception, where nonlinear relationships between perceptual performance and

signal acoustics were taken to indicate a specialized “mode” of processing for speech sounds (e.g., Liberman *et al.*, 1967; Studdert-Kennedy *et al.*, 1970) or general auditory processing of complex sounds (e.g., Locke and Kellar, 1973; Miller *et al.*, 1976). This speech/nonspeech debate generated substantial research on categorical perception itself, but at the same time, it was never limited to categorical perception. The perceptual disentangling of coarticulated phonemes was argued to be accomplished by accessing the underlying speech gestures (e.g., Mann, 1980) or via neural mechanisms emphasizing acoustic differences between sounds (e.g., Lotto and Kluender, 1998). More than a half-century since its inception, this intellectual debate continues today. Various research programs continue to compare perception of speech signals against perception of nonspeech signals, as evidenced by the results of a Google Scholar search for citations in the calendar year 2021 that include both “speech” and “nonspeech” (or when written “non-speech”).<sup>1</sup> The search returned more than 3500 results, of which more than 800 contain the phrase “speech and nonspeech” and 26 contain the more confrontational “speech versus nonspeech.”

While this debate has unquestionably been productive and educational, there are likely simplifying assumptions that merit deeper consideration. It is widely accepted that speech sounds are extremely acoustically variable, but there exists an upper bound on this variability. Acoustic properties of speech (and the variability in those properties) are restricted by the physical constraints imposed by human vocal tracts, the common source to this sound class. Nonspeech sounds do not share a common source nor a common set of physical constraints on sound production.

<sup>a)</sup>Portions of this work were presented at the “180th Meeting of the Acoustical Society of America, Acoustics in Focus.” This paper is part of a special issue on Reconsidering Classic Ideas in Speech Communication.

<sup>b)</sup>Present address: 317 Life Sciences Building Louisville, KY 40292. Electronic mail: christian.stilp@louisville.edu

As such, their acoustic diversity far exceeds that of human speech, ranging from being maximally random (white noise) to maximally sparse in frequency (a pure tone) or in time (a click) (Stilp *et al.*, 2018). This results in a sizable number of experimenter degrees of freedom for stimulus selection in any study, and consequently, how closely the perception of nonspeech sounds might resemble or inform perception of speech sounds. This matter has received varying amounts of attention over the years (e.g., see Rosen and Iverson, 2007; Bent and Pisoni, 2008; Rosen *et al.*, 2011 for discussions), but is the primary focus here.

What follows is a brief review of extremely diverse nonspeech sounds that have been used as stimuli in auditory perception experiments. Three guiding principles shape this review. First, while most of these signals have been and continue to be studied in their own rights (e.g., psychoacoustics, music perception, etc.), here we focus on examples where their perception was directly compared to and/or modeled after perception of speech sounds and related phenomena. We have deliberately refrained from listing the outcome or conclusion of each study (i.e., whether results were statistically significant or not; whether the authors argued that nonspeech perception did or did not resemble speech perception) as to avoid any potential outcome bias. Instead, the nonspeech stimulus selected to ask that question is of central interest. Second, the comprehensiveness in this review is in documenting the variety of nonspeech stimuli employed in various studies, not in citing every study that used a particular nonspeech stimulus. Therefore, each type of nonspeech stimulus listed in Sec. II lists a sampling of studies to keep the focus on the types of stimuli rather than attempting to recreate their entire histories. Third, a variety of experimental paradigms have been utilized to compare perception of speech and nonspeech; two prevalent paradigms cited below ask whether nonspeech sounds are perceived in a categorical manner similar to that of speech sounds (i.e., categorical perception) and how acoustic properties of surrounding sounds influence perception of a (target) speech or nonspeech sound (i.e., acoustic context effects). These are clearly not the only experimental paradigms that have compared perception of nonspeech sounds to perception of speech sounds, but they provide a wealth of examples to fuel the current review. While spectrotemporal complexity is the prevailing organizational principle for this review, it is not the only relevant factor driving nonspeech stimulus selection. Section III discusses the roles played by the acoustic ecology of nonspeech sounds and matters of listener experience and/or recognition for this research. Section IV then offers considerations and recommendations to improve the precision of future work on nonspeech perception before ultimately concluding in Sec. V.

## II. NONSPEECH STIMULI

The following nonspeech stimuli are organized to begin with sounds that are the furthest from speech in terms of spectrotemporal complexity, then progressively increase in

their acoustic complexity, and concurrently, their similarity to the acoustic structure of human speech. The nonspeech sounds reviewed are as follows: noise, pure tones, multitone complexes, environmental sounds, music, speech excerpts not recognized as speech, and sinewave speech (Fig. 1).<sup>2</sup>

### A. Noise

The simplest stimulus with which to begin this review is random (white) noise. Some studies of acoustic context effects have seen the preceding acoustic context stimulus (typically a sentence) replaced by wideband noise. Watkins (1991) presented a context of signal-correlated noise (Schroeder, 1968), where half of a sentence's samples were randomly selected and their polarity was reversed (i.e., multiplying the amplitude value by  $-1$ ), preserving the sentence's amplitude envelope but flattening its spectrum. These stimuli were subsequently processed by a low-pass filter (to approach the long-term average spectrum of speech) and then spectral-envelope difference filters (to capture the spectral difference between the target vowels (Watkins, 1991). Holt (2006) utilized a notched-noise context in which brief spectrotemporal notches (70 ms duration, 100 Hz bandwidth) were excised from 2.2 s of wideband noise. This produced a spectrally complementary stimulus to a sequence of pure tones used in other experiments (described in Sec. II B). To better understand categorical perception of stop consonants, Eimas (1963) examined the identification and discrimination of noise bursts varying in duration (starting at 300 ms), which listeners classified as "short" or "long." In other research, Mirman and colleagues (2004) examined categorization and discrimination of 300 ms white noise bursts with two 200 Hz-wide spectral notches and varying onset/offset ramps.

