JASA TUTORIAL

How to vocode: Using channel vocoders for cochlear-implant research

Margaret Cychosz,[1,a)] Matthew B. Winn,[2] and Matthew J. Goupell[3]

[1]*Department of Linguistics, University of California, Los Angeles, Los Angeles, California 90095, USA*

[2]*Department of Speech-Language-Hearing Sciences, University of Minnesota, Minneapolis, Minnesota 55455, USA*

[3]*Department of Hearing and Speech Sciences, University of Maryland, College Park, College Park, Maryland 20742, USA*

The channel vocoder has become a useful tool to understand the impact of specific forms of auditory degradation—particularly the spectral and temporal degradation that reflect cochlear-implant processing. Vocoders have many parameters that allow researchers to answer questions about cochlear-implant processing in ways that overcome some logistical complications of controlling for factors in individual cochlear implant users. However, there is such a large variety in the implementation of vocoders that the term "vocoder" is not specific enough to describe the signal processing used in these experiments. Misunderstanding vocoder parameters can result in experimental confounds or unexpected stimulus distortions. This paper highlights the signal processing parameters that should be specified when describing vocoder construction. The paper also provides guidance on how to determine vocoder parameters within perception experiments, given the experimenter's goals and research questions, to avoid common signal processing mistakes. Throughout, we will assume that experimenters are interested in vocoders with the specific goal of better understanding cochlear implants. © *2024 Acoustical Society of America*.

https://doi.org/10.1121/10.0025274

(Received 19 July 2023; revised 14 February 2024; accepted 23 February 2024; published online 3 April 2024)

[Editor: Emily Buss] Pages: 2407–2437

TABLE OF CONTENTS

I. INTRODUCTION

The channel vocoder can be used as a method to reduce the complexity of audio information in a highly controlled way so that we can learn about crucial aspects of signal

[a)] Email: mcychosz4@ucla.edu.

transmission and auditory perception. In general, a rich input signal is replaced by a sparse output signal. The listener can then attempt to recover enough of the original information to perform a task such as understanding a word. In the case of speech understanding, a high level of intelligibility can remain in the vocoded signal despite extremely sparse spectro-temporal resolution (Shannon *et al.*, 1995). There are numerous research questions (and perhaps entire fields of research) that can be pursued with vocoders. However, the term "vocoder" is not specific enough to describe the signal processing used in an experiment, because vocoding is an entire class of signal processing and encompasses many different variations. Details of vocoder algorithms can differ dramatically, in ways that have meaningful implications for how we interpret experimental results. Toward the goal of addressing common misunderstandings and reducing common mistakes, this paper describes the processing stages in a vocoder. It also provides guidance on how to set vocoder parameters in a way that is aligned with experimenters' research goals and questions.

Although vocoders have become synonymous with cochlear implant (CI) research, the history of vocoders does not begin with speech processing in CIs. The channel vocoder was first developed in the 1930s by Homer Dudley of Bell Labs as a means to compress and transmit speech via underwater telephone lines (Dudley, 1939). Although never employed for that application, the United States military did use speech vocoding readily from the Second World War through the Vietnam War as a telecommunication encryption method. Currently, speech encoders are prominent in many music genres and can be heard in the work of artists like Daft Punk, Tupac, and Kavinsky. Within the scientific community, the vocoder is often used to degrade aspects of the auditory signal and answer questions about auditory processing and speech perception. The most common use of a vocoder in this realm is to simulate aspects of CI processing (Shannon *et al.*, 1995).

The CI is one of the most successful neural prostheses to date, with over one million ears implanted, including more than 65 000 devices implanted in children (NIDCD, 2021; Zeng, 2022). One of the main purposes of CIs is to partially restore access to speech information for communication. While research with CIs has significant public health benefits, it has numerous complications and logistical challenges. CI users vary greatly in relevant individual factors such as duration of deafness, the level and etiology of hearing loss, and remaining cochlear and neural function, as well as cognitive abilities that underlie speech recognition (Boisvert *et al.*, 2020; Pisoni *et al.*, 2017). Effects of implantation age, degree of hearing loss, and auditory system health are often confounded within individual users—for example, children with greater levels of hearing loss are more likely to be implanted at a younger age—and so these variables are rarely under researcher control or available to isolate and examine empirically (Blamey *et al.*, 2013). The consequence of this subject-level variability, as well as the inability to isolate certain variables, is the highly variable response data often seen in research with CI users [e.g., Friesen *et al.* (2001)]. These facts, combined with the challenges of recruiting members of a relatively low-incidence clinical population and the numerous choices in the signal processing that are made in vocoder-centric speech processing, result in a need for approaches and tools to better control and investigate the parameters that contribute to speech perception with electrical stimulation.

Given these challenges, an alternative approach is to conduct CI simulations in individuals with normal hearing (NH) who are presented with signals processed to emulate aspects of the electric hearing experience of CI users with a channel vocoder. This approach has become an important and practical tool to better understand how individuals with hearing loss and CIs perform on auditory, speech, and language tasks, as well as to assess how the hearing experience of CI users [for example, by presenting a vocoded signal to the NH ear of a CI user with single-sided deafness (Dorman *et al.*, 2020)]. Vocoder CI simulations permit explicit experimental control over variables that are confounded within individual CI users and allow larger sample sizes that can help elucidate important effects. While acoustic simulations are never meant to *exactly* mimic CI users' experiences, results from vocoder studies have clarified details about hearing with a CI that researchers might otherwise be unable to disentangle. For example, channel vocoders have been used in CI research to suggest that shallower CI array insertion depths appear to result in poorer speech perception outcomes [e.g., Rosen *et al.* (1999) and Shannon *et al.* (1998)]. They have also been used to outline how interactions between channels in the CI signal and listeners' dynamic range impact speech recognition (Grange *et al.*, 2017; Oxenham and Kreft, 2014; Stafford *et al.*, 2014), as well as how degraded speech signals may impact language development [e.g., Newman *et al.* (2020)], listening effort (DeRoy Milvae *et al.*, 2021; Winn *et al.*, 2015), and auditory pattern recognition skills that support word learning (Grieco-Calub *et al.*, 2017).

Despite the advantages of vocoder simulations for CI research, it is important to emphasize that many aspects of the CI experience are not accounted for in vocoder CI simulations and cannot be simulated via signal processing. There are inherent differences between the electromechanical transduction of an acoustic-hearing system and the purely electrical transduction of a CI system. For example, CI electrode stimulation is spectrally imprecise because of current spread and channel interference (Srinivasan *et al.*, 2012). Electrical stimulation of the auditory nerve also results in a reduced dynamic range and high temporal synchrony (Brown *et al.*, 1995; Hughes *et al.*, 2001). Acoustic simulations of electrical hearing will necessarily produce a traveling wave and cochlear filtering even though those are absent from electrical stimulation.

A vocoder also does not simulate the lifestyle changes and social dynamics that come with living with hearing loss, nor the process of weighing the benefit of engaging in conversation against the cost of listening fatigue (Hughes *et al.*, 2018). Finally, unlike in typical vocoder studies where NH individuals have short-term exposure to vocoded speech stimuli, CI users exhibit great plasticity in performance as they become accustomed to their devices [especially over the first 6–12 months (Lenarz *et al.*, 2012)] and this is an aspect of the CI experience that is difficult to emulate via vocoder simulations. Overall, listening through a vocoder can simulate some aspects of the *sound* quality of a CI, but does not simulate the true *experience* of being a CI user every day. Therefore, it is essential that experimenters do not consider vocoders to be true "CI simulations."

We have structured this paper as follows: we first introduce electrical stimulation of the auditory nerve to explain modern CI processing and configurations (Sec. II). Then, we provide a *general* overview of vocoder design (Secs. III and IV) followed by individual, in-depth explanations of each step in vocoder processing and how it is meant to emulate certain aspects of CI processing. After describing each step of vocoder processing in detail, we discuss how to report vocoder designs in scientific studies (Sec. V) and we devote a section to describing which aspects of CIs can be simulated using vocoders (Sec. VI). Finally, we conclude with a general discussion of how vocoders have been employed in the literature to study the impact of CIs across the lifespan (Sec. VII).

II. OVERVIEW OF ELECTRICAL STIMULATION OF THE AUDITORY NERVE

Standard CI systems receive acoustic input from a microphone attached to a sound processor. The input is then analyzed by a bank of 12–22 frequency channels that cover the frequency range important for speech perception (approximately 200–8000 Hz). The exact number of electrodes (and therefore number of frequency channels) along the array varies by manufacturer and device (e.g., 16 for Advanced Bionics, 22 for Cochlear, and 12 for MED-EL). Consequently, there are relatively few separate places of cochlear stimulation when compared to what listeners with NH can access, so the frequency resolution in a CI system is much worse than in NH (Mehta *et al.*, 2020).

In CI processing, the temporal envelope is extracted from each spectral channel, conveying the slowly varying changes in intensity that reflect syllables, pauses, consonant onsets and, sometimes the fundamental voice frequency (f0). Faster changes in sound pressure—called temporal fine structure—are usually discarded in this phase [for more details about the nature of these timing categories see Rosen (1992)]. However, information from the slowly varying envelope is sufficient to understand speech in quiet (Drullman, 1995). Those temporal envelopes are used to modulate a series of periodic electrical pulses (generally ≥ 900 pps)) whose fixed rate does not reflect the temporal fine structure of the original input signal.[1] Those modulated electrical pulse trains are emitted from electrodes in a multi-electrode array that was implanted in the cochlea, ordered in a way that is designed to capitalize on the tonotopy of the cochlea to recreate a sparse version of the incoming sound spectrum. In doing so, the CI bypasses the outer- and middle-ear structures and therefore bypasses the sound filtering and processing that is normally done in those parts of the auditory system.

CIs capitalize on the tonotopic organization of the cochlea by stimulating electrodes at the apical region when low-frequency energy is detected by the microphone, and stimulating electrodes at the basal region when high-frequency energy is detected. However, the electrodes in a CI are generally not perfectly aligned with the frequency regions that they are meant to represent. This pattern leads to some amount of tonotopic mismatch that will vary by individual (Dillon *et al.*, 2022; Dorman *et al.*, 1997a).

A number of stimulation techniques have been developed to increase the frequency resolution available to CI users, usually by adjusting the simultaneous stimulation of CI electrodes. For example, current steering re-directs stimulation to an area along the cochlea between two conventional electrode contacts, perhaps permitting activation of intermediate sites of stimulation, while current focusing stimulates regions along the cochlea that are narrower than what would conventionally be excited (Berenstein *et al.*, 2008).

Similarly, a variety of processing techniques help alleviate issues with channel interaction along the electrode array. The continuous interleaved sampling strategy employs interleaved pulses to ensure non-simultaneous activation of (adjacent) electrodes, which prevents supra-additive interactions at overlapping areas of stimulation. Channel peak picking is a strategy that ensures that only a subset of electrodes are selected and activated during each stimulation cycle, again helping address channel interaction by removing low-intensity information that might other "clutter" the spectrum (Winn *et al.*, 2015). See Sec. III C for detail.

In summary, a CI reduces the spectral resolution of an acoustic input like speech by having a finite number of spectral channels. It replaces the temporal fine structure of each channel by modulating an unrelated carrier, typically a periodic high-rate electrical pulse train. This degraded spectro-temporal information is conveyed directly to the auditory nerve in a roughly tonotopically organized fashion akin to NH. A full review of how present-day CIs process acoustic signals can be found in Loizou (2006) and Macherey and Carlyon (2014).

CIs can be appropriate for a number of hearing configurations that can be simulated using vocoders. Individuals could have access to sound in both ears, either electrical or a mix of acoustic and electrical. For example, a person could have bilateral CIs. Or, individuals with severe-to-profound hearing loss in only one ear (i.e., single-sided deafness) could have a CI in one ear and typical acoustic hearing in the other. Or that acoustic hearing ear could have some hearing loss; it is common to have access to low-frequency acoustic hearing only in one ear, which might only be accessible through a hearing aid. In some instances, a person can be implanted with a hybrid CI, which has a shorter electrode array and shallower insertion depth than standard implantations, allowing the user possibly better access to low-frequency residual hearing in the same ear that is implanted. These are all configurations that can be simulated via various vocoder designs (Dillon *et al.*, 2022; Qin and Oxenham, 2006; Stilp *et al.*, 2016).

III. OVERVIEW OF VOCODER CONSTRUCTION: ANALYSIS STAGE

Like a CI, a vocoder typically divides the input speech signal into a set of frequency analysis bands (called the ANALYSIS PHASE) and extracts the slowly varying temporal envelope (amplitude contour) from each band. This is the primary similarity between CIs and vocoders. However, the delivery of that signal (called the SYNTHESIS PHASE) via a vocoder is with acoustic rather than electrical simulation, implying a number of constraints that will be explored later in this paper. For each channel in the vocoder, the temporal envelope is imposed upon a CARRIER, such as a sinewave or noise-band, that often corresponds loosely to the frequency of the original input analysis band (in a way that is similar to the envelope modulating electrical pulse trains in a CI). Finally, the filtered carriers are summed together to create the vocoded sound that is presented to listeners. Figure 1 illustrates the typical main components and order of vocoding, proceeding sequentially from the input signal to the output. These steps outline a relatively basic vocoder design. However, there are several implementation choices within each step, including ways to add complexity to the signal and simulate various components of CI signals. In the following, we will describe the steps behind vocoding, outline choices that practitioners can make to simulate different CI configurations and settings, and how choices at each stage may impact the final vocoded signal.

A. Filterbank and number of channels

The first step taken in vocoder construction is to apply a bank of digital filters that divides the frequency spectrum into channels. Some elements of a vocoder's filterbank that are commonly manipulated in vocoder CI simulations are the (1) number of filters/channels, (2) lower and upper limits of the frequency range, and (3) filter slope. Usually anywhere from 4 to 32 channels are created—the exact number will partially determine the spectral resolution of the output signal and is often used to simulate "number of activated electrodes" (see Sec. VI A).

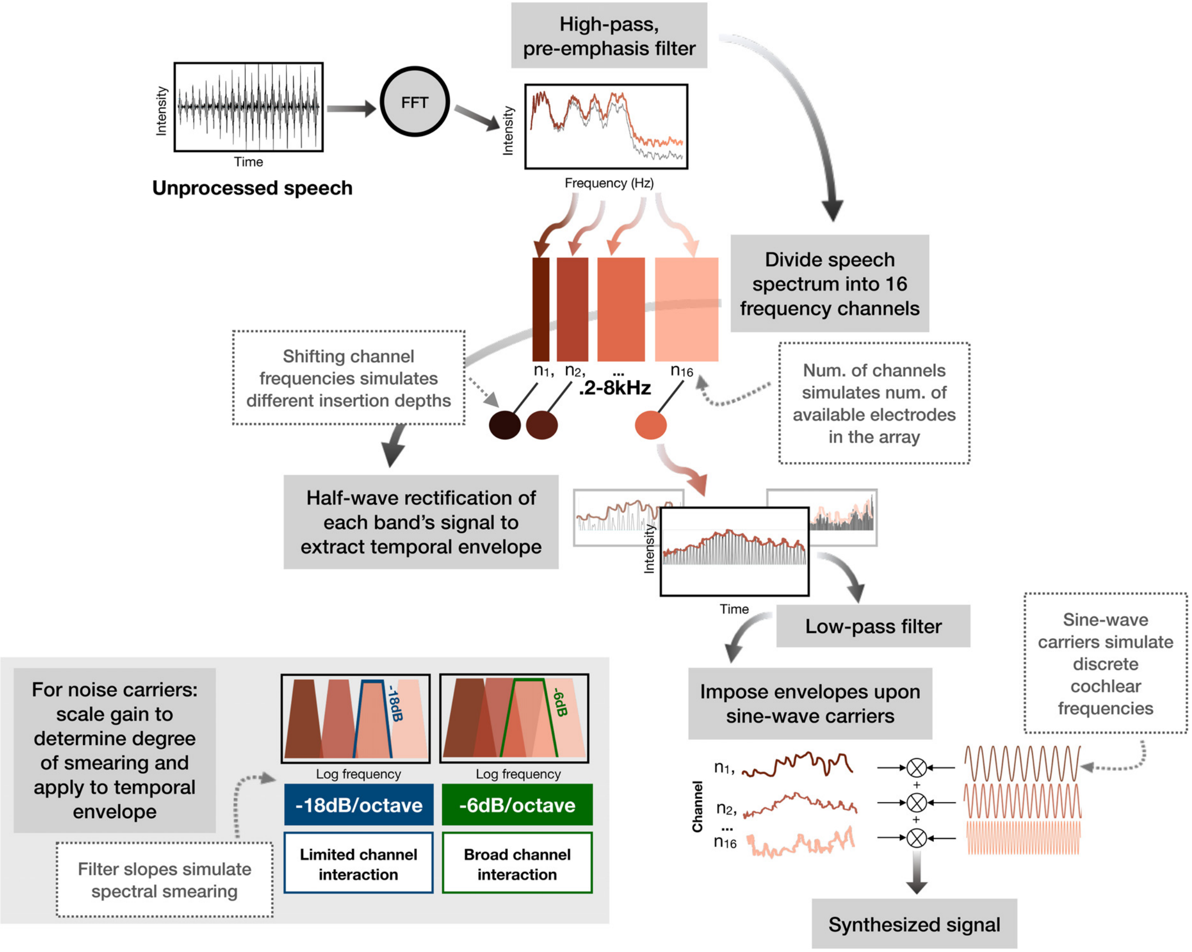

High-pass, pre-emphasis filter
Intensity
Time
Unprocessed speech
FFT
Intensity
Frequency (Hz)
Divide speech spectrum into 16 frequency channels
Shifting channel frequencies simulates different insertion depths
n1, n2, ... n16
.2-8kHz
Num. of channels simulates num. of available electrodes in the array
Half-wave rectification of each band's signal to extract temporal envelope
Intensity
Time
Low-pass filter
Sine-wave carriers simulate discrete cochlear frequencies
For noise carriers: scale gain to determine degree of smearing and apply to temporal envelope
-18dB
Log frequency
-18dB/octave
Limited channel interaction
-6dB
Log frequency
-6dB/octave
Broad channel interaction
Filter slopes simulate spectral smearing
Impose envelopes upon sine-wave carriers
Channel
n1,
n2,
...
n16
+
+
Synthesized signal

FIG. 1. (Color online) Overview of vocoder construction. Each step in the signal processing is listed in light gray boxes. Vocoder settings that practitioners can adjust are listed in white boxes with dotted borders.

Most vocoder implementations use logarithmically scaled frequency channels, reflecting that the auditory system (the basilar membrane) is roughly logarithmically scaled (Cheung *et al.*, 2016). For example, the first channel in the digital filterbank could include the energy between 100 and 200 Hz, the second channel the energy between 200 and 400 Hz, and so on. The frequency-to-electrode allocation is typically based on the Greenwood (1990) map of the basilar membrane, which assumes equal spacing of channels in physical cochlear space. This methodological tendency is well motivated but imperfect for at least two reasons. First, CIs bypass the basilar membrane, instead stimulating the spiral ganglion directly, and the spiral ganglion has a slightly different frequency-to-place mapping, with the largest deviations between those frequency maps found near the apex (Stakhovskaya *et al.*, 2007). Second, the electrodes in a CI are not necessarily equally spaced. Nevertheless, the Greenwood map is a common model for choosing channel analysis and carrier frequencies.