Given the distinctiveness of these studies, it is not surprising that they prioritized modeling different acoustic properties of speech. Eimas' (1963) modeling of speech acoustics was the coarsest, as noise durations were not meant to capture temporal properties of speech but merely to determine whether categorical patterns of identification and/or discrimination would be observed in nonspeech stimuli. Also, on this shorter timescale, the noise bursts of Mirman *et al.* (2004) tested acoustic properties that were analogous to amplitude rise time and spectral modulation (i.e., variation between energy and notches in the spectrum) in speech. Mirman and colleagues noted, "Although the cues that distinguish these stimuli are abstractly similar to cues that distinguish speech sounds, these stimuli were perceived as bursts of noise and not as speech" (p. 1200). On the longer timescale, correspondence to speech acoustics was even broader. Watkins' (1991) noise stimuli preserved the amplitude envelope, sentence-length duration, and long-term average spectrum of speech (except for the modifications imposed by the spectral-envelope difference filter). Holt's (2006) objective was to provide a spectral complement to a sequence of pure tones of varying frequencies. In doing so, this created a stimulus that retained a sentence-

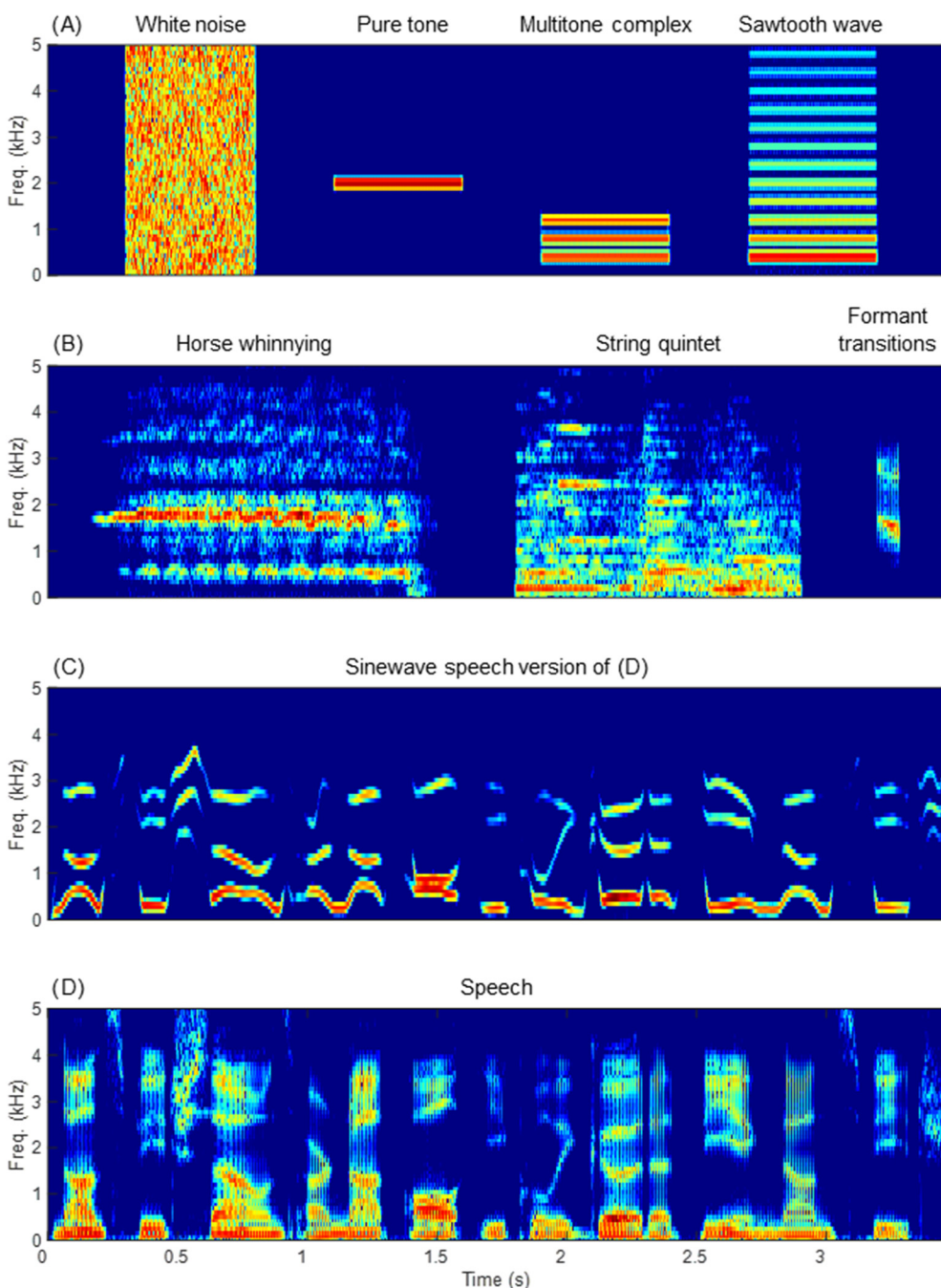


FIG. 1. (Color online) Samples of each nonspeech stimulus class in this review. (A) From left to right, white noise, a pure tone, a multitone complex, and a sawtooth wave. (B) From left to right, a horse whinnying (from the database described in Gygi and Shafiro, 2010), an excerpt of a string quintet (from Stilp *et al.*, 2010), and the excised  $F_2$  and  $F_3$  transitions from the consonantal onset of /da/ (from Stephens and Holt, 2011). (C) A sinewave speech version of the sentence in (D). (D) A recording of the first author reading the title of this manuscript.

length duration but little else of speech acoustics. Given the stark difference in spectrotemporal structure between noise and speech, it is unsurprising that noise’s ability to broadly model this structure is relatively crude.

## B. Pure tones

Pure tones have long and productive histories of use in psychoacoustics and neurophysiology, but also have been utilized frequently in comparisons with speech perception.

Cutting and Rosner (1974; Cutting, 1982) tested categorical perception of pure tones varying in their rise times, akin to the difference in rise times across affricate and fricative consonants. Diehl and Walsh (1989) modeled the perception of /ba/ and /wa/ using a pure tone that mimicked their first formant trajectories and amplitude profiles. Pisoni (1977) presented a sequence of two overlapping pure tones with asynchronous onsets to test a general mechanism underlying perception of voice onset times in syllable-initial position. Siegel and Siegel (1977) presented two pure tones in

sequence to form musical intervals, measuring the degree to which listeners perceived them categorically.

In research on acoustic context effects, [Lotto and Kluender \(1998\)](#) compared perception of target syllables (varying from /da/ to /ga/) when they followed liquid consonants (/l/ or /r/) or pure tones that modeled the trajectory or mean frequency of  $F_3$  transitions in /l/ or /r/ (see also [Kingston et al., 2014](#)). Later, Holt and colleagues presented a sequence of short-duration pure tones before listeners categorized target syllables perceptually varying from /da-/ga/ ([Holt, 2005](#)) or /ba-/wa/ ([Wade and Holt, 2005b](#); see also [Bosker, 2017](#)).

The sparse spectra of pure tones provide exquisite flexibility for modeling specific frequency characteristics of speech. These characteristics span a formant transition ([Lotto and Kluender, 1998](#); [Kingston et al., 2014](#)), formant center frequency ([Lotto and Kluender, 1998](#)), or a spectral peak in the long-term average spectrum ([Holt, 2006](#)). The simple structure of tones is also amenable to modifying their temporal properties. In the studies reviewed above, individual tone durations broadly spanned phoneme-length to syllable-length, but have been further modified to model onset asynchrony across two frequency regions ([Pisoni, 1977](#)), different rise times ([Cutting and Rosner, 1974](#); [Cutting, 1982](#)), formant transitions ([Diehl and Walsh, 1989](#)), and syllable durations and timing at different speaking rates ([Wade and Holt, 2005b](#); [Bosker, 2017](#)). This flexibility provides a marked improvement in the ability to model acoustic properties of speech beyond what is possible through the use of noise stimuli.

### C. Multitone complexes

The stimuli reviewed thus far are considered relatively simple because they occupy the extremes of acoustic structure ([Stilp et al., 2018](#)). Between these extremes, multitone complexes represent the next increment in complexity because of the introduction of (potential) harmonicity. [Healy and Repp \(1982\)](#) examined categorization of non-speech timbres (as “low” or “high”) using brief (50 ms) sounds that were generated in a speech synthesizer. Only the second formant was synthesized, producing stimuli that were a band of energy with a center frequency varying between 2156 and 2837 Hz. [Locke and Kellar \(1973\)](#) arranged three pure tones into triads to assess categorical perception of major and minor musical chords. [Pisoni et al. \(1983\)](#) conducted a study similar to [Diehl and Walsh \(1989\)](#) reviewed in Sec. IIB but used pure tones to model all three formants of /ba/ and /wa/, the end result of which retained some speechlike characteristics. [Wade and Holt \(2005a\)](#) created hybrid stimuli (a lower-frequency component based on a square wave carrier, a higher-frequency component based on either a sawtooth carrier or a noise carrier) with speech-like spectral kinematics (onset or offset trajectories plus steady-state frequencies) to investigate how listeners developed categories for novel spectrally non-invariant sounds.

Beyond adding frequency components together, harmonic complexes have also been used to model speech

perception. Like their experiments using individual pure tones discussed in Sec. IIB, [Cutting and Rosner \(1974\)](#) varied the rise times of sawtooth waves to examine whether they were perceived categorically. [Miller and colleagues \(1976\)](#) measured whether listeners would categorically perceive the asynchrony between the onsets of bandpass filtered noise and a 100 Hz square wave. [Diehl and colleagues](#) used square wave stimuli in multiple studies to model aspects of the perception of intervocalic stop consonants ([Parker et al., 1986](#); [Kluender et al., 1988](#)). Finally, [Tao et al. \(2021\)](#) measured the influences of various preceding contexts on perception of  $f_0$  contour in the final test word in Cantonese, ranging from intelligible speech to a triangle wave that followed the  $f_0$  and intensity profiles of the intelligible speech.