The absolute lower and upper frequency limits contextualize the number of channels; 12 frequency channels spanning the range 100 Hz to 10 000 Hz would provide poorer resolution than 12 channels spanning 200 to 5000 Hz, so it is not sufficient to merely report the number of channels used. Depending on the implementation of the filterbank, there are practical limits for the number of channels that can be used before encountering undesirable technical limitations. These limitations are discussed in detail later, but a general heuristic is that more channels likely requires a higher filter order (i.e., steeper filter slopes), which runs the risk of creating unstable filters that cause distortion (see Sec. VI B and the Appendix).

Although one can filter the spectrum into discrete "rectangular" bands, this is not always the goal. The carriers can instead be shaped to mimic the non-rectangular spread of excitation that results from large intracochlear electrical fields generated at each electrode (see Sec. VI B). For the carrier filter slope, anywhere from −6 to −24 dB/octave slopes are common (with larger numbers corresponding to steeper slopes that emulate less interaction and therefore a clearer signal). Figure 2 illustrates the difference between rectangular and sloping carrier filter carrier channels as they would represent a vowel sound.

B. Envelope extraction

After constructing the filterbank, the amplitude envelope is extracted from each channel. This extraction is either done by calculating the Hilbert envelope of the signal or by half- or full-wave rectification [this is a minor detail; there is no evidence of a benefit for one method or the other (Loizou, 2006)]. Then, a low-pass filter is typically applied to each envelope. Typical low-pass filter cut-off values of envelope extraction range from 50 to 400 Hz, with filter orders often ranging from 1 to 4 (–6 to –48 dB/octave). The chosen cut-off frequency for the filter depends upon the research question being asked and which acoustic cues the experimenter wishes to preserve (Fig. 3), as well as the carrier. During speech production, the human jaw oscillates at approximately 4 to 8 Hz, reflecting the average rate of syllables per second, so temporal signal frequencies around 6 Hz are critical for recognizing most speech. Similarly, the fundamental frequency of the human voice source ranges from 80 Hz (adult cisgender men) to upwards of 400 Hz (infants). So, temporal modulation frequencies within this range can be important cues for pitch identification and discrimination. Higher cut-off frequencies (300 to 400 Hz) will preserve fast modulations that include periodicity cues to the voice pitch, while lower cut-offs around 50 Hz will eliminate those pitch cues while preserving phonetic contrasts for different manners of articulation as well as slower cues for syllables timing. Preservation of very high-rate envelope modulations does not necessarily mean that they will be useful for speech perception, particularly because temporal modulation detection thresholds begin to decrease for modulation rates

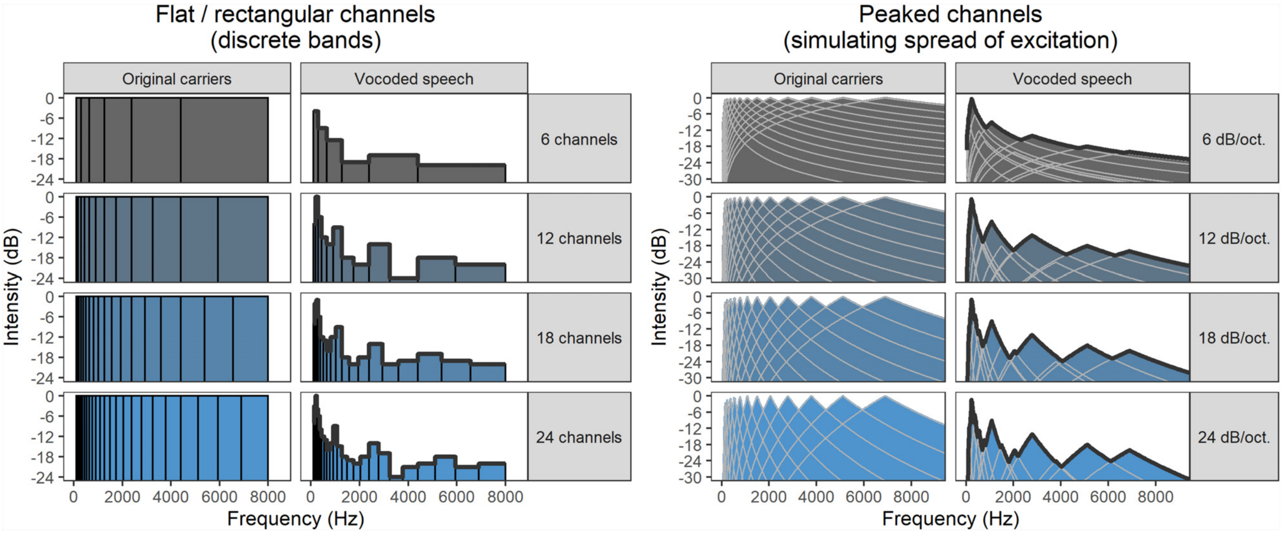


FIG. 2. (Color online) Two styles of filtering carrier bands in a vocoder. On the left, better spectral resolution results from an increasing number of bands. On the right, the same number of channels are present in each row, and better spectral resolution instead results from steeper filter slope.

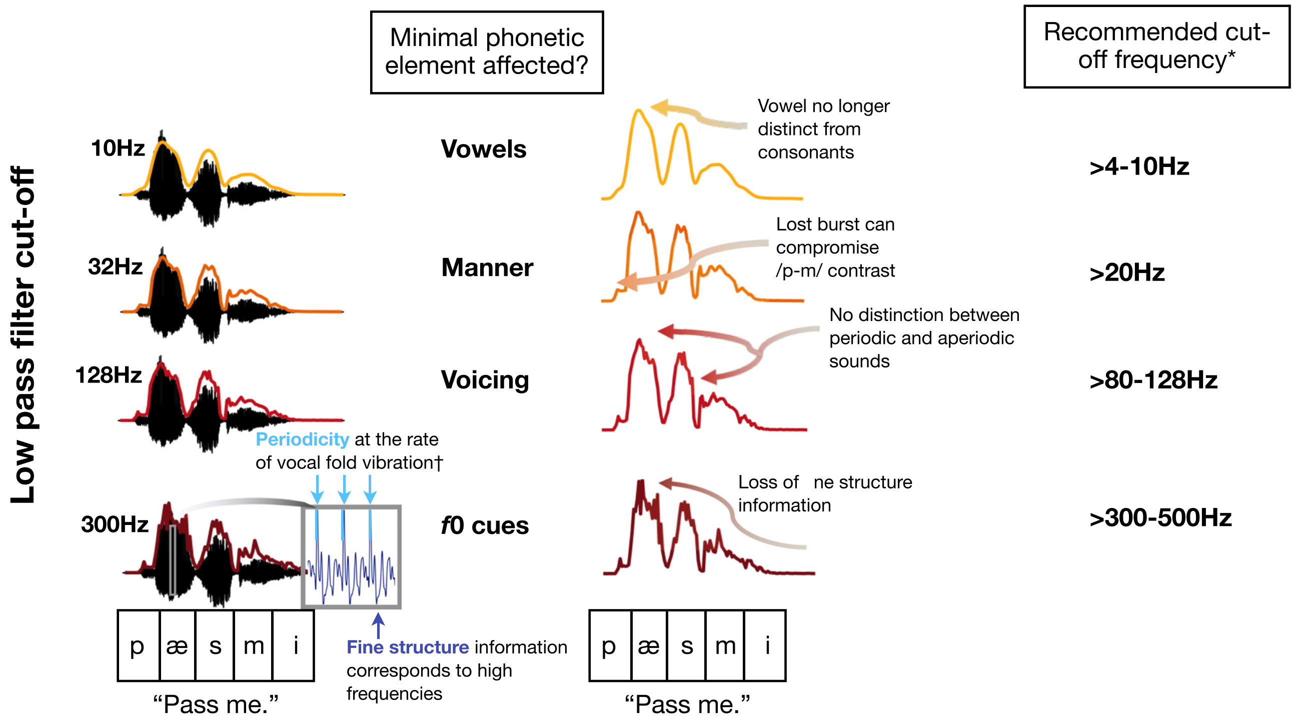

Minimal phonetic element affected?
Recommended cut-off frequency*
Low pass filter cut-off
10Hz
Vowels
Vowel no longer distinct from consonants
>4-10Hz
32Hz
Manner
Lost burst can compromise /p-m/ contrast
>20Hz
128Hz
Voicing
No distinction between periodic and aperiodic sounds
>80-128Hz
Periodicity at the rate of vocal fold vibration†
300Hz
f0 cues
Loss of fine structure information
>300-500Hz
p æ s m i
Fine structure information corresponds to high frequencies
"Pass me."
p æ s m i
"Pass me."

FIG. 3. (Color online) The low-pass filter cut-off will affect different phonetic elements of speech. Cut-off frequencies form an implicational hierarchy: 4-Hz cut-off includes vowels and everything below, 32-Hz cut-off includes manner and everything below, etc. *Always stimulus-dependent. See text for detail.

greater than 10 Hz (Bacon and Viemeister, 1985; Shannon, 1992), become rather poor around 100 Hz, and completely disappear by around 1000 Hz. The amplitude envelope is therefore low-pass filtered at this stage in vocoder design to only include frequencies below a target range (e.g., modulations between 0 and 300 Hz).[2]

The range and depth of amplitude modulations in the signal will be limited by the spectral bandwidth of the channel analyzed. Modulations occur when two or more spectral components fall within the same filter, with modulations emerging at the rate equal to the linear difference of those components. Therefore, vocoders with a large number of channels are likely to have narrower channels and therefore have a limitation in the modulation rates that can be encoded in each channel. Figure 4 illustrates this concept using vocoders with different numbers of channels reflecting the same overall frequency range. In the 4-channel case, each channel spans a wide range, and the envelope (in red) therefore shows deep modulations reflecting the periodicity of the original speech (displayed above the channel breakout). This happens because each of the 4 channels is wide enough to capture many of the harmonics from the voice, each strengthening the periodic modulations. Conversely, the lower-frequency channels in the 16-channel vocoder only include one or two harmonics of the signal, and therefore only shallow modulations emerge when multiple spectral components interact. Consistent with this, modulations in the 16-channel vocoder are limited to the upper channels, which are wide enough (because they are logarithmically spaced) to include a larger number of linearly spaced harmonics. However, the relatively narrower bandwidth of the analysis channels in the 16-channel vocoder, combined with the relatively lower intensity of the upper-frequency components of speech, render these modulations less perceptible even if they can be visualized on the plot.

Previous work has experimentally manipulated envelope cut-off frequencies (e.g., 16, 64, 256 Hz) to evaluate the effect of amplitude envelope information bandwidth (temporal cues) upon speech perception in vocoder CI simulations (Shannon *et al.*, 1995; Xu and Pfingst, 2003; Xu *et al.*, 2005b). These manipulations have demonstrated that phoneme recognition may improve as a function of cut-off frequency from 1 to 512 Hz (in octave steps) and that fine structure information (>500 to 600 Hz) may be important for lexical tone perception [see Xu and Pfingst (2003); although in this work resolved harmonics also likely play a role]. These improvements are seen most dramatically when the spectral resolution is poorer (i.e., most apparent in 1-channel vocoders,[3] but are apparent up to as many as 12 channels). Thus, there could be a trade-off relationship between spectral and temporal fidelity, although evidence suggests that listeners also flexibly use those cues that are available to them (Xu *et al.*, 2005b).

C. Pre-emphasis and channel peak-picking

A common peak-picking strategy is to select the 8 highest energy channels out of 22 eligible channels at any particular time, with the channel evaluation updating on each cycle of electrode activation. The idea is to improve the salience of spectral peaks, and the overall spectral resolution, by de-cluttering the other intermediate channels whose

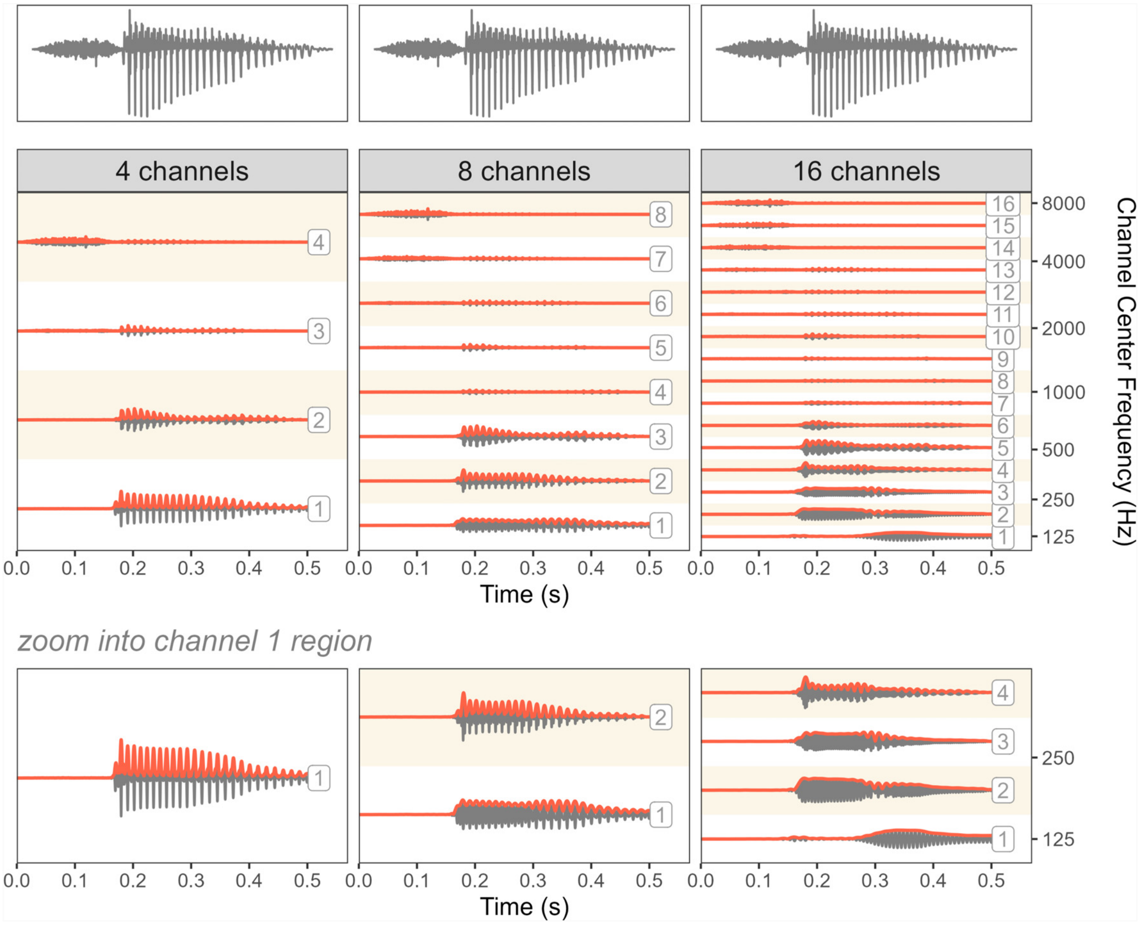


FIG. 4. (Color online) Envelope (in red) of individual spectral channels in speech (the word "sail"). The top row shows the input waveform. In the left column, the relatively small 4-channel filterbank results in wide spectral bandwidths; each channel captures a large number of spectral components (e.g., harmonics) that interact to produce periodic modulations. Conversely, in the right column, the relatively large 16-channel filterbank results in narrower spectral bandwidths; each channel captures fewer spectral components and therefore produce shallower modulations.

activation patterns might smear the informative valleys between spectral peaks.

There are two important caveats to this peak picking procedure that deserve consideration. First, the speech envelope is characterized by a −6 dB/octave loss of energy, or "energy decay," across the frequency range (Sun, 2000). Because of this, we would expect that the 8 highest energy channels would often be clustered near the low-frequency end of the spectrum, neglecting any important high-frequency information that is naturally lower in intensity. In order to preserve the salience of those high-frequency spectral peaks, which still contain information highly relevant for speech perception, one could counteract the natural roll-off in intensity via a PRE-EMPHASIS filter that compensates for this energy loss *before* the channels are evaluated for peak picking. In practice, there are ready-made pre-emphasis functions in most computing languages (and can be as simple as implementing a high-pass filter to the signal), or signal processing libraries (e.g., LIBROSA for PYTHON or native pre-emphasis filtering function in PRAAT). Readers are invited to consult the Appendix of this article for details of the underlying computation, and to understand how different amounts of pre-emphasis could impact their signal.

Pre-emphasis does not guarantee accurate representation of CI processing strategies, but it should be considered if you want to emulate a peak-picking strategy in your vocoded signal. If there is any effect of including pre-emphasis in CI simulations, it would likely only matter for broadband stimuli—particularly when attempting to emulate peak-picking strategies. Figure 5 illustrates the impact of pre-emphasis on channel peak picking. In this spectrogram representation of a woman saying /ɑ/, there are strong concentrations of energy (peaks) around 900, 1450, and 3000 Hz. However, that 3000-Hz spectral peak is a *local* peak and has relatively less absolute intensity compared to the lower-frequency peaks due to the sloping nature of the speech spectrum. Without pre-emphasis, the eight channels picked during a peak-picking processing strategy might all be below the third formant[4] (F3, or the 3000 Hz peak near channel 17), and the output might not manifest an energy peak for F3 at all. Figure 5 illustrates this possibility by displaying the channel outputs for a vocoder that picks the top 8 out of 22 channels for each 30-ms time bin (see Fig. 15 in

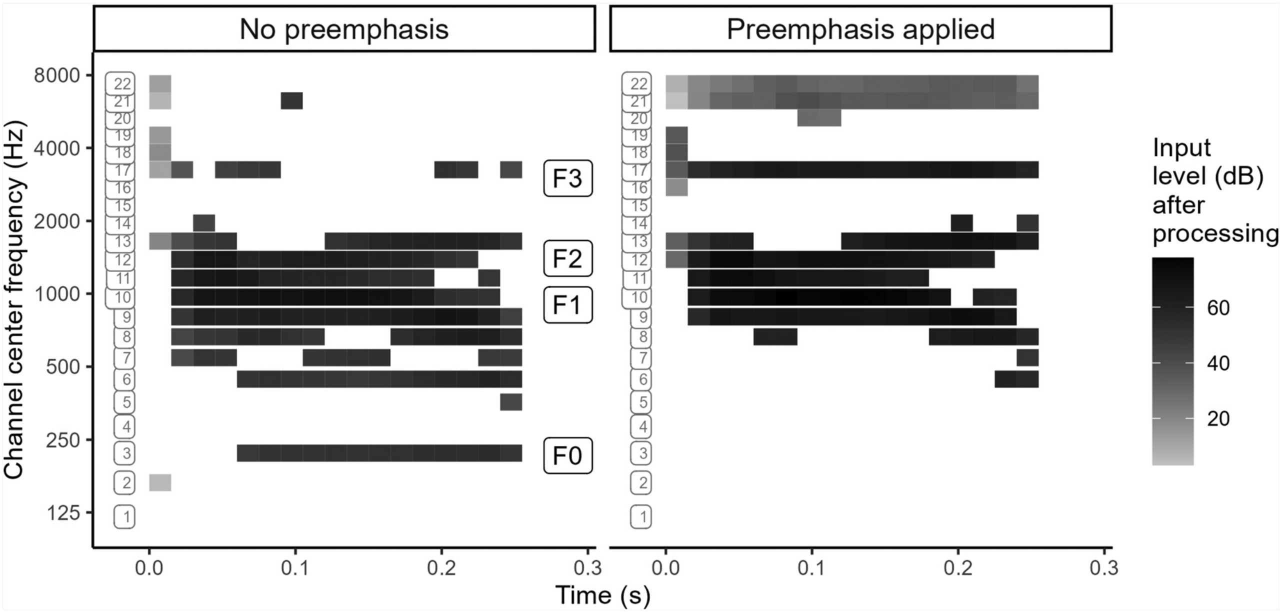


FIG. 5. A female adult's vocoded /ɑ/, processed using a peak-picking strategy that selects the top 8 of 22 channels with the most energy. Each tile represents simulated channel output discretized into successive 30-ms time bins. The shading of the tile reflects the output intensity on a dB scale. Higher frequencies are less likely to be selected during peak-picking if the signal is not pre-emphasized; lower frequencies are less likely to be selected when the signal *is* pre-emphasized; mid-frequencies are selected whether or not the signal is pre-emphasized.

the Appendix for similar illustrations at the word and sentence level).