Multitone complexes can capture harmonicity, which is a key characteristic of speech (and other natural sounds). Multitone complexes also allow for variation in spectral density, from extremely sparse (e.g., a single formant) ([Healy and Repp, 1982](#)) to a sparse spectrum (e.g., three simultaneous pure tones in [Pisoni et al., 1983](#)) to a denser spectrum (e.g., sawtooth waves in [Cutting and Rosner, 1974](#)). Strategic positioning of harmonic complex components to align with formant center frequencies can brush up against producing a speech-like stimulus [see [Pisoni et al. \(1983\)](#) and [Healy and Repp \(1982\)](#)]. This exceptional case is reviewed separately (see Sec. IIG for sinewave speech). Finally, as was true for pure tones, multitone complexes retain sufficient flexibility to allow for various temporal modifications. On a shorter timescale, [Cutting and Rosner \(1974\)](#) varied the rise times of sawtooth waves to model consonant rise times that vary as a function of manner of articulation. On a longer timescale, [Tao et al. \(2021\)](#) presented triangle waves of sentence-length duration (all stimuli >800 ms) to convey the  $f_0$  frequency and intensity contours of an entire sentence.

### D. Environmental sounds

This point in the review marks a sizable increase in the acoustic complexity and diversity of nonspeech stimuli. Broadly stated, environmental sounds are naturally occurring sounds that are neither speech nor music (as defined in [Shafiro and Gygi, 2004](#)). This sound class contains many subcategories, such as animal sounds, vehicles, tools and machinery, nonspeech human sounds (e.g., laughing, crying, sneezing), household sounds, nature sounds, and impact sounds, to name a few (e.g., see [Gygi and Shafiro, 2010](#) for discussion). Given this diversity, spectral and temporal properties of environmental sounds are highly variable across individual sound types, much more so than noise, tones, or multitone/harmonic complexes. To date, this research has not participated in the debates surrounding categorical perception or acoustic context effects that other nonspeech sounds have. Instead, broad parallels have been forged between environmental sounds and speech. Both sound classes exhibit a variety of spectral dynamics, varying from little and slow spectral change to abrupt and rapid spectral change ([Reddy et al., 2009](#)). Recognition of both environmental sounds and speech in the presence of

background noise have been suggested to rely on similar frequencies (1200–2400 Hz) (Gygi *et al.*, 2004) and a common ability to segregate familiar sounds from complex backgrounds (Kidd *et al.*, 2007). Finally, several acoustic features that listeners use to recognize environmental sounds are also important in speech sound recognition: harmonicity, amplitude envelope shape, periodicity, and coherence of temporal changes across frequency bands (for detailed analyses see Ballas, 1993; Gygi *et al.*, 2004). In other work, Fowler and Rosenblum (1990) used the sound of slamming doors to investigate a possible nonspeech version of duplex perception of the initial consonant in consonant–vowel syllables (Rand, 1974; Liberman *et al.*, 1981) (see Sec. II F).

### E. Music

Several studies mentioned thus far examined phenomena related to music perception, but used stimuli that were acoustically simpler than music (pure tones, multitone complexes). Here, we review comparisons to speech perception that used actual musical sounds or excerpts as stimuli. Investigations of spectral context effects were extended to nonspeech signals by Stilp *et al.* (2010), who presented a nonspeech context (string quintet excerpt) before a musical instrument target (edited renditions of a French horn or a tenor saxophone). In their study listed in Sec. II C, Tao *et al.* (2021) compared the influence of a speech context on target  $f_0$  contour perception to that of a series of piano notes at pitches closest to those in the syllables of the speech sample. In other work, Shorey *et al.* (2022) modified a talker adaptation paradigm by having participants classify the pitch of a tone when played by one instrument (tenor saxophone) or multiple instruments in random orders (marimba, piano, clarinet, French horn). Finally, Vatakis and Spence (2006) compared perception of audiovisual asynchrony for speech (isolated phonemes, consonant-vowel syllables) to that of music (individual notes or dyads played on a guitar or piano).

The closest that these studies came to modeling speech acoustics is Tao *et al.* (2021), in which piano notes were selected to follow the pitch contour of a four-syllable sentence. Otherwise, these studies did not model speech acoustics so much as they modeled listening situations common to speech perception and nonspeech perception. Stilp *et al.* (2010) modeled spectral properties of earlier sounds that influenced identification of later sounds. Shorey *et al.* (2022) selected musical instruments not to model speech but to extend an experimental paradigm used to explore speech perception amidst talker variability to nonspeech perception. Vatakis and Spence (2006) were not investigating speech perception versus music perception but audiovisual synchrony. This is evidenced by their experiments also containing a third condition that measured resilience to audiovisual asynchrony for video clips of object actions.

### F. Speech excerpts not recognized as speech

The final two sound classes reviewed here carry the distinction of best approximating the spectrotemporal

complexity of speech because these sounds are (or were) speech. Investigations of potential speech and nonspeech modes of perception have utilized speech signals but with the stipulation that they were not recognized as such. Mattingly *et al.* (1971) excised the second formant transition from stop consonants and presented them in isolation, which resulted in percepts of chirps. This laid the groundwork for later work on duplex perception where the excised formant transition was presented to the opposite ear as the rest of the otherwise intact syllable (Rand, 1974; Liberman *et al.*, 1981). Stephens and Holt (2003) compared context effects in the perception of /da-/ga/ target syllables to those in the perception of just the  $F_2$  and  $F_3$  transitions from the /d-/g/ consonant, which were not recognized as speech. Collectively, these studies leveraged the fact that English stop consonant formant transitions sound like frequency-modulated glides when isolated from the rest of the syllable. Other research explored discrimination of Zulu click consonants by native English speakers with no prior exposure to Zulu (Best *et al.*, 1988). Listeners described the click consonants as nonspeech “mouth sounds,” but their lack of experience with Zulu led them to perceive objectively (nonnative) speech sounds as nonspeech. Use of speech sounds or excerpts that are not recognized as speech raise important questions about the role of listening experience in nonspeech perception, addressed in Sec. III of this review.

### G. Sinewave speech

The nonspeech signal closest to speech reviewed here is sinewave speech. While it is more spectrally sparse than speech (or some of the other nonspeech stimuli listed here), it is the only stimulus class in this review that can be perceived as intelligible speech. Sinewave speech is created by tracking the first three formants in speech and replacing them with frequency-modulated and amplitude-modulated sinewaves (Remez *et al.*, 1981). While some listeners immediately recognize what is being said in sinewave speech, others do not (at least initially). This has yielded insightful experimental paradigms where different patterns of neural activation [as recorded via functional magnetic resonance imaging (fMRI)] are evident before and after recognizing the speech content in the same sinewave-speech stimulus (e.g., Möttönen *et al.*, 2006). Sinewave speech served as a starting point for Rosen *et al.* (2011), who investigated acoustic characteristics of speech that were essential for sentence intelligibility. They created noise-vocoded renditions of two-formant sinewave speech, then permuted whether the sentences had the frequency and/or amplitude modulations of natural speech. While two-formant stimuli without frequency modulations or amplitude modulations were “obviously unintelligible,” they measured keyword intelligibility and differential patterns of neural activation in positron emission tomography scans for these other renditions of these stimuli.

## III. ECOLOGICAL AND EXPERIENTIAL FACTORS

This review organized nonspeech stimuli according to their spectrotemporal complexity, but stimulus acoustics is

not the only key consideration when comparing nonspeech perception to speech perception. First, one ought to consider the acoustic ecology of auditory perception. The auditory system is not optimally designed to process every sound imaginable, nor does every sound imaginable occur in the sensory environment. Instead, sensory systems have adapted and evolved to encode the (types of) sounds that are commonly encountered. The best way to understand how sensory systems operate is to present stimuli that are naturalistic or as naturalistic as possible (Felsen and Dan, 2005; Einhauser and König, 2010). The nonspeech stimuli reviewed above vary dramatically from one another on acoustic grounds, but also in terms of their naturalness. The frequency with which one encounters sounds outside the lab and their role in everyday perception are valuable considerations when seeking to understand auditory (and speech) perception on the whole.

A second consideration is that of listening experience and/or recognition. While listeners have incomparable experience hearing speech, the amount of experience hearing the different nonspeech sounds reviewed above varies widely, from extremely high (speech segments that are not immediately recognized as speech) to considerable (music, environmental sounds) to very low (pure tones, random noise). Further, listening experience can be carefully separated from recognition when presenting speech segments that are not immediately recognized as speech (cf. Sec. II F). Whether this dissociation qualifies these sounds as nonspeech is a difficult question, as failure to recognize a particular sound does not immediately discount previous perceptual experience with it. Historically, part of the motivation in studying perception of nonspeech sounds is to control for the prodigious experience listeners have hearing speech. However, selection of a particular nonspeech stimulus raises a separate, important question of “how much less experience?” These questions become thornier in cases where sounds belonging to an unfamiliar nonnative language are initially perceived as nonspeech (e.g., native English listeners hearing Zulu click consonants) (Best *et al.*, 1988). The same stimulus can transition from being perceived as nonspeech to being perceived as speech as experience learning the nonnative language accumulates; identifying when exactly that switch from nonspeech to speech occurred is difficult. In all, listening experience varies widely across the gamut of nonspeech sounds, and it merits deep reflection during stimulus selection for future experiments.

#### IV. CONSIDERATIONS AND RECOMMENDATIONS

The sound class “nonspeech” contains tremendous acoustic diversity, even when limiting these sounds to those whose perception has been directly compared to perception of speech. We have deliberately refrained from using the results of these experiments to justify stimulus selection. Even from this neutral standpoint, there are clear considerations and recommendations for future research that seeks to compare perception of nonspeech sounds to perception of speech sounds.