For much of the vowel, F3 is only encoded when pre-emphasis is applied. When pre-emphasis is not applied, the first two harmonics are encoded at the expense of F3. The perceptual consequence is that the listener might not get a strong sense of the speaker's vocal tract size (partly indicated by F3) and might therefore have a more difficult time interpreting the other formants within that speaker's vowel space. Figure 6 further illustrates this same concept in the frequency domain, where the spectral shape and relevant landmarks of the underlying vowel are more visible.

IV. OVERVIEW OF VOCODER CONSTRUCTION: SYNTHESIS STAGE

A. Type of carrier

One of the most significant choices that an experimenter can make is the type of carrier used to synthesize the channels. The two most common carrier types—sinewave and noise-band—are best characterized by their spectro-temporal features. For sinewaves, this refers to faithfully replicated amplitude envelopes. For noise-bands, it refers to the ability to maintain a non-tonal quality of sound output while being able to carefully manipulate spectral shape and thus simulate spread excitation in the cochlea. Consequently, as with many other degrees of freedom in vocoder design, carrier choice should only be made after carefully considering the spectro-temporal characteristics of the outcome to be measured. Below, we review some common and less-common choices for vocoder carriers.

1. The most common carriers: Sinewave and noise-band

Sinewave carriers have a *flat* envelope, devoid of any inherent modulations that would interfere with the envelope that is intended to be imposed upon each channel. In this way, sinewave carriers are especially useful for experimenters who are interested in a listener's ability to perceive finely controlled temporal cues. However, sinewave carriers can only stimulate narrow frequency band regions so they cannot simulate the effects of wide electrical fields that characterize real CI use (there are clever ways to simulate channel interaction with sinewave carriers that will be discussed later). Noise-band vocoders, on the other hand, permit better representation and manipulation of spectral shape, and thus channel interaction, precisely because they have a wider bandwidth. However, noise carriers contain inherent random fluctuations in the envelope that can distort the intended signal envelope (Whitmal *et al.*, 2007).[5] Figure 7 illustrates these fluctuations, which are especially noticeable when the noise is filtered into a narrower band.

Carriers interact with other elements of vocoding, such as the number of channels employed and the low-pass filter (LPF) cut-off (Dorman *et al.*, 1997b; Fu *et al.*, 2004). In the absence of reliable temporal cues (low LPF cut-off) listeners may tend to rely on spectral cues, and these are represented more faithfully in the denser spectrum that noise carriers generate. Similar interactions occur when the number of channels is changed. When spectral and temporal cues are sparse, as in noise carriers with few (< 8) channels, sine carriers result in (relatively) better speech perception outcomes. However, given a sufficient number of channels, the relative benefit of sinewave over noise-band carriers disappears

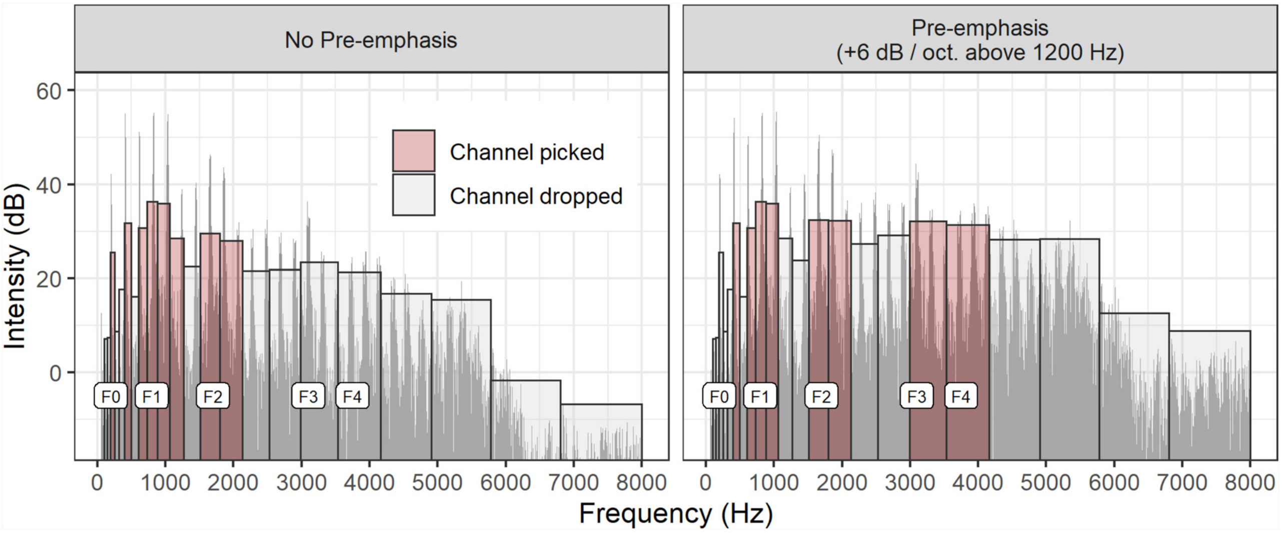


FIG. 6. (Color online) Spectra of a vowel in raw form (left) and after pre-emphasis (right). In each panel, the 8 channels (out of 22) with the highest energy are highlighted in red, reflecting *n-of-m* peak picking. The pre-emphasized spectrum contains an increase in 6 dB per octave starting at 1200 Hz, making higher-frequency energy more likely to fall into the 8 highest energy channels. The pre-emphasized vowel has its F3 and F4 reflected in the selected channels, while the non-pre-emphasized vowel neglects these channels in favor of higher-intensity low-frequency energy, such as the lower harmonics. Without peak-picking, all of the spectral landmarks would be represented at this analysis stage (including the empty gray channel bars), with only the coarse spectral shape being affected by the pre-emphasis.

because an 8- or 16-channel noise vocoder more densely samples the spectrum, giving listeners more robust spectral cues even if temporal cues are masked (Winn and O'Brien, 2022). This spectro-temporal cue trade-off explains why noise carriers benefit much more than sine carriers from increasing channel number.

2. Low-fluctuation noise carriers

In response to some of the limitations inherent to sine and noise vocoders, a number of additional carrier types have been designed and compared (Strydom and Hanekom, 2011). For example, low-fluctuation noise-band carriers, sometimes referred to as low-noise noise carriers, intentionally flatten out the problematic modulations inherent in standard noise carriers (Whitmal *et al.*, 2007). Even so, these carriers are still intermediaries between noise and sine, and thus have limitations that could interfere with an experimenter's intended design. They still cannot, for example, simulate excitation spread beyond a single auditory filter, because the concept of low-fluctuation noise is only meaningful within a single filter. If a low-fluctuation noise bandwidth extends beyond a single auditory filter, uncontrolled modulations will quickly return to the narrowband noise (Hartmann and

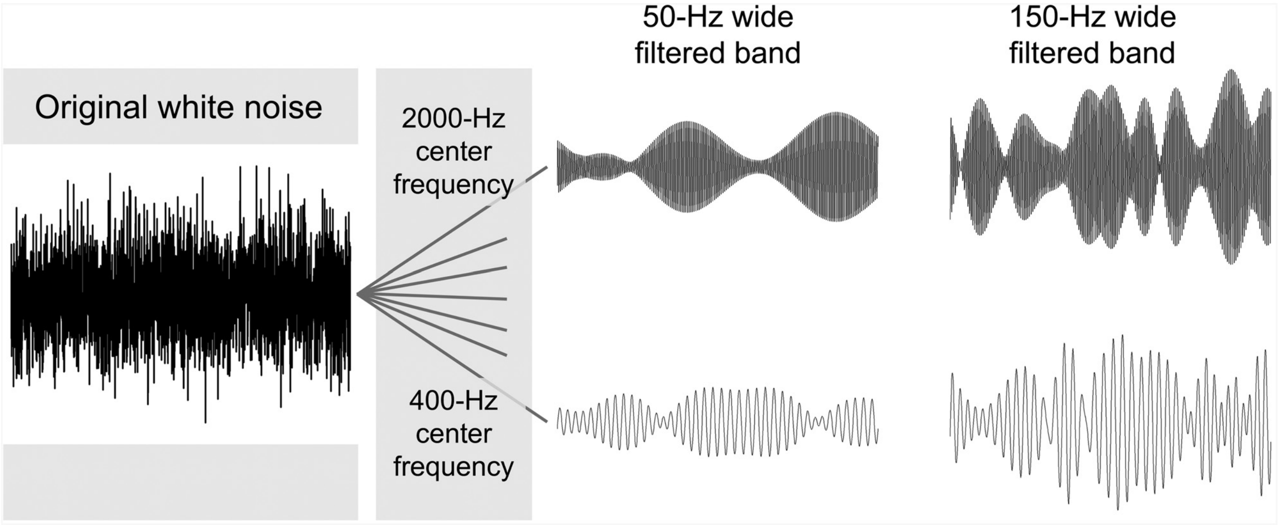


FIG. 7. White (Gaussian) noise contains random fluctuations in the amplitude envelope that emerge visually when the noise is filtered into a narrow frequency range, as for a vocoder carrier channel. In this figure, white noise is filtered into 50-Hz (center) or 150-Hz (right) spectral bands that are centered at 2000 Hz (top) or 400 Hz (bottom). The filtered bands contain random fluctuations in amplitude even before any speech-related envelope is imposed upon them.

Pumplin, 1988). Whitmal *et al.* (2007) provides an illustration of this. The authors used a constant, 100-Hz low-fluctuation noise carrier bandwidth. However, the auditory filter bandwidth below 1000 Hz can be less than 100 Hz (e.g., approximately 80-Hz bandwidth for the filter centered at 500 Hz). Consequently, the authors only truly achieved a low-fluctuation noise carrier for some higher-frequency channels. Experimenters typically try to restrict the bandwidth of a low-fluctuation carrier to a single ERB to control for envelope properties, but because phase properties are normally left unconstrained, and because ERB is a simplified heuristic rather than a biological boundary. It is likely that there is envelope fluctuation leakage into neighboring channels.

The amplitude of the residual modulations in low-fluctuation noise is much less than in random filtered noise carriers, but slightly more than in sine vocoders. The practical consequences of the band-limiting process in creating low-fluctuation noise are that (1) low-fluctuation noise carriers cannot create carriers that have contiguous spectral coverage and (2) the experimenter may believe they solved the issue of inherent carrier modulations, but if the carrier bandwidth exceeds a critical bandwidth, the benefit of a low-fluctuation noise carrier is undermined. The experimenter who absolutely prioritizes fidelity of the envelope will likely prefer sinewave carriers. However, the low-level jitter in the fine structure of low-fluctuation noise carriers can benefit experimenters who wish to simulate the lack of fine-structure synchrony between ears, as might be relevant when simulating bilateral CIs, and which would be unattainable when using sinewave carriers.

3. SPIRAL vocoder

The spectral precision of tonal/sinewave carriers can make it difficult to reflect the decay of current stimulated along the spiral ganglion. The SPIRAL vocoder was designed in response to this and other limitations of carriers previously outlined (Grange *et al.*, 2017). The objective of the SPIRAL vocoder was to modulate complex tonal carriers by *weighting* the amplitude envelopes. This technique still introduces spurious amplitude modulations in the envelope, but, in contrast to noise carriers, the fluctuations in the SPIRAL carrier are orderly rather than random. The resulting vocoder simulated both the continuous nature of the spiral ganglia and current spread without introducing fluctuations from envelope modulations [see Oxenham and Kreft (2014) and Crew *et al.* (2012) for similar approaches to envelope calculation]. The combination of a tonal carrier with weighted envelopes results in a vocoder that can simulate some aspects of the electrode-nerve interaction and isolate effects of current spread independently of electrode activation. (Vocoders typically implement continuous frequency boundaries, without gaps, so the bandwidth varies proportionally to number of channels activated; fewer channels equates to wider bandwidths making it difficult to isolate effects of current spread from electrode activation.) Indeed, the authors found better speech recognition thresholds for the SPIRAL vocoder than a similar noise vocoder [akin to that employed in Fu and Nogaki (2005)].

4. Acoustic pulse trains

Because actual CI processors use amplitude-modulated electrical pulse trains, some experimenters have employed a pulsatile carrier, in hopes of better approximating the CI listening experience in basic psychophysical experiments such as rate or binaural discrimination (Carlyon *et al.*, 2008; Faulkner *et al.*, 2000; Goupell *et al.*, 2010; Goupell *et al.*, 2013). The width and shape of the pulses in the carrier can be manipulated to further approximate CI stimulation. For example, Goupell *et al.* (2013) manipulated pulse width in the time domain under the assumption that larger widths more closely approximated CIs' electrical stimulation.

Although pulsatile carriers may appear to emulate the CI's pulsatile stimulation, they have significant limitations that prevent their widespread utility. Although electric pulses can be transmitted at a seemingly arbitrary rate, the pulse rate imposes an inescapable limitation on the frequency content of the carrier that can be transmitted acoustically. The simple rule is that one cannot represent any spectral frequency that is lower than the rate of acoustic pulses.[6] This limitation is problematic for two reasons. First, electrical pulse rates in CIs are often around 900 pulses per second (pps) or more, which when transmitted acoustically would omit any frequency lower than 900 Hz, and thereby eliminate a significant part of the useful frequency spectrum of speech. Additionally, the harmonics of such a pulsed signal would be spaced 900 Hz apart, resulting in very sparse sampling of the spectrum. Given these limitations, if critical low-frequency spectral regions (and general details of the spectral shape) are to be represented in the vocoded signal, the pulse rate must be so low (around 100 pps) as to no longer be representative of a real CI. Second, for any frequency lower than approximately 6–8 times the pulse rate, listeners will perceive pulses as pure tones at the harmonic frequencies of the pulse rate. Pulsatile vocoders are therefore more appropriate for simulations of *non*-speech phenomena such as sound localization (Goupell *et al.*, 2010).

5. Pulse-spreading harmonic complexes

Carriers can also be filtered from wideband signals such as a harmonic complex (a signal consisting of sinewaves whose frequencies are integer multiples of the fundamental). Harmonic complexes are able to be shaped into specific spectral shapes to simulate spread of activation, do not contain spurious amplitude modulations that result from using filtered noise bands, and produce an output with "pulses" (because harmonics interact to produce temporal modulations at the rate of the fundamental). However, harmonic complexes impose a very strong and specific fundamental frequency (giving a strong pitch percept, as opposed to a weak or absent pitch percept that would be more reflective of CI listening): a harmonic complex vocoded signal with an f0 of 100 Hz will contain strong harmonic pitch cues indicating 100 Hz regardless of the voice's actual f0, resulting in a spurious and heavily misleading pitch cue. Additionally, the density of the spectral sampling trades off with the rate of the envelope; a 100 Hz harmonic complex will pulse 100 times per second, and its spectral components will be 100 Hz apart. For spectral components to be sampled more densely—say, 20 Hz apart—the experimenter gains the control of a very precise spectral shape but at the cost of producing a sound with slow and noticeable 20-Hz amplitude modulations.

Pulse-spreading harmonic complex (PSHC) carriers are another response to the inherent limitations of sine and noise carriers, which also address the limitations of conventional harmonic complexes [Mesnildrey *et al.* (2016); see Gaudrain (2016)]. PSHC carriers use a harmonic complex with a low f0, rendering densely spaced harmonics and enabling greater control of each channel's spectral shape. However, PSHCs are designed to escape the constraint of modulation at the low rate of the f0. The basic premise of the PSHC is that the temporal pulses that result from combining in-phase harmonics can be *spread* through the period of the f0 by shifting groups of harmonics slightly out of phase relative to each other. The harmonics still interact with each other, producing periodic modulations in between the modulations of the original f0. Ultimately, these groups of harmonics combine to produce a wideband signal which appears to have a pulse rate much higher than what one would anticipate from the f0 comprising the harmonic partials. Importantly, the overall modulation is flatter than what would result from in-phase harmonics, but more regular than what would result from filtered noise.

Behavioral evidence presented by Mesnildrey *et al.* (2016) corroborates this intuition: listeners have better modulation detection thresholds for PSHC carriers than broadband noise carriers, suggesting the spurious amplitude modulations were minimized compared to the noise carriers (modulation detection thresholds were still better with sine carriers). Adult listeners showed better speech reception thresholds (SRTs) with PSHC carriers than noise carriers.

Given the apparent advantages of PSHC carriers, one might wonder why this carrier has not come to dominate vocoder CI simulation research. Perhaps a reason for this is the additional complexity and computational load that PSHC carriers require. Unlike more traditional carriers, PSHC carriers employ channel-specific settings (e.g., LPF cut-off) and, most importantly, channel-specific harmonic phase shifts. These shifts depend on the signal's f0 and require optimization of harmonic grouping. By optimizing each channel's modulation rate in this way, relatively flat envelopes can be generated with minimal modulation (a reduced CREST FACTOR, or peakiness of the waveform, like one would seek with a sinewave carrier) while maintaining dense (nearly continuous) spectral coverage (like one might seek with a noise carrier). Collectively, these features allow the experimenter to control spectral spread and also maintain an envelope with reduced irregular fluctuations.

Additionally, although the composite waveform from a PSHC carrier appears to have high-rate pulsatile components, the waveform crests are not necessarily perceptible as pulses, since the auditory system still passes the signal through a bandpass auditory filterbank. These limitations to passing the intended signal to the auditory system may explain why the simpler sinewave and noise carriers continue to be more widely used.