Recently, Schutz and Gillard (2020) surveyed 1017 experiments from 443 published articles from leading journals (one of which was *JASA*) on the perception of nonspeech sounds. Their analysis focused primarily on the amplitude envelopes of these nonspeech stimuli, reporting that 89% of the stimuli surveyed had amplitude envelopes that lacked the dynamic variation typical of natural sounds (including speech). This raises a critically important point. For nonspeech perception to efficaciously model speech perception, the stimuli employed should model at least some of the acoustic properties of speech (and the more, the better); these have been highlighted toward the ends of many subsections of Sec. II. Conclusions that speech is perceived materially differently from nonspeech might be reached hastily if the nonspeech stimuli have far less spectrotemporal complexity than speech (as noise, pure tones, and at least some multitone complexes do). This is, in essence, an “apples to dump trucks” comparison. Conversely, selecting nonspeech stimuli whose spectrotemporal complexity better approximates that of speech will ensure that both stimulus sets are engaging similar underlying (at least lower-level) processing, making it more of an “apples to apples” comparison. Studies that report similar patterns of perception for speech and spectrotemporally simple stimuli (judiciously chosen to capture key acoustic properties of speech) should be replicable using more spectrotemporally complex (and thus, more ecological) stimuli that still capture those key acoustic properties. While replication of past work is a valuable enterprise, a stronger recommendation for future research is to design experiments that test multiple nonspeech stimuli varying in spectrotemporal complexity. This would move beyond treating nonspeech as a dichotomous alternative to speech, instead asking “where along this gradient of spectrotemporal complexity does the parallel between nonspeech perception and speech perception break down (if anywhere)?”

Which acoustic properties of the speech signal should nonspeech signals model? The timescale of the to-be-modeled speech is critically important. On brief timescales, the acoustic variability of human speech runs rampant. Given the acoustic diversity of shorter-duration speech signals, there will be instances in which a particular nonspeech stimulus matches key acoustic characteristics of speech rather well. This is not a universal justification for that particular nonspeech stimulus, as other instances exist where it will be a very poor match for the target speech. For example, a frequency-modulated pure tone might be selected to model the trajectory of tone four in Mandarin Chinese, but the same sound would be a poor choice to model the frication noise of /s/ in American English. Similarly, band limited random noise might be selected to model the /s/ spectrum but is a poor choice to model a Mandarin tone. Therefore, for nonspeech–speech comparisons on brief timescales, it matters to what one is comparing.

While the speech signal is highly acoustically variable on short timescales, it exhibits structured composition on slightly longer (1 s plus or sentence-length) timescales. When selecting nonspeech stimuli, this longer-timescale

structure of speech should be considered in the following ways. First, the long-term average speech spectrum has a characteristic shape defined by energy that is prominent at lower frequencies (<500 Hz) and drops off steadily across higher frequencies, a pattern that holds across different languages (Byrne *et al.*, 1994). Second, the speech signal is amplitude modulated at rates closely tied to speaking rate (Houtgast and Steeneken, 1985; Singh and Theunissen, 2003). Third, the speech spectrum also exhibits modulations, both at slow rates tied to formant-related harmonics and at faster rates tied to the talker's fundamental frequency (Elliott and Theunissen, 2009). To test the closest comparisons possible, nonspeech stimuli should exhibit at least some (if not most) of the acoustic properties characteristic of speech. Other entries may be readily added to this list, but the point remains that researchers should explicitly consider the ways in which their nonspeech stimuli mirror speech. This is a nontrivial task, as it invokes tradeoffs between experimenter control and acoustic ecology (Winn and Stilp, 2019), but such sacrifices in control may be desirable.

## V. CONCLUSION

Speech is far from being a homogenous sound class, well known for its extreme acoustic variability. This brief review reveals the same to hold true for nonspeech sounds, if not more so. Nonspeech perception has enjoyed a long history of comparisons to speech perception. However, the diversity and variability of nonspeech stimuli result in not all comparisons to speech perception being equally effective. One actionable proposal to improve the precision of future research in this area is to not simply motivate a particular nonspeech stimulus "because it is not speech," but enumerate which aspects of the speech signal (if any) are being modeled by that particular stimulus. Discussion of why a particular nonspeech stimulus was selected over other alternatives would be highly illuminating and improve methodological rigor; comparing perception of multiple different nonspeech stimuli to speech in the same experiment even more so. Going forward, these speech/nonspeech comparisons may be strengthened and solidified through careful considerations of the nonspeech stimuli's faithfulness to speech acoustics, role within everyday hearing, listening experience, and recognition.

## ACKNOWLEDGMENTS

The authors thank Eleanor Chodroff, Alex Francis, and an anonymous reviewer for invaluable feedback on previous versions of this manuscript. The authors also thank the anonymous reviewer for suggesting discussion of nonspeech-speech comparisons on brief timescales (Sec. IV).

<sup>1</sup>Some authors wrote "nonspeech" as "non-speech." For simplicity, we conducted two literature searches, one with either spelling, then reported the upper limit of citations across searches. Summing the citation counts

across the two searches double-counts some articles (e.g., the search "speech and nonspeech and non-speech" in 2021 returned 277 results). The most recent literature searches were conducted on June 23, 2022.

<sup>2</sup>Some readers might argue that these stimuli should be organized in a different order; others might argue that some additional stimulus or stimuli should be added to this list. In both cases, the stimuli presented here satisfy the authors' main objective of highlighting the wide range of spectrotemporal complexity that exists under the moniker "nonspeech."

- Ballas, J. A. (1993). "Common factors in the identification of an assortment of brief everyday sounds," *J. Exp. Psychol. Hum. Percept. Perform.* **19**(2), 250–267.
- Bent, T., and Pisoni, D. B. (2008). "Comparisons in perception between speech and non-speech signals," in *Handbook of Clinical Linguistics*, edited by M. J. Ball, M. R. Perkins, N. Muller, and S. Howard (Blackwell, Malden, MA), pp. 400–411.
- Best, C. T., McRoberts, G. W., and Sithole, N. M. (1988). "Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants," *J. Exp. Psychol. Hum. Percept. Perform.* **14**(3), 345–360.
- Bosker, H. R. (2017). "Accounting for rate-dependent category boundary shifts in speech perception," *Atten. Percept. Psychophys.* **79**(1), 333–343.
- Byrne, D., Dillon, H., Tran, K., Arlinger, S., Wilbraham, K., Cox, R., Hagerman, B., Hetu, R., Kei, J., Lui, C., Kiessling, J., Kotby, M. N., Nasser, N. H. A., El Kholly, W. A. H., Nakanishi, Y., Oyer, H., Powell, R., Stephens, D., Meredith, R., Sirimanna, T., Tavartkiladze, G., Frolenkov, G. I., Westerman, S., and Ludvigsen, C. (1994). "An international comparison of long-term average speech spectra," *J. Acoust. Soc. Am.* **96**(4), 2108–2120.
- Cutting, J. E. (1982). "Plucks and bows are categorically perceived, sometimes," *Percept. Psychophys.* **31**(5), 462–476.
- Cutting, J. E., and Rosner, B. S. (1974). "Categories and boundaries in speech and music," *Percept. Psychophys.* **16**(3), 564–570.
- Diehl, R. L., and Walsh, M. A. (1989). "An auditory basis for the stimulus-length effect in the perception of stops and glides," *J. Acoust. Soc. Am.* **85**(5), 2154–2164.
- Eimas, P. D. (1963). "The relation between identification and discrimination along speech and non-speech continua," *Lang. Speech* **6**(4), 206–217.
- Einhäuser, W., and König, P. (2010). "Getting real-sensory processing of natural stimuli," *Curr. Opin. Neurobiol.* **20**(3), 389–395.
- Elliott, T. M., and Theunissen, F. E. (2009). "The modulation transfer function for speech intelligibility," *PLoS Comput. Biol.* **5**(3), e1000302.
- Felsen, G., and Dan, Y. (2005). "A natural approach to studying vision," *Nat. Neurosci.* **8**(12), 1643–1646.
- Fowler, C. A., and Rosenblum, L. D. (1990). "Duplex perception: A comparison of monosyllables and slamming doors," *J. Exp. Psychol. Hum. Percept. Perform.* **16**(4), 742–754.
- Gygi, B., Kidd, G. R., and Watson, C. S. (2004). "Spectral-temporal factors in the identification of environmental sounds," *J. Acoust. Soc. Am.* **115**(3), 1252–1265.
- Gygi, B., and Shafiro, V. (2010). "Development of the database for environmental sound research and application (DESRA): Design, functionality, and retrieval considerations," *EURASIP J. Audio Speech Music Process.* **2010**, 1–12.
- Healy, A. F., and Repp, B. H. (1982). "Context independence and phonetic mediation in categorical perception," *J. Exp. Psychol. Hum. Percept. Perform.* **8**(1), 68–80.
- Holt, L. L. (2005). "Temporally nonadjacent nonlinguistic sounds affect speech categorization," *Psychol. Sci.* **16**(4), 305–312.
- Holt, L. L. (2006). "The mean matters: Effects of statistically defined nonspeech spectral distributions on speech categorization," *J. Acoust. Soc. Am.* **120**(5), 2801–2817.
- Houtgast, T., and Steeneken, H. J. M. (1985). "A review of the MTF concept in room acoustics and its use for estimating speech-intelligibility in auditoria," *J. Acoust. Soc. Am.* **77**(3), 1069–1077.
- Kidd, G. R., Watson, C. S., and Gygi, B. (2007). "Individual differences in auditory abilities," *J. Acoust. Soc. Am.* **122**, 418–435.
- Kingston, J., Kawahara, S., Chambless, D., Key, M., Mash, D., and Watsky, S. (2014). "Context effects as auditory contrast," *Atten. Percept. Psychophys.* **76**, 1437–1464.