6. Effect of carrier type and envelope filtering on fidelity of envelope modulations

Collecting all of the descriptions of common carrier types described above, Fig. 8 illustrates the capacity of each carrier type to faithfully transmit a speech envelope. This illustration zooms in to a small segment of the word "sail" to see the onset of the vowel, which contains periodicity reflecting the talker's f0 of 220 Hz. When the envelope is low-pass filtered at 300 Hz (red lines), this periodicity is maintained, and reflected very faithfully in the modulated sinewave channel. The periodic modulations are present but somewhat distorted in the filtered noise carrier, and less distorted for the low-fluctuation noise carrier. For the vocoder with a harmonic complex carrier, the 100-Hz f0 of the tone complex replaces the f0 of the original speech signal. When the envelope is low-pass filtered at 50 Hz (blue lines), the periodicity of the original speech signal is effectively lost regardless of the carrier, because it is physically no longer present in the envelope that modulates the carrier. In that case, only the slowly varying cues to syllable structure remain.

7. Effect of carrier type upon speech perception outcomes

Previous comparative work has established that the benefit of one carrier type over another for speech perception outcomes depends upon the carrier's parameter settings and which acoustic cues the experimenter wishes to preserve: if the experimenter is interested in the ability to perceive and use timing cues that are finely controlled to convey periodicity or consonant bursts, sinewave carriers preserve those parameters more faithfully, whereas noise carriers could have inherent amplitude fluctuations that overpower the modulations. Sine carriers can outperform noise carriers for a variety of speech-related outcomes including intelligibility, consonant and vowel identification, and speaker and gender identification, especially if the envelope includes frequencies up to at least 160 Hz (i.e., possible f0 values for the human voice that cue gender perception) (Fu *et al.*, 2004; Gonzalez and Oliver, 2005). The effect of carrier on these outcomes is generally predictable once the experimenter has a firm understanding of how different carriers impact spectro-temporal cues.

Pitch cues that are relevant to speaker and gender identification are more robust when carriers are free from spurious modulations, as sinewave carriers are. This fact reflects the utility of spectral cues from resolved sidebands[7] [i.e., frequencies that are separated from the carrier by one or more equivalent rectangular bandwidths (Kohlrausch *et al.*,

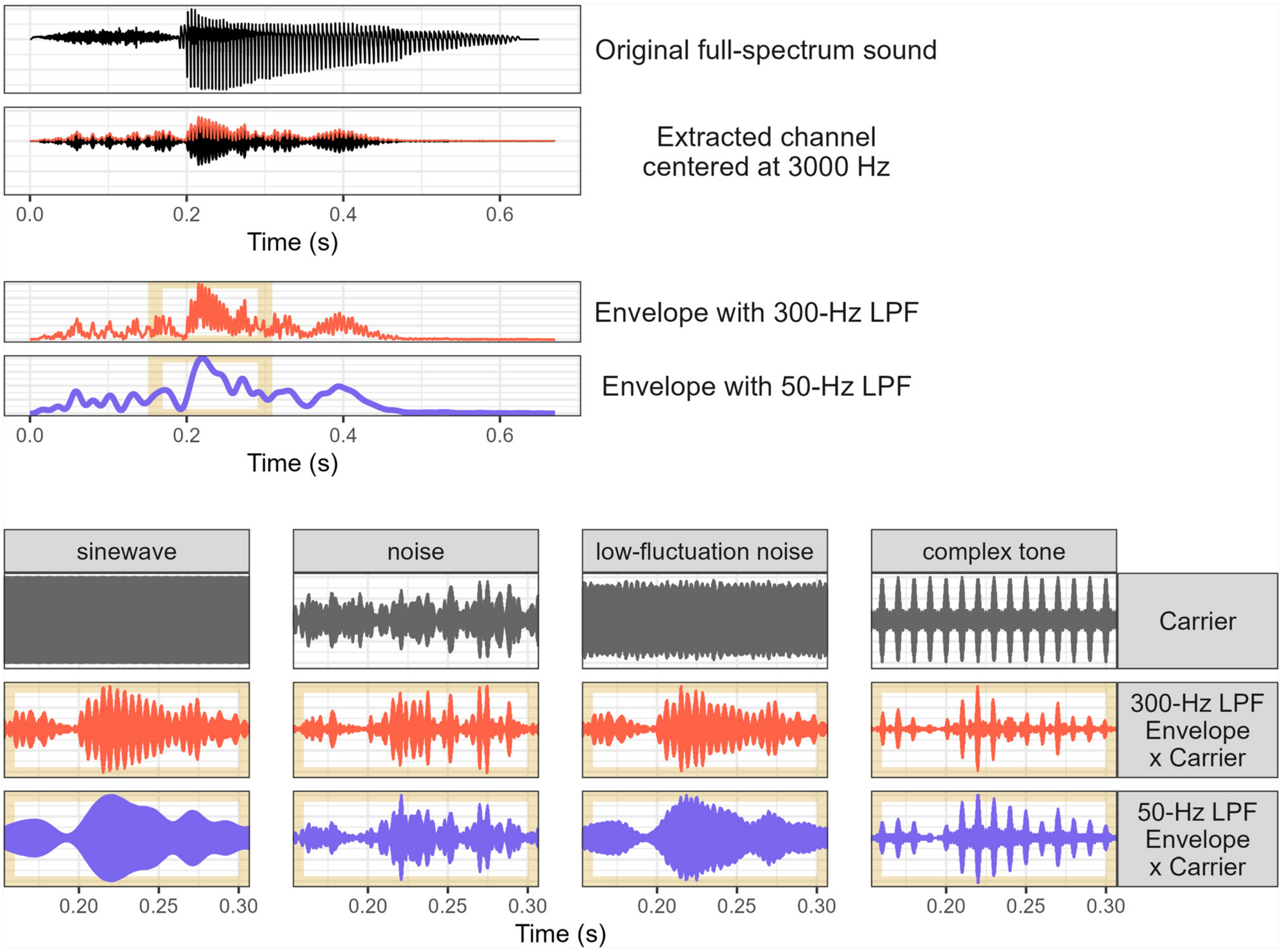


FIG. 8. (Color online) A variety of carrier channels representing a vocoded channel extracted from an utterance of the word "sail." Starting from the top the figure, the full speech waveform is filtered into a spectral band centered at 3000 Hz, from which the envelope is calculated and outlined in red. That envelope is low-pass filtered with cut-off frequencies of 300 Hz (red) or 50 Hz (blue) to either maintain or not maintain periodicity modulations, respectively. Those two filtered envelopes are used to modulate four different vocoder carrier types—sinewave, noise, low-fluctuation noise, and a tone complex with a 100-Hz f0. Each panel in the lower portion of the figure represents a small segment of time indicated by the gold box in the upper portion. The envelope modulations from the original speech sound are shown to be reflected very accurately for the sinewave carrier, with varying levels of envelope distortion imposed by the other carriers.

2000)]. In noise carriers, sidebands would compete with fluctuations in noise, and be further masked by the aperiodicity and spreading spectral energy of the noise carrier (Fogerty *et al.*, 2016; Souza and Rosen, 2009). However, even sinewave carriers will not transmit pitch cues if the f0 is higher than the envelope LPF cut-off of the vocoder (see Fig. 8).

Harmonic complex carriers are a special case, as the spurious amplitude modulations are structured rather than random, resulting in a very strong pitch percept that overrides whatever f0 was originally in the signal. That is, a harmonic complex vocoded signal with an f0 of 100 Hz will contain a strong 100-Hz pitch regardless of the voice that is being processed by the vocoder.

Acoustic cues to speech contrasts such as manner (i.e., contrasts that depend on how air flows or is obstructed in the vocal tract as in /b-w/) and voicing (i.e., contrasts that depend on whether or not vocal folds are vibrating as in /k-g/) are not immune to carrier choice, but the largest impact of carrier is upon place of articulation contrasts. For place of articulation, the superiority of sine carriers is very small for contrasts such as /k-t/ because these contrasts rely on fuller representations of the spectral envelope and sine carriers produce sparser spectral shapes (Churchill *et al.*, 2014). Noise carriers may result in better rates of fricative identification because these carriers can more faithfully mimic wideband turbulent, fricative-like noise. Nevertheless, since place of articulation is signaled by frequency contrasts, vocoded speech will likely lead to misperceptions of place of articulation regardless of the choice of carrier. If there are any effects on perception of consonant voicing, one might expect a slight bias toward hearing sinewave vocoded speech as being voiced (because it is periodic) and perceiving noise vocoded speech as being voiceless (because it is aperiodic), but outcome measures have generally focused on other issues since voicing perception is so well preserved overall.

Because the choice of carrier is so central to vocoder construction, a number of studies have tried to evaluate the

TABLE I. Effects of carrier type upon speech perception outcomes. PSHC = pulse-spreading harmonic complex; SRT = speech recognition threshold.

Author (year)	Carriers compared	Stimuli	Outcome	Result
Dorman *et al.* (1997b)	Sinewave, noise-band	CVC syllables and words	% vowels correct	No difference
		Multi-talker vowels	% vowels correct	Sinewave > Noise = Pulse train
		/ɑCɑ/ syllables	% consonants correct	No difference
		HINT sentences	% words correct	No difference
		/ɑCɑ/ syllables	% manner info transmitted	No difference
			% place info transmitted	Noise > Sinewave
			% voicing info transmitted	No difference
Faulkner *et al.* (2000)	Noise-band, pulse train, f0-dependent and f0-constant hybrid pulse-noise	/ɑCɑ/ syllables	% manner info transmitted	No difference
			% place info transmitted	High-LPF noise > f0-Dep. Hybrid
			% voicing info transmitted	f0-Dep. Hybrid = High-LPF noise = f0-Constant Hybrid > Low-LPF Noise > Pulse train
		/bVd/ words	Vowel identification	No difference
		BKB sentences	% words correct	High-LPF noise > f0-Constant Hybrid
		Standardized texts	Connected discourse tracking rate	High-LPF noise > Low-LPF noise
		Sawtooth wave glides	Pitch salience	f0-Dependent Hybrid > Noise > f0-Constant Hybrid
Fu *et al.* (2004)[a]	Sinewave, noise-band	/hVd/ words	Gender discrimination	Sinewave > Noise
Gonzalez and Oliver (2005)	Sinewave, noise-band	Sentences	Gender identification	Sinewave > Noise
			Speaker identification	Sinewave > Noise
Whitmal *et al.* (2007)	Sinewave, noise-band	/ɑCɑ/ syllables	% manner info transmitted[b]	Sinewave > Noise
			% place info transmitted	Sinewave > Noise
			% voicing info transmitted	Sinewave > Noise
	Sinewave, noise-band	Words in sentences	% words correct	Sinewave > Noise
	Sinewave, noise-band, 100-Hz noise, low-fluctuation noise	/ɑCɑ/ syllables	% consonants correct	Sinewave = Low-fluctuation noise > Noise > 100-Hz noise
Stone *et al.* (2008)	Sinewave, noise-band	IEEE sentences	% words correct	Sinewave > Noise
Souza and Rosen (2009)	Sinewave, noise-band	/ɑCɑ/ syllables	% consonants correct	Low-LPF sinewave = Low-LPF noise; High-LPF sinewave > High-LPF noise
Hervais-Adelman *et al.* (2011)	Sinewave, noise-band, pulse train	Sentences	% words correct	Sinewave > Noise = Pulse Train
Rosen *et al.* (2015)	Typical sinewave, dense sinewave, noise-band	Open-set sentences	% words correct	Dense Sinewave > Noise > Typical sinewave
Mesnildrey *et al.* (2016)	Sinewave, noise-band, PSHC	Sentences	SRT	Sinewave > PSHC > Noise

[a] Presented as pilot data in the publication.

[b] The only exception is for fricatives where noise carriers outperformed tonal. See text for detail.

effect of carrier type upon speech and pitch outcomes (see Table I for a summary of these results). The original channel vocoder proposed by Dudley (1939) was a combination of sine and noise vocoding. In that case, the carrier choice depended upon the perception of voicing in the signal. When no voicing was present, a carrier akin to noise-band ("hiss") was used; when voicing was present a carrier more akin to sinewave ("buzz") was used. The advantage of this hybrid vocoder is enhanced intelligibility of some consonants, such as fricatives, likely because noise carriers more accurately simulate the phonemes' high-frequency, noisy profiles (Dorman *et al.*, 1997b). Nevertheless, in contemporary research, combination carriers such as these are rarely implemented [but see Faulkner *et al.* (2000)].

The first direct comparison between vocoders using noise-band and sinewave carriers for speech recognition was performed by Dorman *et al.* (1997b). The authors found that carrier type only made a small difference in speech recognition performance, and only for some outcomes; sine carriers were better for multi-talker vowel identification, noise carriers were better for consonant place of articulation. For other outcomes, such as percent words correct in HINT sentences (Nilsson *et al.*, 1994) or percent vowels (real and synthesized) correctly identified, the authors found no effect of carrier.

Whitmal *et al.* (2007) likewise compared carrier type, but in more challenging listening conditions (speech spectrum noise-maskers and two-talker babble). Results of that work showed that sinewave vocoders resulted in a larger percentage of words recognized, and consonant features identified, compared to noise vocoders. Whitmal *et al.* (2007) also investigated the effect of carrier amplitude fluctuations upon speech perception by creating different types of noise carriers: contiguous wideband Gaussian noise, narrowband (100-Hz bandwidth) Gaussian noise, and narrowband low-fluctuation noise (Pumplin, 1985) (described above). Narrowband Gaussian noise carriers resulted in the worst speech perception, followed by the wideband Gaussian noise carrier—suggesting that the modulations (present in the wideband, and stronger in the narrowband) were a key factor that decreased performance. Best performance occurred for the sinewave and low-fluctuation noise carriers. These results highlight how the intrinsic modulations of each carrier affect temporal envelope cues that are important for speech perception.

An additional complicating factor relating to the comparison of carriers is the rate of envelope modulation. All of the studies mentioned in the previous paragraph used relatively high cut-off frequencies on their envelope extraction (160–400 Hz). However, as discussed in Sec. III B, LPF cutoff and carrier type interact. For example, a small number of channels with sinewaves carried at reduced envelope bandwidth (30 Hz) actually produce *lower* speech intelligibility than noise carried at the same rate (Rosen *et al.*, 2015; Souza and Rosen, 2009).

It is important to remember that although we have summarized the literature comparing carrier choice (Table I), we encourage experimenters to carefully consider their choice of carrier in the context of the actual research goals. We encourage experimenters to consider the signals, and not just compare listeners' performance. Just because one carrier results in better speech recognition or any other auditory outcome does *not* mean that the signal generated actually reflects true CI processing.

In sum, the choice of carrier depends upon the outcome being measured. If the vocoder experiment focuses on disruption of temporal fine structure, or distortions to spectral shape, a noise carrier may be most appropriate. However, it is important to remember that this carrier will sacrifice certain temporal properties of the signal because noise-bands introduce random temporal modulations. Conversely, if the experiment prioritizes temporal processing, sinewave carriers will likely be preferred, despite their sparse, unrealistically tonal quality, since they convey temporal envelope cues more faithfully.

B. Interaction between envelope filtering and channel bandwidth

There is an unavoidable interaction between temporal envelope filtering and the channels' spectral bandwidth. If a carrier sinewave with frequency X is multiplied by an envelope that contains modulations at rate Y, the spectrum of that carrier will not only include a component at frequency X, but also sideband components at frequencies $X - Y$ and $X + Y$. This will be especially noticeable in sinewave carriers which lack the random temporal modulations in a noise band that might mask the newly introduced sidebands. The sidebands can be resolved in the auditory system as cues to periodicity and other frequency modulations (Souza and Rosen, 2009).[8] Conversely, envelope modulation rate interacts less strongly with noise carriers because noise has its own (albeit random) modulation rate that could interfere with the envelope modulation rate cues (Stone *et al.*, 2008). A lower envelope cut-off frequency theoretically just limits the *rate* of modulations, but effectively also results in *shallower* modulations within the temporal envelope, which will affect the availability of various acoustic-phonetic cues in the signal (see Fig. 3 for illustrations). For example, stop sounds are signified by extremely rapid changes in the envelope, and when those modulations are filtered out, perception of consonant manner of articulation cues is accordingly weakened (Xu *et al.*, 2005b).

C. Order of operations for filtering and imposing the envelope

The amplitude envelope can be imposed on a noise carrier either before or after filtering that carrier into the desired frequency range for the channel, but the order of these operations affects the temporal and spectral fidelity of the output. If the researcher prioritizes maintaining the exact frequency of the channel bandwidth, they will want to impose the amplitude envelope on the full white noise and *then* filter the modulated noise to match the frequency band of the channel. In doing so, any modulations whose rate exceeds the spectral bandwidth will be lost.

However, if the researcher prioritizes maintaining the full range of envelope modulations, they can filter the noise to the channel's frequency range, and then impose the amplitude envelope. In this scenario, the spectral bandwidth will be expanded if the envelope contains modulations whose rate exceeds the bandwidth of the filter. For example, even if one filters a noise band to have a 100-Hz bandwidth and then imposes an envelope that contains 500-Hz modulations, the channel will now have a 1100-Hz bandwidth (the energy outside the original frequency range will reflect the strength of the faster modulations). This sideband effect is observed most clearly in sinewave carriers, where each modulation frequency is visible as an additional sideband component that is separated from the carrier frequency by ± the modulation rate. For example, if we begin with a sinewave whose center frequency is 1000 Hz and then impose an envelope that is modulating both at 5 Hz (the rate of syllable production) and also at 200 Hz, the output will have components at 800, 995, 1000, 1005, and 1200 Hz—as well as interacting sideband components at 795, 805, 1195, and 1205 Hz.

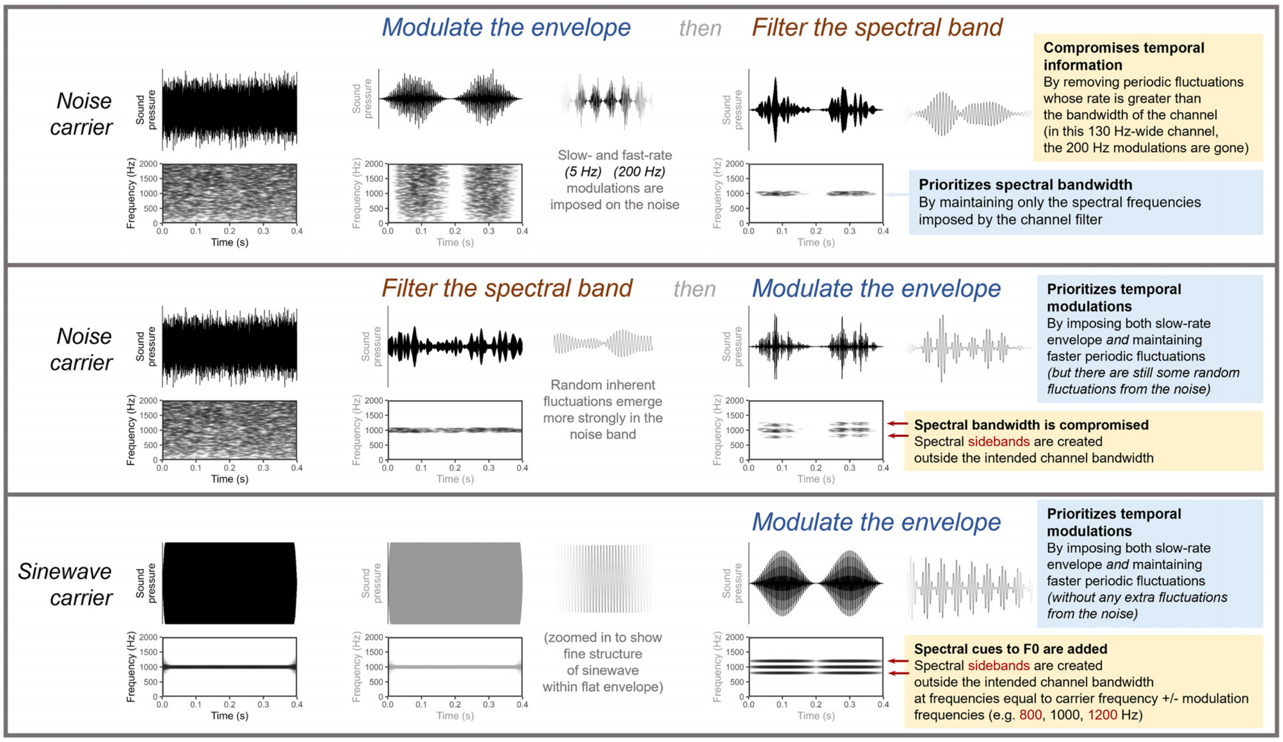
Figure 9 illustrates the impact on the order of filtering and envelope operations for filtered noise and sinewave carriers. There is no singular perfect choice that both preserves perfect spectral and temporal fidelity because the spectral shape is a reflection of the temporal modulations and vice versa. The choice should reflect the experimenter's preference to either prioritize control over the spectral shape or over the temporal envelope modulations in the carrier.