- Kluender, K. R., Diehl, R. L., and Wright, B. A. (1988). "Vowel-length differences before voiced and voiceless consonants: An auditory explanation," *J. Phon.* **16**(2), 153–169.
- Lieberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). "Perception of the speech code," *Psychol. Rev.* **74**(6), 431–461.
- Lieberman, A. M., Isenberg, D., and Rakerd, B. (1981). "Duplex perception of cues for stop consonants: Evidence for a phonetic mode," *Percept. Psychophys.* **30**(2), 133–143.
- Locke, S., and Kellar, L. (1973). "Categorical perception in a non-linguistic mode," *Cortex* **9**(4), 355–369.
- Lotto, A. J., and Kluender, K. R. (1998). "General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification," *Percept. Psychophys.* **60**(4), 602–619.
- Mann, V. A. (1980). "Influence of preceding liquid on stop-consonant perception," *Percept. Psychophys.* **28**(5), 407–412.
- Mattngly, I. G., Liberman, A. M., Syrdal, A. K., and Halwes, T. (1971). "Discrimination in speech and nonspeech modes," *Cogn. Psychol.* **2**(2), 131–157.
- Miller, J. D., Wier, C. C., Pastore, R. E., Kelly, W. J., and Dooling, R. J. (1976). "Discrimination and labeling of noise–buzz sequences with varying noise-lead times: An example of categorical perception," *J. Acoust. Soc. Am.* **60**(2), 410–417.
- Mirman, D., Holt, L. L., and McClelland, J. L. (2004). "Categorization and discrimination of nonspeech sounds: Differences between steady-state and rapidly-changing acoustic cues," *J. Acoust. Soc. Am.* **116**(2), 1198–1207.
- Möttönen, R., Calvert, G. A., Jääskeläinen, I. P., Matthews, P. M., Thesen, T., Tuomainen, J., and Sams, M. (2006). "Perceiving identical sounds as speech or non-speech modulates activity in the left posterior superior temporal sulcus," *Neuroimage* **30**(2), 563–569.
- Parker, E. M., Diehl, R. L., and Kluender, K. R. (1986). "Trading relations in speech and nonspeech," *Percept. Psychophys.* **39**(2), 129–142.
- Pisoni, D. B. (1977). "Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops," *J. Acoust. Soc. Am.* **61**(5), 1352–1361.
- Pisoni, D. B., Carrell, T. D., and Gans, S. J. (1983). "Perception of the duration of rapid spectrum changes in speech and nonspeech signals," *Percept. Psychophys.* **34**(4), 314–322.
- Rand, T. C. (1974). "Dichotic release from masking for speech," *J. Acoust. Soc. Am.* **55**, 678–680.
- Reddy, R. K., Ramachandra, V., Kumar, N., and Singh, N. C. (2009). "Categorization of environmental sounds," *Biol. Cybern.* **100**, 299–306.
- Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. (1981). "Speech-perception without traditional speech cues," *Science* **212**(4497), 947–950.
- Rosen, S., and Iverson, P. (2007). "Constructing adequate non-speech analogues: What is special about speech anyway?," *Dev. Sci.* **10**(2), 165–168.
- Rosen, S., Wise, R. J. S., Chadha, S., Conway, E. J., and Scott, S. K. (2011). "Hemispheric asymmetries in speech perception: Sense, nonsense and modulations," *PLoS One* **6**(9), e24672.
- Schroeder, M. R. (1968). "Reference signal for signal quality studies," *J. Acoust. Soc. Am.* **44**(6), 1735–1736.
- Schutz, M., and Gillard, J. (2020). "On the generalization of tones: A detailed exploration of non-speech auditory perception stimuli," *Sci. Rep.* **10**(1), 9520.
- Shafiro, V., and Gygi, B. (2004). "How to select stimuli for environmental sound research and where to find them?," *BRMIC* **36**, 590–598.
- Shorey, A. E., King, C. J., Theodore, R. M., and Stilp, C. E. (2022). "Talker adaptation or 'talker' adaptation? Musical instrument variability impedes pitch perception," *J. Acoust. Soc. Am.* **151**, A222.
- Siegel, J. A., and Siegel, W. (1977). "Categorical perception of tonal intervals: Musicians can't tell sharp from flat," *Percept. Psychophys.* **21**(5), 399–407.
- Singh, N. C., and Theunissen, F. E. (2003). "Modulation spectra of natural sounds and ethological theories of auditory processing," *J. Acoust. Soc. Am.* **114**(6 Pt. 1), 3394–3411.
- Stephens, J. D., and Holt, L. L. (2003). "Preceding phonetic context affects perception of nonspeech," *J. Acoust. Soc. Am.* **114**(6 Pt. 1), 3036–3039.
- Stephens, J. D. W., and Holt, L. L. (2011). "A standard set of American-English voiced stop-consonant stimuli from morphed natural speech," *Speech Commun.* **53**(6), 877–888.
- Stilp, C. E., Alexander, J. M., Kieffe, M., and Kluender, K. R. (2010). "Auditory color constancy: Calibration to reliable spectral properties across nonspeech context and targets," *Atten. Percept. Psychophys.* **72**(2), 470–480.
- Stilp, C. E., Kieffe, M., and Kluender, K. R. (2018). "Discovering acoustic structure of novel sounds," *J. Acoust. Soc. Am.* **143**(4), 2460–2473.
- Studdert-Kennedy, M., Liberman, A. M., Harris, K. S., and Cooper, F. S. (1970). "Motor theory of speech perception: A reply to Lane's critical review," *Psychol. Rev.* **77**(3), 234–249.
- Tao, R., Zhang, K., and Peng, G. (2021). "Music does not facilitate lexical tone normalization: A speech-specific perceptual process," *Front. Psychol.* **12**(14), 717110–717111.
- Vatakis, A., and Spence, C. (2006). "Audiovisual synchrony perception for music, speech, and object actions," *Brain Res.* **1111**(1), 134–142.
- Wade, T., and Holt, L. L. (2005a). "Incidental categorization of spectrally complex non-invariant auditory stimuli in a computer game task," *J. Acoust. Soc. Am.* **118**(4), 2618–2633.
- Wade, T., and Holt, L. L. (2005b). "Perceptual effects of preceding non-speech rate on temporal properties of speech categories," *Percept. Psychophys.* **67**(6), 939–950.
- Watkins, A. J. (1991). "Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion," *J. Acoust. Soc. Am.* **90**(6), 2942–2955.
- Winn, M. B., and Stilp, C. E. (2019). "Phonetics and the auditory system," in *The Routledge Handbook of Phonetics*, edited by W. F. Katz, and P. F. Assmann (Routledge, New York), pp. 164–192.