Experimenters have attempted to avoid the envelope/bandwidth trade-off by simulating channel interaction at the "front end," in the analysis phase. To do this, different degrees of channel interaction can be simulated by introducing temporal envelope information extracted from increasingly wider analysis bands and imposing them upon the envelope of the carrier for that band [cf. Crew *et al.* (2012)]. Therefore, that band will contain information from neighboring bands (hence the interaction), but the carrier can be created without any concurrent limitation of that analysis phase; see more in Sec. V B 1. As in many other elements of vocoding, simulations of current spread interact with other choices that an experimenter may make, such as the number of vocoder channels since reducing current spread allows CI users, and individuals with NH listening to vocoder CI simulations, to benefit from more channels (Bierer and Litvak, 2016; Bingabr *et al.*, 2008; Gaudrain and Başkent, 2015; Grange *et al.*, 2017).

D. Summation and normalization

The final stages in signal vocoding are to sum together the envelope-modulated carriers and set the overall intensity of the summed signal. When experimenting with vocoder parameters, researchers will benefit greatly from listening to, visualizing, and examining the vocoded stimuli to avoid unintended signal processing consequences before and after summing as the final combined signal may hide some problematic aspects that arose during vocoding such as unstable or leaky filters. Otherwise, summation is relatively straightforward.

Normalization may require more attention. One option is to ensure that the original unprocessed and new vocoded signals are equal in terms of root-mean-square (RMS) energy. However, considering that some intensity of the original signal stems from frequencies outside the range that are included in the vocoder analysis channels, a more subtle equalization would be to equate the intensity of the range



Compromises temporal information
By removing periodic fluctuations whose rate is greater than the bandwidth of the channel (in this 130 Hz-wide channel, the 200 Hz modulations are gone)

Prioritizes spectral bandwidth
By maintaining only the spectral frequencies imposed by the channel filter

Prioritizes temporal modulations
By imposing both slow-rate envelope *and* maintaining faster periodic fluctuations *(but there are still some random fluctuations from the noise)*

Spectral bandwidth is compromised
Spectral sidebands are created outside the intended channel bandwidth

Prioritizes temporal modulations
By imposing both slow-rate envelope *and* maintaining faster periodic fluctuations *(without any extra fluctuations from the noise)*

Spectral cues to F0 are added
Spectral sidebands are created outside the intended channel bandwidth at frequencies equal to carrier frequency +/- modulation frequencies (e.g. 800, 1000, 1200 Hz)

FIG. 9. (Color online) The sequence of operations (filtering and imposing the amplitude envelope) has implications for your control of the spectral bandwidth and preservation of amplitude modulations in the output. Top row: Modulating then filtering noise maintains the intended spectral bandwidth at the cost of excluding any modulations whose rate exceeds the linear bandwidth of the spectral filter. Middle row: filtering then modulating noise preserves the intended amplitude envelope fluctuations at the cost of introducing spectral sidebands that might not have been intended to be within that spectral channel. Bottom row: modulating a sinewave introduces very specific and potentially resolvable sidebands in the spectrum, but faithfully maintains the intended amplitude envelope, without any extra random fluctuations.

that is analyzed (i.e., if the vocoder spans 100–6000 Hz, equate the vocoded signal to a 100–6000 Hz band-passed version of the original signal rather than the full-spectrum signal).

The same applies for different types of vocoded signals to avoid experimental confounds. For example, in a study manipulating the boundary placement of an eight-channel vocoded signal, one condition may allot a greater number of channels to lower frequencies and another a greater number at higher frequencies. Yet given the sloping nature of the speech spectrum (−6 dB/octave energy decay), if the stimuli across the two conditions were not energy-normalized, any differences between conditions could be attributed to sound levels and not boundary placement. Consequently, researchers should (1) consider how the vocoder parameters might shape the energy profile of speech spectrum in ways that could create confounds with the unprocessed signal and (2) strategically energy-normalize across all conditions, including the unprocessed signal.

V. HOW TO CREATE A VOCODER FOR A SCIENTIFIC STUDY

There are several methods available to create vocoded stimuli. Readers familiar with computing languages such as MATLAB or PYTHON can follow the steps outlined in Secs. III and IV to implement a vocoder. Readers who are familiar with the Praat software can explore most of the parameters described in this tutorial by using the script available online at https://github.com/ListenLab/Vocoder online (Winn, 2024). This page contains a brief user guide and instructions for creating specific kinds of vocoded stimuli as well as visually inspecting specific components of a vocoder, such as the envelope processing and individual carrier channels. Users can also follow a full vocoder feature demonstration hosted on the same page. Another approach is to create vocoded stimuli using ANGELSIM (Fu, 2012), a freely available graphical user interface where users can specify many desired vocoder parameters and generate vocoded stimuli without any knowledge of the back-end computation. Although ANGELSIM does not currently offer all of the features and options described in the current paper, it can be a valuable method to generate stimuli and to quickly explore relevant parameter ranges during pilot testing.

A. Reporting vocoder design in a scientific study

Vocoder design varies widely between studies. To ensure that the design can be understood and replicated, authors should minimally specify the following parameters in their vocoder description:

(1) Type of carrier.
(2) Number of channels.
(3) Frequency range (upper and lower limit).
(4) Corner frequencies of each analysis and carrier band.
(5) Analysis and carrier filter slopes.
(6) Low-pass filter cut-off of the amplitude envelope.

Additional settings that are ideally specified but may be less essential than those in the previous list:

(1) Pre-emphasis high-pass filter cut-off.
(2) Shifting the center frequency of each analysis and carrier band.
(3) Type of filter (e.g., Butterworth).
(4) Envelope extraction method (e.g., Hilbert transform, half-wave rectification plus low-pass filtering).

VI. WHAT ASPECTS OF COCHLEAR IMPLANTATION CAN WE SIMULATE WITH A VOCODER?

A. Number of activated electrodes

The number of activated electrodes is one of the most common parameters simulated in CI vocoder research (Ananthakrishnan and Luo, 2022; Ananthakrishnan *et al.*, 2017; Başkent, 2006; Dorman *et al.*, 1997b; Eisenberg *et al.*, 2000; Friesen *et al.*, 2001; Fu *et al.*, 2004; Fu and Shannon, 1999; Gonzalez and Oliver, 2005; Shannon *et al.*, 2004, 1995; Winn *et al.*, 2015). The cardinal rule with channel number manipulations is that a larger number of channels will yield greater spectral detail and will thus usually result in better performance on most speech perception outcomes.[9] In the simulations, anywhere from 4 to 32 channels are typically used. However, the exact number will depend greatly on the outcome measure, so it is important to consider how many channels are required for performance saturation. Some outcomes see performance saturation after just 8 channels, while other outcomes require greater spectral resolution (see Table II for examples).

It is also important to consider how channel number may interact with other aspects of the signal. For manipulations that degrade the signal in other ways such as channel overlap/spectral smearing (Fu and Nogaki, 2005; Grange *et al.*, 2017) or more difficult SNRs (Başkent, 2006), a greater number of channels does not necessarily provide more benefit. The stimuli will also impact the number of channels that an experimenter may wish to use. There is a potentially greater speech recognition benefit from higher channel numbers for spectral than temporal acoustic cues (Dorman *et al.*, 1997b). In practice, this usually corresponds to a substantial effects of number-of-channels for consonant place of articulation as well as vowel contrasts, but virtually no effect for perceiving consonant voicing. Finally, the listener population will of course impact the ideal number of channels to employ: children as old as 7 years require more channels than adults to reach equivalent levels of performance (Eisenberg *et al.*, 2002; see Sec. VII B).

It may appear intuitive to simply match the number of vocoder channels to the number of electrodes in a CI. However, studies show that this is an ineffective simulation strategy because most acoustic simulations lack the important factor of channel interaction. Instead, a common practice in the literature has been to increase the number of discrete vocoder channels to approximate the performance score of better-performing CI users. Once the vocoder

TABLE II. More channels in the vocoded signal will often result in stronger perceptual and processing outcomes. However, the number of channels required for saturation varies by outcome. Note: research presented is not comprehensive. Only results from monolingual adult listeners with NH are included. When multiple SNRs were studied, performance at 0-dB SNR is reported. PRT = phoneme recognition threshold; FFR = frequency following response.

Author (year)	Num. of channels	Stimuli	Outcome	Channels required for saturation
Dorman *et al.* (1997b)	2, 3, 4, 5, 6, 7, 8, 9	Iowa vowels	% vowels correct	6
		Synthetic b/V/t words	% vowels correct	8
		Multitalker vowels	% vowels correct	8
		h/V/d words	% vowels correct	8
		a/C/a syllables	% consonants correct	6
			% place info transmitted	6
			% manner info transmitted	2
			% voicing info transmitted	3
		HINT sentences	% words correct	5
Shannon *et al.* (1998)	1, 2, 3, 4	h/V/d words	% vowels correct	Not attained
		a/C/a syllables	% consonants correct	Not attained
			% place info transmitted	Not attained
			% manner info transmitted	2
			% voicing info transmitted	3
		Sentences	% words correct	Not attained
Fu and Shannon (1999)	4, 8, 16	h/V/d words	% vowels correct	Not attained
Loizou *et al.* (1999)	2, 3, 4, 5, 6, 8, 10, 12	TIMIT sentences	% words correct	8
Friesen *et al.* (2001)	2, 4, 6, 8, 12, 16, 20	h/V/d words	PRT	16
		a/C/a words	PRT	16
		HINT sentences	PRT	12
Gonzalez and Oliver (2005)	3, 4, 5, 6, 8, 10, 12, 16	Sentences	Gender identification	10
	3, 4, 8, 16	Sentences	Speaker identification	Not attained
Xu *et al.* (2005b)	1, 2, 3, 4, 6, 8, 12, 16	h/V/d words	% vowels correct	12
		C/a/ syllables	% consonants correct	8
Başkent (2006)	2, 4, 6, 8, 10, 12, 16, 24, 40	h/V/d words	% vowels correct	8
		a/C/a syllables	% consonants correct	8
Winn *et al.* (2015)	4, 8, 16, 32	IEEE sentences	Rate and size of pupillary response in key analysis window	Not attained
Ananthakrishnan *et al.* (2017)	1, 2, 4, 8, 16, 32	/u/	FFR	8

performance matches the CI performance, the number of vocoder channels is taken to be the number of *effective spectral channels* in the CI. This is a heuristic concept that should not be taken literally; instead, it is a useful proxy that can allow experimenters to estimate effects on perception without accounting for all of the many variables inherent to the CI experience. One to two decades ago, it was thought that activating anything more than 7–8 electrodes in CI arrays did not result in measurable improvements in speech perception, recognition, or sound localization [e.g., Friesen *et al.* (2001) and Goupell *et al.* (2008)]. However, more recent work has provided some nuance on this topic. There is greater recognition that the number of functional spectral channels is limited primarily by the spread of excitation rather than the literal number of electrodes that are active.

The exact placement of the array within the cochlea has implications for the resolution of the device, owing to larger current fields generated for electrodes farther from the spiral ganglion cells. CI users with well-positioned perimodular arrays show more benefit from increasing the number of channels beyond 8 (up to 22) (Croghan *et al.*, 2017), while lateral wall arrays (which would typically yield greater channel interaction) show plateau in performance around 8 channels (Berg *et al.*, 2019, 2021). With these results in mind, the likely explanation for the pattern of NH listeners benefiting from greater number of channels is that vocoded stimuli typically preserve the independence of each spectral channel, while activation of a real electrode in a CI interacts with activation from neighboring electrodes. Consistent with this, consonant recognition thresholds plateaued after 16 channels in the noise vocoder employed by Friesen *et al.* (2001), but plateaued after just 6 electrodes among CI users in the same study. These studies collectively support the notion that spectral resolution in a CI is limited primarily by the spread of excitation rather than the literal number of electrodes that are active.

B. Channel interaction and current spread

As suggested in the previous section, channel interaction can explain a substantial amount of variability in CI users' outcomes (DeVries *et al.*, 2016; Fu and Nogaki, 2005; Henry *et al.*, 2000; Oxenham and Kreft, 2014), and is perhaps the more realistic representation of the factor that limits spectral resolution in a CI (as opposed to the *number* of channels). As such, this aspect of CI functioning is frequently simulated in vocoder research by varying the slope, or roll-off, of the filters that produce the analysis bands, or the filters that shape the carrier bands. Shallower slopes simulate more channel interaction and steeper slopes simulate less interaction [Fig. 14(B)]. It is common to see anywhere from −2 to −24 dB/octave slopes employed in simulations (Fu and Nogaki, 2005; Litvak *et al.*, 2007; Winn, 2020; Winn *et al.*, 2015). Anything steeper than −24 dB/octave is unrealistically precise for CI simulations, although it may serve other theoretical purposes such as demonstrating how current spread interacts with other CI parameters like array insertion depth, electrode activation, or interaural mismatch (Bingabr *et al.*, 2008; Cychosz *et al.*, 2023), or simply as a starting point to demonstrate an effect of increased channel interaction (Litvak *et al.*, 2007). Reducing channel interaction leads to benefit for CI users, and individuals with NH listening to vocoder CI simulations, for a variety of speech perception outcomes including consonant and vowel perception (Bierer and Litvak, 2016), speech recognition thresholds (Grange *et al.*, 2017), sentence recognition (Bingabr *et al.*, 2008; Grange *et al.*, 2017), and perceived vocal tract length discrimination (Gaudrain and Başkent, 2015).

Simulation of current spread is complicated for many reasons. First, the degree of current spread is not uniform across the electrode array (Bierer, 2007; Pfingst *et al.*, 2004). Electrical current spreads more widely at the apex of the cochlea than the base, leading to poorer resolution at lower than higher frequencies. The proximity of healthy neurons to activated electrodes likewise varies non-monotonically across the cochlea, and varies person-to-person (Bierer, 2010). Nevertheless, most simulation work in this area assumes similar degrees of overlap throughout the speech frequency range; in other words, simulations do not tend to vary the filter slope by vocoder channel (Fu and Nogaki, 2005). Additionally, the exact nature of channel interaction is different in acoustic versus electric hearing, as signal level will impact basilar membrane excitation in acoustic hearing, but there is no basilar membrane involvement in a CI.

1. An alternative method to simulate channel interaction with sinewave vocoders

Earlier, we observed how channel interaction was easy to implement via noise vocoders where carriers could be synthesized with varying spectral bandwidths. This is seemingly an impossible task when using sinewave vocoders, since sinewaves by definition have just one frequency component (setting aside for a moment the sidebands that emerge when that sinewave is modulated). However, there is strong attraction to sinewave vocoders because of their capacity to preserve fidelity of the temporal envelope. Fu and Nogaki (2005) proposed a solution to this conundrum. To simulate different degrees of spectral smearing, the authors used carrier frequencies whose *analysis* filters were manipulated to have roll-off of −24 or −6 dB/octave, creating a mixture of envelopes within each channel envelope. Different degrees of channel interaction can be simulated by introducing temporal envelope information extracted from increasingly wider analysis bands and imposing them upon the envelope of the carrier for that band. Figure 10 illustrates this concept by showing how frequency bands from a speech spectrum can be either contiguous or overlapping, and that the energy from those bands can be imposed on a band of

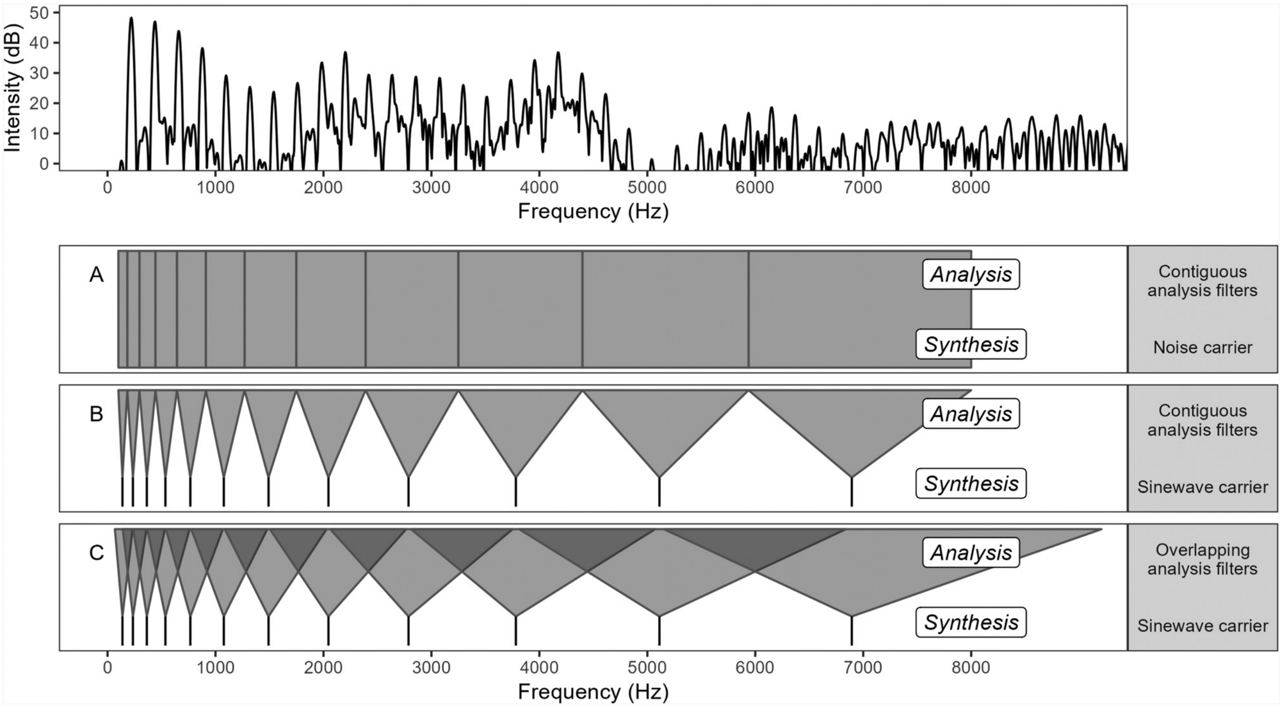


FIG. 10. Comparison of analysis filter regions that are represented by carrier channels. Starting with a spectrum of a moment of speech (top row), there are three configurations shown. Row A shows channels for a noise carrier whose frequency bandwidths are perfectly matched to the analysis filters. Row B shows those same analysis bandwidths being represented by sinewaves. Row C shows sinewave carriers that represent wide overlapping analysis bands, thereby imposing channel interaction on the output signal (each carrier represents information that is overlapping with neighboring carriers).

noise or a single-frequency sinewave. Although Fu and Nogaki originally performed this with a noise carrier, later work simulated different amounts of spectral smearing by adding variable amounts of temporal envelope information extracted across analysis bands with a sine vocoder (Crew *et al.*, 2012), thereby alleviating one of the primary limitations of sine carriers, namely, that they could not simulate channel interaction.

2. Spectral holes

Spectral holes result from areas of complete hair cell/auditory neuron loss along the cochlea (Moore, 2001; Moore and Glasberg, 2004). Spectral holes are typically accompanied by elevated CI mapping thresholds, as the stimulation will only be detected when the electrical current is increased enough to spread to neighboring neural populations that still have sensitivity. However, this spreads excitation to locations surrounding the area of loss, thus distorting the expected frequency mapping. The resulting tonotopic mismatch impacts speech recognition performance (Shannon *et al.*, 2002; Turner *et al.*, 1999; Won *et al.*, 2015). This is an element of electric hearing that researchers may wish to simulate.

There are surface-level similarities between simulating spectral holes and simulating spread of excitation. However, manipulating digital filter roll-off rate (slope) is not sufficient to simulate spectral holes because even though spectral holes can result in channel interaction, channel interaction stemming from complete neuron loss (spectral holes) is different from channel interaction stemming from irregular neuron survival adjacent to activated electrodes. In addition, spectral holes vary in size and are generally irregularly placed over the cochlea. Instead of simulating various filter slopes, as is more traditional in acoustic simulations of channel interaction, spectral holes are simulated by either completely removing channel frequencies from the final vocoded signal (*dropped* channels) and/or re-assigning frequencies to more basilar or apical channels (*redistributed* channels) (Başkent and Shannon, 2006; Shannon *et al.*, 2002). Figure 11 illustrates these two concepts. The exact simulated "location" of spectral holes along the cochlea, as well as their size (Won *et al.*, 2015), can also be manipulated on the basis of the research question. For example, DiNino *et al.* (2016) were interested in how channel interaction impacted vowel and consonant identification. In an acoustic simulation, the authors placed spectral holes in regions approximately corresponding to the first (apical), second (middle), or third (basal) formant frequencies to examine how the absence of certain phonetic cues would impact vowel confusion patterns, finding that perception of vowel formants gravitated away from the spectral holes in a systematic fashion [see also Kasturi *et al.* (2002)].

The signal processing behind spectral holes is relatively straightforward, but there are a few elements to carefully consider. In redistributed conditions, channels adjacent to spectral holes "carry" more frequencies which could result in greater amplitudes for the adjacent channels, masking surrounding frequencies. Consequently, it may be important

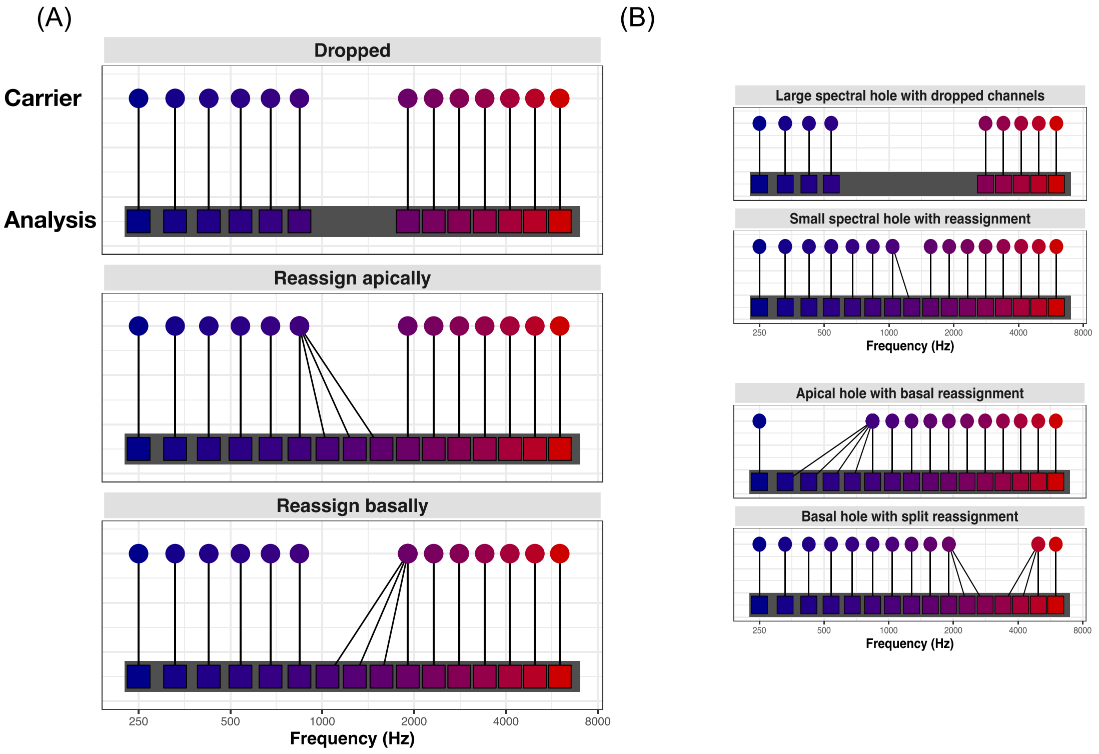


FIG. 11. (Color online) Illustration of spectral holes that can be simulated by dropping or re-assigning channel frequencies. Colored boxes represent analysis band frequencies; circles represent carrier band frequencies. (A) Frequencies can simply be dropped or reassigned. (B) Various combinations of spectral hole size and reassignment.

to normalize the intensity at the boundary channel by, for example, dividing the summed envelope by the number of carrying bands (Shannon *et al.*, 2002).

C. Insertion depth

Implant electrode arrays are typically inserted anywhere from 22 to 30 mm[10] into the cochlea (approximately 35 mm in length). This insertion depth means that implants may not stimulate the most apical places in the cochlea. Using a standard frequency-to-electrode allocation approach that distributes about 200–8000 Hz across the available electrodes, the relatively shallow insertion depth results in frequencies being shifted upward, or more toward the basal end of the cochlea (FREQUENCY-TO-PLACE MISMATCH) (Başkent and Shannon, 2005; Dorman *et al.*, 1997a; Shannon *et al.*, 1998). In practice, upward-shifted representation of low-frequency energy is the default for clinical frequency-to-electrode allocation. Therefore, although the CI listener with a shallow insertion will not experience *stimulation* of low-frequency areas of the cochlea, low-frequency energy is still represented, albeit at a cochlear position associated with acoustic transduction of higher frequencies in a NH ear.

The effect of insertion depth on outcomes is mixed and appears to depend on several co-varying factors. Recent reports in cases with fully inserted arrays placed at the lateral wall in adult CI users at least 12 months post-operation suggest that deeper insertion depths may slightly improve patient outcomes (Canfarotta *et al.*, 2022), particularly after time to adapt to the frequency-to-place mismatch. These results in CI users can be contrasted with large decrements in speech intelligibility in acute frequency-to-place mismatch experiments with vocoders and the subsequent rapid but incomplete recovery with feedback and training (Rosen *et al.*, 1999; Waked *et al.*, 2017).

Consider the vowel /u/, which is characterized by peaks in the spectrum around 500 and 1200 Hz, corresponding to the two lowest-frequency vocal tract resonances for an adult woman. An implant inserted 28 or 29 mm into the cochlea will appropriately stimulate sites along the cochlea that roughly correspond to these frequencies, but an array inserted 22 mm might stimulate parts of the auditory system that correspond to frequencies 900 and 2000 Hz (the exact values depend upon electrode array configuration). Although vowel identification can be robust to some frequency shifts because formant frequency scaling is *relative* and not absolute, the mismatch between the *expected* and actual spectral peaks can be a barrier during speech recognition, especially in the context of the speaker's f0, perceived height, etc. (Barreda and Nearey, 2012). There is likewise evidence from vocoder CI simulations that shallower array insertions systematically bias listeners to perceive higher-frequency phonemes (e.g., /s/ instead of /ʃ/), and that some listeners are less capable of recalibrating their perception to accommodate that frequency shift (Smith and Winn, 2021). In theory, a clinician could balance the choice of either shifting the frequency information upwards (to preserve representation of low-frequency information), or to drop the low-frequency information from the analysis phase altogether by reassigning the most-apical electrodes to carry higher frequencies (to better preserve tonotopic match). Difficulty with frequency-place mismatch invites ideas about how to balance the risk of tonotopic mismatch with the risk of dropping the frequencies that would otherwise be shifted upward [see Fitzgerald *et al.* (2013) for experimental exploration of the individual differences in preference].

Frequency-to-place mismatch appears to be most detrimental for post-lingually deafened CI users who established a frequency-to-place function based on typical, acoustic input before deafness onset (Canfarotta *et al.*, 2020). For these users, expectations for speech sounds must then be recalibrated and "shifted" once the CI is activated. However, even young pediatric CI recipients (<4 years) often receive one or both implants after years of progressive hearing loss (Warner-Czyz *et al.*, 2022), so the concerns over frequency mismatch can extend to children as well. In the case of children, the problems of frequency-to-place mismatch may be even more severe because those children not only need to establish a new frequency-to-place function to understand speech, but are also in the process of learning the language skills that are critical for supporting speech recognition from a degraded input (Grieco-Calub *et al.*, 2017).

To shift a frequency by a specific cochlear distance along the length of basilar membrane, the experimenter will want to first convert the frequency into a cochlear position, then shift that position in mm, then finally convert that position back into a frequency. For example, if one has a sinewave carrier of 1000 Hz and wanted to shift it by 2 mm along the basilar membrane, we would perform the following computation using the variables A = 165.4, α = 2.1, length = 35, and k = 0.88 from Greenwood (1990).

First, calculate the characteristic position in the cochlea for the frequency 1000 Hz:

$$\begin{aligned} frequency_{original} &= 1000\,\text{Hz},\\ position_{original} &= \log_{10}((1000/A)+k)*length/\alpha,\\ position_{original} &= 14.008\,\text{mm}. \end{aligned}$$

Then, the shift that position by 2 mm

$$position_{shifted} = 14.008 + 2\,\text{mm}.$$

Finally, calculate the characteristic frequency for that shifted cochlear position

$$\begin{aligned} frequency_{shifted} &= A*(10^{\alpha*(14.008+2)/length} - k),\\ frequency_{shifted} &= 1364.5\,\text{Hz}. \end{aligned}$$

Although it is computationally straightforward to shift output frequencies by a fixed distance along the basilar membrane, there are some challenges to simulating frequencies that correspond to an implanted CI electrode array. Devices differ in array length and spacing between electrodes, with consequences for frequency-to-place mapping. Even individuals with the same device vary in device insertion depth. Real devices have non-uniform spacing between electrodes and often a mismatch between each channel's frequency analysis range and the corresponding frequency of the auditory nerve that is stimulated (Landsberger *et al.*, 2015). These two factors are usually discarded in vocoder studies in favor of simplicity, but some studies have been designed to directly address the issue of frequency shifting and warping (Rosen *et al.*, 1999; Smith and Winn, 2021). Therefore, although the frequency analysis parameters are known for each manufacturer, and we have estimates of *average* angular insertion depth for various arrays (Landsberger *et al.*, 2015),[11] simulation of electrode placement is fraught with complication so experimenters often simplify. For example, frequency mismatch in a real CI is greater at the apex than the base (Landsberger *et al.*, 2015), but many simulations of insertion depth assume uniform frequency mismatch over the cochlea (Rosen *et al.*, 1999; Smith and Winn, 2021). Additionally, CIs stimulate the spiral ganglion, but acoustic simulations of frequency shifting typically operate using the Greenwood (1990) map of the basilar membrane which has different frequency spacing (Dillon *et al.*, 2021).

To determine the desired center frequencies of the analysis and carrier bands, the experimenter must choose the spacing between center frequencies. This decision could depend on if the experimenter wishes to emulate a CI company's processing strategy. For example, both MED-EL and Advanced Bionics employ log spacing between frequencies (akin to scaling using the Greenwood function), and, as previously mentioned, this is the method most commonly employed in vocoder research. Alternatively, a user may wish to emulate Cochlear's alternative processing strategy which employs a hybrid spacing: lower frequencies are linearly spaced (similar to the Bark scale) and higher frequencies are log spaced. Linear spacing should, in theory, improve resolution for low-frequency cues such as the first vowel formant. There is some evidence from CI users themselves for this: Loizou (2006) points out that vowel recognition scores among CI users are worst in a Mel-frequency spacing condition, and attributes it to the fact that Mel spacing only permits four channels within the 0 to 1000 Hz (i.e., F1, or first formant) range. Other works, however, have found more limited effects of frequency spacing upon speech perception outcomes (Fourakis *et al.*, 2004).

A number of experimental conditions have been designed to explore frequency-to-place mismatch: *unshifted*—where the center frequencies of analysis and carrier bands are tonotopically matched, as in "typical" vocoding, *shifted up* (e.g., 1, 2, and 3 mm)—where the carrier frequencies are higher than the analysis frequencies to simulate shallower insertions, and *compressed*—where a wide range of input frequencies are represented by a smaller range of carrier frequencies, covering less of the frequency spectrum, as often occurs when the frequency-to-electrode table is re-programmed (see Fig. 12 for illustration and additional examples). Several studies have been published that incorporate these conditions [e.g., Baskent and Shannon (2003), Başkent and Shannon (2004), Fu *et al.* (2004), and Goupell *et al.* (2008)]. Additional experimental manipulations could include the filterbank's span, to simulate how the CI covers the frequency spectrum (Fitzgerald *et al.*, 2013), or uniform versus non-uniform frequency mismatches over the cochlea (Li and Fu, 2010). An example experimental design manipulating frequency coverage and tonotopic mismatch, and their combined impact upon a speech stimulus, is outlined in Fig. 13. The reader is cautioned to interpret these diagrams as characterizations of acoustic simulations rather than descriptions of real frequency warping in a CI, which is both highly variable and also typically expressed along the dimension of insertion *angle* rather than insertion depth (Canfarotta *et al.*, 2020).

All vocoder CI simulations are limited by the fact that NH listeners do not have the extended experience of listening through a CI, though they can adapt (Davis *et al.*, 2005). With time, especially during the first 6 to 9 months, CI users adapt to the novel sound of their devices to reach a performance plateau (Wilson and Dorman, 2008)—a phenomenon that is nearly impossible to replicate via short-term studies in the lab among listeners with NH. However, the issue of listening experience is especially noteworthy for tonotopic mismatch simulations because, with time, actual CI users learn

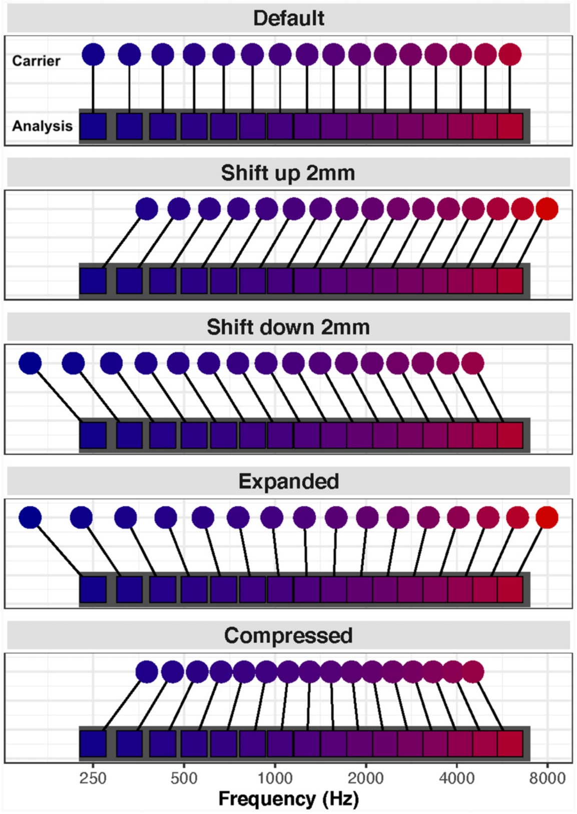


FIG. 12. (Color online) Various methods for shifting vocoder channels' center frequencies. Colored boxes represent analysis band frequencies; circles represent carrier band frequencies. Each plot represents a different manipulation of analysis to carrier band frequency. The default condition simulates no analysis to carrier shifting, or ideal frequency-to-place correspondence. See text for detail.

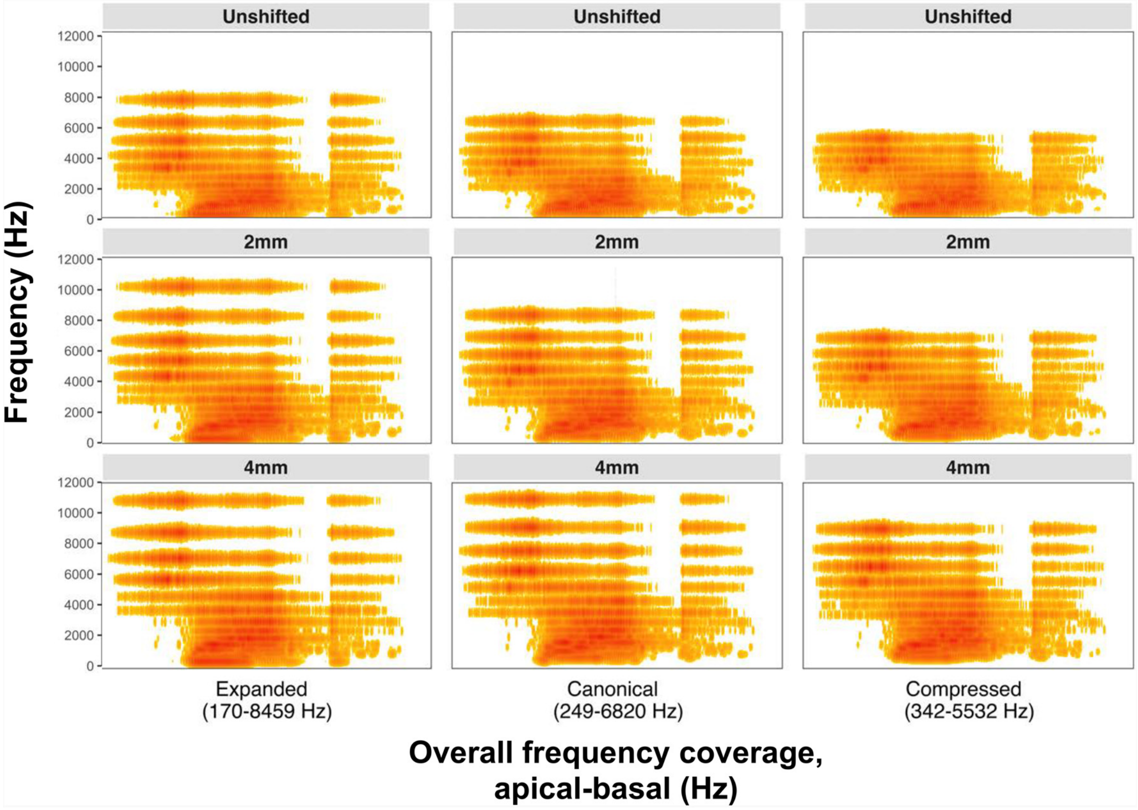


FIG. 13. (Color online) Example vocoder experimental design to simulate two parameters of tonotopic mismatch upon spectral resolution of the word "shack." Each panel shows a spectrogram where the *x*-axis is time. Plots along the horizontal axis demonstrate effects of different filterbank spans [simulating overall coverage of the frequency spectrum; frequency settings taken from Fitzgerald *et al.* (2013)]. Plots along the vertical axis demonstrate effects of different analysis to carrier band mismatch (simulating different insertion depths and basalward shifting). Greater basalward shift in "shack" biases listeners to misperceive sounds as higher frequency, possibly resulting in perception of the word "set" (/ʃ/ to [s], /ae/ to [ɛ], and /k/ to [t]).

to normalize for basalward shifts in a way that is reminiscent of how listeners with NH normalize to a new talker's voice via vocal tract length normalization and formant frequency scaling. For example, CI users typically comment that incoming speech sounds "high-pitched" or "tinny" at first activation [see Dorman *et al.* (1997a) for similar discussion], but that sensation is temporary and can disappear after even a few days of device use without adjusting programming strategies. Although Fu *et al.* (2002) showed that CI users do not *completely* adjust to tonotopic mismatches (2–4 mm) within a three-month window, the listeners' consonant, word, and sentence recognition scores improved over that time suggesting that CI users adapt to even relatively large amounts of frequency-to-place mismatches with increased listening experience. Some listeners are also better than others at recalibrating to frequency shifts (Smith and Winn, 2021; Waked *et al.*, 2017), with no clear explanation at this time for how to tell whether an individual will find it easy or difficult.

Another issue of concern for vocoder studies on tonotopic mismatch is the shifting of upper channels into a very high-frequency region, creating a potential confound of high-frequency audibility, especially for older listeners. The most common manipulation—a basalward shift of 2 to 6 mm—might shift some frequencies into an inaudible range. Given the prevalence of age-related, high-frequency hearing loss, one must carefully consider the carrier frequencies, the audibility of the carriers, and the high-frequency speech information, and check these elements against the hearing thresholds of the participants in the perception study. For example, choosing a maximum frequency no higher than 4 kHz for an unshifted condition could ensure that no stimulus in a shifted condition rises above 10 kHz, or a similarly audible frequency range. This constraint would be unusual for a study that is not interested in spectral shifting but could be a useful precaution against the audibility concern. Additional solutions to the age confound for tonotopic mismatch studies are, of course, to age-match the CI users and listeners with NH, or to measure audibility for each stimulus component at the actual presentation level in the experiment. Alternatively, one could apply a correction factor to amplify the higher frequencies to ensure audibility (Waked *et al.*, 2017). Overall, the best course is to ensure that stimuli do not extend into a range that is inaudible for the listener group.

D. Barriers to acoustic CI simulations

Most simulations of CI processing suffer from three major challenges that vocoding cannot overcome. First, because most CI users rely entirely on information conveyed by brief electrical pulses, it would be ideal to simulate CI listening in listeners with NH using stimuli consisting of acoustic pulse trains. However, it is exceedingly difficult (usually impossible) to generate acoustic stimuli that produce a click-like response and simultaneously activate a specific array of auditory nerve fibers because that click will be filtered by the mechanical tonotopy of the cochlea rather than being delivered to a specific cochlear region directly via electrode stimulation. Because of the duration-bandwidth trade-off, it is not possible to generate an acoustic pulse that is both brief enough not to temporally overlap with a subsequent pulse while also maintaining the desired bandwidth (Goupell *et al.*, 2010).

Second, the CI users with the strongest perceptual outcomes can resolve unilaterally presented rates up to 800 pps, but the majority can only resolve rates up to 300 pps (Kong and Carlyon, 2010). However, for listeners with NH, auditory nerve fibers phase-lock to resolved frequency components up to at least 2 kHz (perhaps even 4–10 kHz Verschooten *et al.*, 2019), providing information in the neural code that is not available to most CI users. Although it is unclear if NH listeners process temporal fine structure for cues besides binaural hearing, the underlying differences in neural encoding of temporal fine structure make it challenging to convey the same type of information to listeners with CIs and NH, and thus difficult to compare across the groups.

Finally, another barrier to realistic simulation is the complexity of frequency-to-place mismatch (Landsberger *et al.*, 2015) and the complication of basilar membrane filtering. As mentioned elsewhere in this paper, simulating exact electrode position is not straightforward. Additionally, the rate of electrical stimulation in a CI can be fixed at a specific rate at any part of the cochlea, but when an acoustic signal is filtered by the basilar membrane (as for a vocoded signal), the rate of temporal fine structure will necessarily result in the stimulation in the basilar membrane corresponding to that frequency rather than any other intended place.

These confounds pose significant barriers for *accurate* simulation of CI signals with vocoding. Nonetheless, a theoretical understanding of the perceptual cues provided by vocoded stimuli can guide future development of CI processors even when not all cues provided in the simulations can be practically implemented in a CI device. Furthermore, as outlined in Sec. VI, acoustic CI simulations allow researchers to fully isolate the effects of individual CI parameters, such as degree of channel interaction or insertion depth, that are confounded within individual CI users and typically not under experimental control in that population. Simulations also allow researchers to evaluate CI parameters without changing CI users' speech processing strategies, an often undesirable experimental manipulation since CI users require time to adapt to new strategies which may affect their listening experiences.

VII. VOCODERS IN CI RESEARCH ACROSS THE LIFESPAN

A. Age- and performance-related confounds in CI-NH comparisons

Studies employing vocoded stimuli often compare the performance of NH listeners to CI users. While there are numerous differences between how individuals with CIs process electric signals and individuals with NH process acoustic vocoded signals, one difference between the populations stands out conspicuously: age. In many studies, adult CI listeners are middle-aged or older (>50 years), while NH control vocoder groups in the same experiments tend to be younger adults (18–25 years). This creates a potential confound in experimental design because the ability to perceive vocoded stimuli decreases with age (Jaekel *et al.*, 2018; Schvartz *et al.*, 2008; Sheldon *et al.*, 2008; Tinnemore *et al.*, 2022). Some studies have suggested that the age effect could be a result of small differences in hearing thresholds (Shader *et al.*, 2020), but there are alternative explanations. Older NH listeners seem to have particular difficulties with temporal cue processing (Goupell *et al.*, 2017), which is especially relevant for vocoder research because vocoding typically diminishes spectral cues, relegating the listener to rely more heavily on temporal cues. Recent studies have used one-to-one age-matching between CI to NH listeners to attempt to address this confound (O'Neill *et al.*, 2021), with some studies finding reduced differences between testing groups when controlling for multiple potential confounds (Bhargava *et al.*, 2016; Tinnemore *et al.*, 2020; Waddington *et al.*, 2020).

B. Vocoders in developmental research

There are numerous scientific questions that can be addressed using vocoders in developmental research, and yet few vocoder studies have included children of any age. Eisenberg *et al.* (2000) found that children aged 10–12 years required 6 channels to recognize phonemes compared to adults who needed just 4. In the same study, 5–7-year-olds' performance on sentence and phoneme recognition did not asymptote before 32 channels, suggesting that they would continue to benefit from greater spectral resolution, whereas performance in adults frequently saturates at lower resolutions. Younger children also showed larger between-subject variability, suggesting that other cognitive-developmental factors, such as working memory, may have impacted their performance. Overall, this seminal work showed that when speech conveys primarily temporal cues, 10–12-year-olds, but not 5–7-year-olds, can recognize it on par with adults.

Dorman *et al.* (1998) similarly concluded that younger children (3–5-year-olds) required a greater number of channels than adults to ensure word recognition: adult performance asymptoted around 10 channels, but children's performance continued to improve between 12 and 20 channels. The authors additionally found interactions with word difficulty: *easy* words (high-frequency, low phonological neighborhood density) required fewer vocoder channels for recognition than *hard* words (low-frequency, high neighborhood density) [see Dorman *et al.* (1998), Eisenberg *et al.* (2002), and Roman *et al.* (2017) for similar discrepancies between "easy" and "hard" word recognition between 5 to 14-year-olds listening to 4-channel vocoded stimuli]. Children's need for increased spectral resolution is also relevant for individual vowel sounds as well, although the patterns of phoneme confusion are similar to those of adults (Jahn *et al.*, 2019). Children 8–10 years of age are affected by frequency mismatch like adults, including the same effect of improved performance across testing sessions. Waked *et al.* (2017) evaluated how different degrees of mismatch (0 [no mismatch], 3, and 6 mm) affected word recognition in adults and 8–10-year-olds. There were no differences by age for either shifted condition (3 or 6 mm), although the adults outperformed the children in the unshifted (0 mm) condition. Furthermore, since effects of frequency-to-place mismatch are especially susceptible to training and adaptation (Sec. VI C), the authors evaluated if the adults' and children's performance improved differently as a function of block number; no such differences by age emerged suggesting that 8–10-year-olds and adults adapt to spectral mismatch at similar rates. Thus, unlike the relatively more established literature evaluating impacts of spectral degradation via channel number manipulation, 8- to 10-year-olds and adults in Waked *et al.* (2017) appeared equally sensitive to spectral mismatch. Clearly, much more work is needed in this area to evaluate additional outcomes (e.g., phoneme versus word recognition) and ages (e.g., early childhood or preschool-hood).

Other vocoding work with children has extended beyond spectral degradation to manipulate channel configuration, namely, designing vocoder channels that do or do not preserve formant structure (Nittrouer *et al.*, 2014a). Here, the assumption is that earlier in life, CI listeners might rely relatively more on formant structure cues to process speech and language, and that this could shift with age and experience, just as phonetic cue-weighting strategies change markedly through early adolescence in children with NH (Hazan and Barrett, 2000). Two 5-channel vocoder conditions were created and presented to children aged 5, 7, and adults: a *standard* Greenwood-spacing channel configuration, and a *speech-preserving* configuration that maximized frequency resolution in the range of the first two formants and compromised resolution in the higher frequency range (cut-off frequencies of 550, 936, 1528, and 2440 Hz). Listeners had higher word and sentence recognition scores in the speech-preserving vocoder condition, approximating speech recognition benefits between 4- and 6-channel vocoded stimuli seen elsewhere (Eisenberg *et al.*, 2000). However, there were no differences between vocoder conditions by age. Thus, all CI recipients' speech recognition could benefit from speech-specific processing strategies (Nittrouer *et al.*, 2014a).

A handful of simulation studies with children have used vocoders to examine how access to spectral and temporal components of the signal that are typically not present in the CI signal might improve certain aspects of speech recognition and even sentence processing (Martin *et al.*, 2022). For example, voice emotion recognition is typically compromised among children with CIs owing to limited access to low-frequency spectro-temporal information (generally cued by frequencies < 400 Hz). Yet school-aged children with NH have better voice emotion recognition in 16- than 8-channel vocoded stimuli (Tinnemore *et al.*, 2018), suggesting that improvements in spectral resolution could supplement missing spectro-temporal information necessary for emotion recognition [see also Chatterjee *et al.* (2015)]. Elsewhere, Nittrouer *et al.* (2014b) created 4-channel noise-vocoded stimuli *with* and *without* low-frequency spectral cues (< 250 Hz)—again, cues that are systematically absent in the CI's signal. Results showed that children aged 5 and 7 years, as well as adults, similarly benefited from low-frequency spectral information.

A limitation of developmental vocoder research has been the age ranges studied: most work has studied mid- and late-childhood (> 5 years). Yet if channel vocoders are meant to approximate the CI listening experience, this work must extend to the ages most typical of cochlear implantation: toddler- and preschool-hood. On the basis of the limited work in this area, we know that 27-month-olds *can* recognize spectrally degraded speech (Newman and Chatterjee, 2013). Specifically, in word recognition tasks, toddlers' performance asymptotes after 8 noise-vocoded channels, with variable recognition at 4 channels, and word recognition failure at 2 channels. Also, 2- to 3-year-olds' word recognition is sensitive to spectral degradation (4- versus 8-channel vocoding) (Newman *et al.*, 2015; Nittrouer and Lowenstein, 2010). There remains a wealth of knowledge yet to be discovered regarding children's ability to perceive speech with the various distortions that are imposed by vocoders. Considering the varying rates at which aspects of auditory perception mature in childhood (Litovsky, 2015), vocoders could be a useful tool for controlling basic acoustic properties in speech, in ways that have this far remained exclusively in the realm on non-speech psychoacoustics.

One interesting line of work has examined which cues from the ambient speech stream might aid in language development. Nittrouer *et al.* (2009) found that children had higher word recognition scores when *spectral* components of the speech signal were maintained (sinewaves simulating the three lowest formant frequencies) than when *temporal* components were maintained (4- and 8-channel vocoding) [but see Newman *et al.* (2015) who found the opposite among 27-month-olds]. The 7-year-olds also, unsurprisingly, recognized fewer words in 4- than 8-channel vocoded sentences, and performed worse than adults in both vocoder conditions. Nevertheless, the study's intent was to compare how children processed temporal versus spectral degradation. Given their results, the authors proposed that children's sensitivity to spectral over amplitude structure in the speech stream might allow infants to discover phonetic units during the first year of life and eventually parse linguistic units such as words and syntactic constituents [cued by prosodic structure (Morgan and Demuth, 2014)] in early childhood [see also Nittrouer and Lowenstein (2010) who found that children relied more on spectral than temporal components of the speech signal, but that this difference decreased with age in 3-, 5-, and 7-year-olds].

Because pre-lingual CI recipients must not only process speech, but also learn language through a CI, a growing body of research has examined the impact of spectral degradation upon language learning tasks. Newman *et al.* (2020) examined the impact of noise vocoder channel number (8 versus 16) upon 34-month-olds' FAST-MAPPING skills, or a child's ability to match a word with a novel referent in their environment. In a preferential looking paradigm, the children were initially taught to associate two novel words with unknown objects and then subsequently tested on the word-object pairing. The children successfully established word-object pairing in the 16-, but not the 8-channel condition, suggesting that spectral degradation disrupts fast-mapping—one of the foremost skills required for early word learning and vocabulary development.

Again, studies like Newman *et al.* (2020) are critical extensions of earlier work examining speech recognition because there are reasons to believe that signal degradation does *systematically* impact trajectories of language learning. For example, a commonly accepted developmental trajectory among children with NH is that those who process incoming words faster develop larger receptive vocabularies, setting the stage for phonological awareness and early literacy. If children who process degraded speech through a CI instead learn to employ different processing strategies during word comprehension, as work with older children with CIs suggests (Blomquist *et al.*, 2021; Klein *et al.*, 2023), then this entire model of phonological developmental needs to be reconsidered for children with CIs. This fact is one reason why work using channel vocoders to understand not simply children's speech recognition, but their speech-language development, will be critical going forward.

Overall, the developmental vocoding literature is sparse, and biased towards middle and late childhood (8–17 years). However, from this work several generalizations can be made:

- Compared to older children and adults, younger children require greater spectral resolution to recognize speech (Eisenberg *et al.*, 2000).
- Children as young as 27 months can recognize spectrally degraded speech (Newman and Chatterjee, 2013; Newman *et al.*, 2015).
- Vocoders can be employed to understand children's speech recognition *and* language learning (Newman *et al.*, 2020; Nittrouer *et al.*, 2009).

Furthermore, we know that if certain modifications are taken into account [e.g., increasing the number of channels for younger children (Newman *et al.*, 2015)], and experimental methods appropriate for early childhood are employed [e.g., preferential looking paradigms (Golinkoff *et al.*, 2013)], that there is no *a priori* reason not to employ vocoders to study speech and hearing development even among toddlers and preschoolers.

VIII. CONCLUSION

Employing channel vocoders for CI research has resulted in increased understanding of the mechanisms of recognizing, processing, and learning from auditory signals that are degraded in specific ways. Several aspects of CIs can be simulated using vocoders, enabling systematic exploration of factors that are normally confounded within patients who vary in numerous meaningful ways. However, vocoder experiments do not automatically offer insight in the experience of listening through a CI, so vocoders are not comprehensive "CI simulations." This paper has encouraged understanding of some of the most common signal processing choices that underlie vocoders, so that experimenters feel more empowered to carefully consider their desired outcome measure—be it speech recognition, sound localization, or word learning—and manipulate the vocoder signal accordingly. Experimenters are encouraged to modify, listen to, visualize, and test with a variety of vocoder settings that could shed light on a wide range of interesting experimental questions.

ACKNOWLEDGMENTS

This work was supported by National Institute on Deafness and Other Communication Disorders Grants Nos. F32DC019539 (M.C.), R01DC017114 (M.B.W.), and R01DC014948 (M.J.G.).

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts of interest to report.

APPENDIX

1. Technical details: Filter slopes

The shape of each frequency filter has implications for the output of the vocoder. Although one can conceptually consider each filter independently, the shapes of adjacent filters can be made to overlap, resulting in interaction between channels. This type of channel interaction can be used to simulate the common situation of electrode activation overlap in real CIs.

Digital filters come in many forms, but are most commonly characterized by the (1) flatness of the passband, (2) attenuation of the stopband, and (3) steepness/attenuation of the cut-off region slope, or FILTER ORDER, from the passband to the stopband (see Fig. 14 for illustration of these concepts).[12] The faster/steeper the attenuation rate, the higher the filter order. An ideal digital filter has a flat passband, completely attenuated stopband, and infinitely steep filter slope—these characteristics would ensure that the filter faithfully contains only the energy between two frequencies. However, there are limitations to digital filter design that prevent this ideal. Attempting to impose perfectly rectangular filters will result in distortions to the signal. For common Butterworth filters, there is a trade-off between flatness in the passband and attenuation slope in the cut-off: minimizing spectral ripples in the passband will result in slower attenuation rates, and thus require a higher filter order to

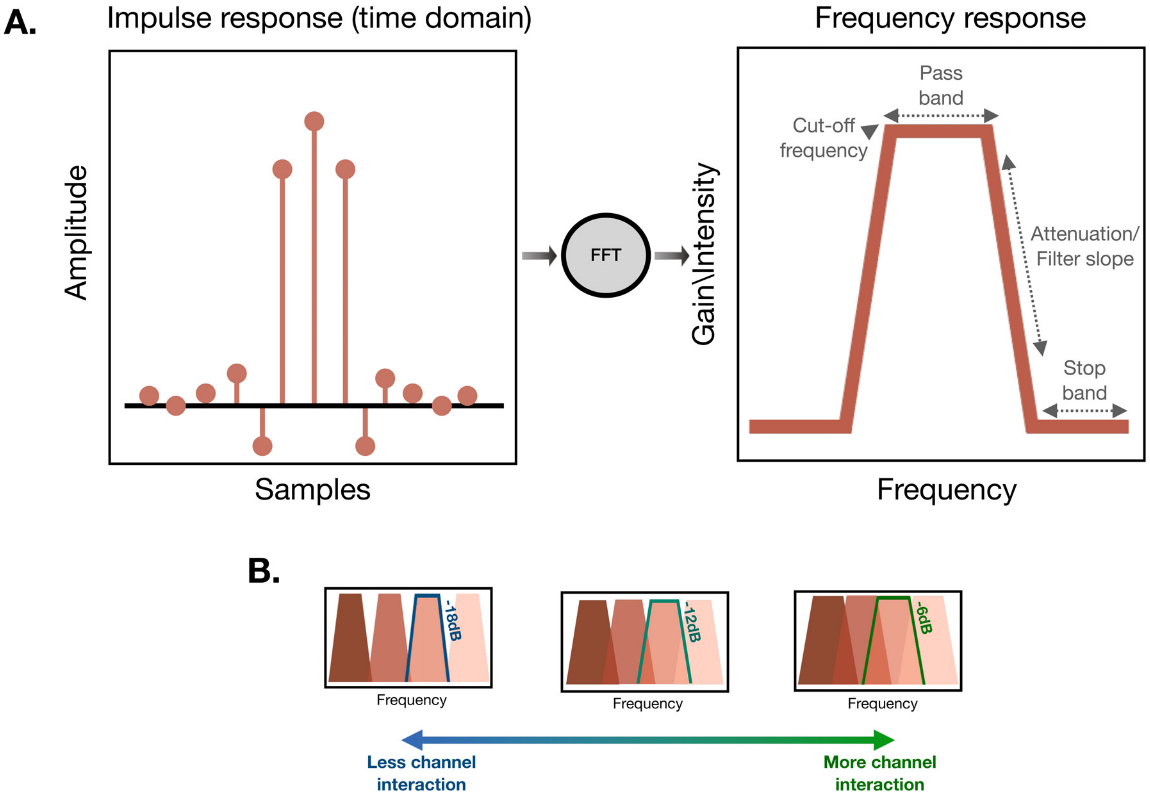


FIG. 14. (Color online) Key terminology and concepts in digital filter construction. (A) Digital filter construction in the time and frequency domains: applying a Fourier transform to the filter's impulse response generates the frequency response, or the relationship between gain and frequency. The frequency response is idealized as it has a maximally flat passband and complete attenuation in the stopband. (B) Idealized visualization of common digital filter slopes. The steeper the filter slope, the higher the filter order. Steeper IIR filter slopes, for example the Butterworth filters commonly employed in vocoder research, also require wider channel bandwidths to ensure filter stability. See text for detail.

achieve the same filter slope. In reality, digital filters have some degree of spectral ripple in the passband, do not have a completely attenuated response outside of the passband (i.e., allow some frequencies beyond the cut-off to pass through in the signal), and might slightly attenuate some of the energy within the nominal frequency bandwidth of each channel. Therefore, the experimenter must decide whether it is more important to ensure that all frequencies within the passband are included (using a lower-order filter), or if it is instead more important to ensure that energy outside the passband is excluded (using a higher-order filter). A common compromise in vocoder CI simulations is to use fourth-order Butterworth filters.

There are different methods of filterbank construction: Cochlear's processing system employs an FFT-based filtering system while AB and MED-EL are thought to use infinite impulse response (IIR) filters. It is beyond the scope of this paper to explore all of the differences between IIR and finite impulse response (FIR) filters. We encourage readers to see Tarr (2018) and Lyons (2004) for accessible written introductions to that topic, including accompanying video tutorials. FIR filters only compute output samples over the array of input samples, while IIR filters compute output samples over both input and previous output samples. The result is that the impulse response either falls to 0 after a finite period of time (FIR filter) or continues falling to 0 infinitely (IIR filter).

Vocoder filterbanks commonly consist of fourth-order (–24 dB/octave) Butterworth filters because Butterworth filters (1) have extremely flat passbands (i.e., there is uniform gain and minimal spectral ripple across the frequency band passed through the filter) and (2) have an IIR which requires fewer coefficients and is less computationally demanding than FIRs (the process of filtering can be slow and we want it to be as fast as possible). Although a maximally flat passband is ideal for many signal processing applications, the effects of passband flatness for vocoder applications is likely minimal. The Butterworth filters commonly employed in vocoder research do have flatter responses than Chebyshev or elliptic filters. However, a number of vocoder constructions instead choose to employ elliptic filters which have similar amounts of spectral ripple in the passbands and stopbands and a maximally steep roll-off into the passband—steeper than what would be achievable for the same Butterworth filter order (Shannon *et al.*, 1998). Nevertheless, the choice of filter does not greatly impact the resulting vocoded signal. The only true concern about vocoder filterbank construction is that when IIR filters, such as Butterworth or elliptic are employed, the practitioner *must ensure that the constructed filters are stable*.

What is digital filter stability? A filter is stable when its impulse response approaches 0 [see Fig. 14(A) for

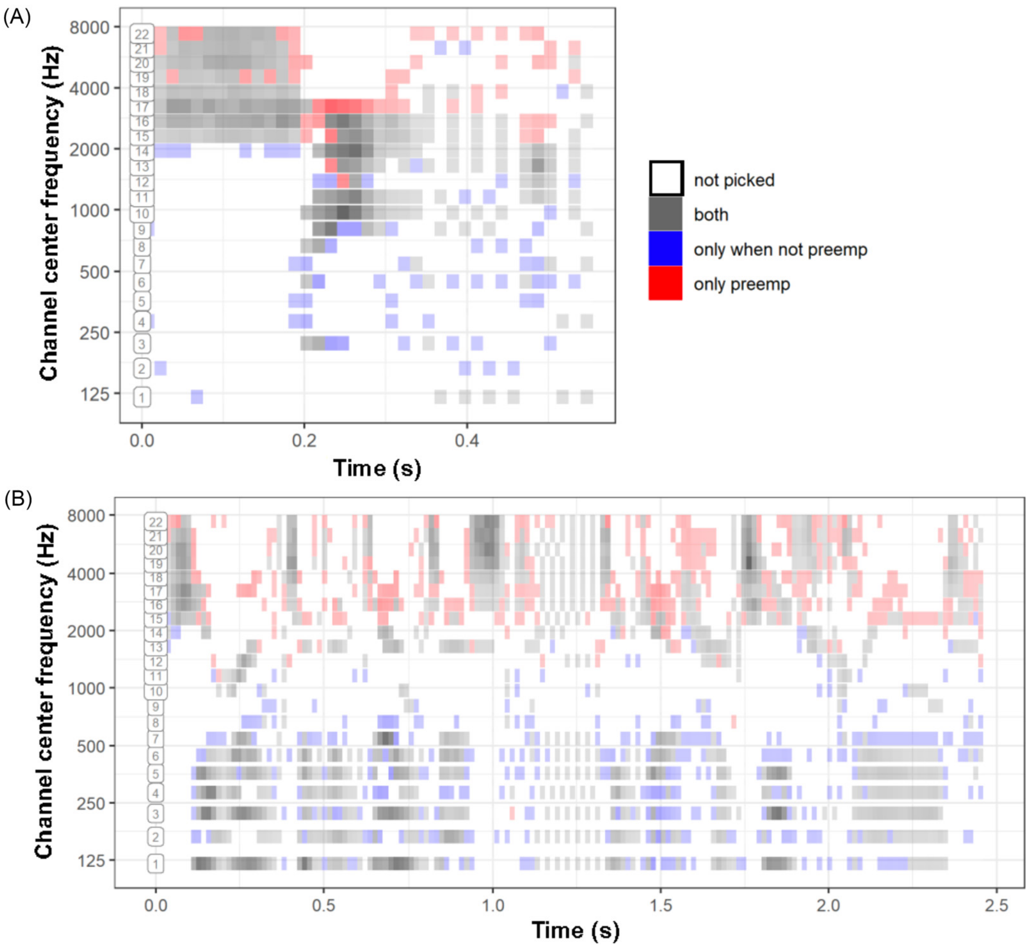


FIG. 15. (Color online) A female adult's vocoded production of "shack" (A) and "She went to the dentist to get her teeth cleaned." (B) Processed using a peak-picking strategy that selects the top 8 out of 22 channels. See Fig. 5 for interpretation.

illustration]. Filter stability is primarily a concern for IIR filters, not FIR filters, which consistently result in impulse and frequency responses that fall to 0 (the correct temporal waveform and spectrum). However, IIR filters risk instability when (1) the filter cut-off frequency approaches the Nyquist frequency and (2) the filter slope is too steep given the channel bandwidth. The latter is of particular concern when channel bandwidths are narrow. Since channel bandwidth varies as a function of the number of channels in the vocoder—more channels entails narrower bandwidths across the frequency spectrum—concerns about filter stability *increase* as a function of the number of vocoder channels.

There are a couple of steps that vocoder practitioners can take to ensure that their IIR filters are stable. We recommend plotting the individual impulse response and frequency spectrum for *each* channel's filter. Crucially, validating the filterbank in this way should be done *prior* to signal summation because the process of combining all the vocoder channels together may make it more difficult to identify an unstable filter. Should a filter appear unstable, the practitioner can employ wider channel bandwidths. They can also downsample the signal (to avoid interactions with the cut-off frequency) and/or employ forward-backward filtering (which will double the filter slope and produce a slightly smaller passband). In forward-backward filtering, two lower-order filters are applied sequentially. For example, a third-order filter applied forward in the time domain and subsequently backward in the time domain would have the cumulative effect of a sixth-order filter, but without the instability that applying a single sixth-order filter would entail.

2. Computing pre-emphasis

We pre-emphasize a signal in the time domain by applying a first-order differencing filter [one of a number of different filters that could be employed at this stage; e.g., Xu *et al.* (2005a)], which entails computing the difference between adjacent samples, say, n and $n-1$, in an input signal. A scalar is applied to one of those samples, which controls the amount of pre-emphasis applied to the overall signal. We can write this formally as

$$y(n) = x(n) - \alpha * x(n-1), \qquad \text{(A1)}$$

where $x(n)$ is a sample in the input signal and $x(n-1)$ is the previous sample, α is the scalar (typically 0–1) to control the amount of pre-emphasis, and y is the output signal.

3. Additional visualizations

A female adult's vocoded production of "shack" (A) and "She went to the dentist to get her teeth cleaned." (B) Processed using a peak-picking strategy that selects the top 8 out of 22 channels (Fig 15).

[1] Some processing strategies from MED-EL attempt to represent temporal fine structure better, especially at the lower frequencies.

[2] In theory this may have been to approximate the estimated upper limit of rate-pitch perception in CIs, but some CI users can resolve rates >1000 Hz.

[3] Temporal fine structure cues in 1-channel vocoders may be difficult to interpret as they may not involve temporal encoding; see Shamma and Lorenzi (2013).

[4] Formants are the resonant frequencies of the vocal tract that, among other things, differentiate vowel quality; see Johnson (2011) for a detailed overview.

[5] Interestingly, heightened amplitude modulation does not appear to interfere with speech perception in CI users. There are few differences in CI users' word recognition performance between noise, tonal, and noise-modulated tonal maskers (Oxenham and Kreft, 2014). The authors attribute this finding to the reduced spectral resolution of electrical hearing.

[6] Envelope modulations higher than the low-pass envelope filter cut-off are greatly reduced.

[7] Resolved sidebands also present an experimental confound in vocoder studies since sidebands provide spectral cues that are unavailable to actual CI users.

[8] There are limits to the accessibility of temporal cues for f0 information; however, because above approximately 200 Hz, listener sensitivity to envelope modulation rate begins to decrease (Chatterjee and Peng, 2008).

[9] Nevertheless, even in vocoded signals with 100+ channels, performance does not reach the same levels as with unprocessed signals suggesting that temporal fine structure also is important for speech perception (Gibbs *et al.*, 2022) (100 noise bands would have spurious modulations that could hurt performance, too). And as always, there is an inverse relationship between channel number and bandwidth. So in simulations with large numbers of channels, practitioners must ensure that filters are stable (see Sec. III A).

[10] Increasingly, insertion depth is expressed as an angle (turns around the cochlea) instead of length (mm) (Landsberger *et al.*, 2015).

[11] In theory, longer arrays could produce less tonotopic mismatch (Landsberger *et al.*, 2015), but it is important to remember that longer arrays/deeper insertions are not uniformly beneficial as they may compromise low-frequency residual hearing.

[12] Filter order, or slope, can likewise be referred to with a Pole Number. So a first-order digital filter (−6 dB/octave), can alternatively be referred to as a one-pole filter. A second-order filter (−12 dB/octave) is a two-pole filter, etc.

Ananthakrishnan, S., and Luo, X. (**2022**). "Effects of temporal envelope cutoff frequency, number of channels, and carrier type on brainstem neural representation of pitch in vocoded speech," J. Speech. Lang. Hear. Res. **65**(8), 3146–3164.

Ananthakrishnan, S., Luo, X., and Krishnan, A. (**2017**). "Human frequency following responses to vocoded speech," Ear Hear. **38**(5), e256–e267.

Bacon, S. P., and Viemeister, N. F. (**1985**). "Temporal modulation transfer functions in normal-hearing and hearing-impaired listeners," Int. J. Audiol. **24**(2), 117–134.

Barreda, S., and Nearey, T. M. (**2012**). "The direct and indirect roles of fundamental frequency in vowel perception," J. Acoust. Soc. Am. **131**(1), 466–477.

Başkent, D. (**2006**). "Speech recognition in normal hearing and sensorineural hearing loss as a function of the number of spectral channels," J. Acoust. Soc. Am. **120**(5), 2908–2925.

Baskent, D., and Shannon, R. V. (**2003**). "Speech recognition under conditions of frequency-place compression and expansion," J. Acoust. Soc. Am. **113**(4), 2064–2076.

Başkent, D., and Shannon, R. V. (**2004**). "Frequency-place compression and expansion in cochlear implant listeners," J. Acoust. Soc. Am. **116**(5), 3130–3140.

Başkent, D., and Shannon, R. V. (**2005**). "Interactions between cochlear implant electrode insertion depth and frequency-place mapping," J. Acoust. Soc. Am. **117**(3), 1405–1416.

Başkent, D., and Shannon, R. V. (**2006**). "Frequency transposition around dead regions simulated with a noiseband vocoder," J. Acoust. Soc. Am. **119**(2), 1156–1163.

Berenstein, C. K., Mens, L. H. M., Mulder, J. J. S., and Vanpoucke, F. J. (**2008**). "Current Steering and current focusing in cochlear implants: Comparison of monopolar, tripolar, and virtual channel electrode configurations," Ear Hear. **29**(2), 250–260.

Berg, K. A., Noble, J. H., Dawant, B. M., Dwyer, R. T., Labadie, R. F., and Gifford, R. H. (**2019**). "Speech recognition as a function of the number of channels in perimodiolar electrode recipients," J. Acoust. Soc. Am. **145**(3), 1556–1564.

Berg, K. A., Noble, J. H., Dawant, B. M., Dwyer, R. T., Labadie, R. F., and Gifford, R. H. (**2021**). "Speech recognition as a function of the number of channels for an array with large inter-electrode distances," J. Acoust. Soc. Am. **149**(4), 2752–2763.

Bhargava, P., Gaudrain, E., and Başkent, D. (**2016**). "The intelligibility of interrupted speech: Cochlear implant users and normal hearing listeners," J. Assoc. Res. Otolaryngol. **17**(5), 475–491.

Bierer, J. A. (**2007**). "Threshold and channel interaction in cochlear implant users: Evaluation of the tripolar electrode configuration," J. Acoust. Soc. Am. **121**(3), 1642–1653.

Bierer, J. A. (**2010**). "Probing the electrode-neuron interface with focused cochlear implant stimulation," Trends Amplif. **14**(2), 84–95.

Bierer, J. A., and Litvak, L. (**2016**). "Reducing channel interaction through cochlear implant programming may improve speech perception: Current focusing and channel deactivation," Trends Hear. **20**, 2331216516653389.

Bingabr, M., Espinoza-Varas, B., and Loizou, P. C. (**2008**). "Simulating the effect of spread of excitation in cochlear implants," Hear. Res. **241**(1-2), 73–79.

Blamey, P., Artieres, F., Başkent, D., Bergeron, F., Beynon, A., Burke, E., Dillier, N., Dowell, R., Fraysse, B., Gallégo, S., Govaerts, P. J., Green, K., Huber, A. M., Kleine-Punte, A., Maat, B., Marx, M., Mawman, D., Mosnier, I., O'Connor, A. F., O'Leary, S., Rousset, A., Schauwers, K., Skarzynski, H., Skarzynski, P. H., Sterkers, O., Terranti, A., Truy, E., Van de Heyning, P., Venail, F., Vincent, C., and Lazard, D. S. (**2013**). "Factors affecting auditory performance of postlinguistically deaf adults using cochlear implants: An update with 2251 patients," Audiol. Neurotol. **18**(1), 36–47.

Blomquist, C., Newman, R. S., Huang, Y. T., and Edwards, J. (**2021**). "Children with cochlear implants use semantic prediction to facilitate spoken word recognition," J. Speech. Lang. Hear. Res. **64**(5), 1636–1649.

Boisvert, I., Reis, M., Au, A., Cowan, R., and Dowell, R. C. (**2020**). "Cochlear implantation outcomes in adults: A scoping review," PLoS One **15**(5), e0232421.

Brown, C. J., Abbas, P. J., Bertschy, M., Tyler, R. S., Lowder, M., Takahashi, G., Purdy, S., and Gantz, B. J. (**1995**). "Longitudinal assessment of physiological and psychophysical measures in cochlear implant users," Ear Hear. **16**(5), 439–449.

Canfarotta, M. W., Dillon, M. T., Brown, K. D., Pillsbury, H. C., Dedmon, M. M., and O'Connell, B. P. (**2022**). "Insertion depth and cochlear implant speech recognition outcomes: A comparative study of 28- and 31.5-mm lateral wall arrays," Otol. Neurotol. **43**(2), 183–189.

Canfarotta, M. W., Dillon, M. T., Buss, E., Pillsbury, H. C., Brown, K. D., and O'Connell, B. P. (**2020**). "Frequency-to-place mismatch: Characterizing variability and the influence on speech perception outcomes in cochlear implant recipients," Ear Hear. **41**(5), 1349–1361.

Carlyon, R. P., Long, C. J., and Deeks, J. M. (**2008**). "Pulse-rate discrimination by cochlear-implant and normal-hearing listeners with and without binaural cues," J. Acoust. Soc. Am. **123**(4), 2276–2286.

Chatterjee, M., and Peng, S.-C. (**2008**). "Processing F0 with cochlear implants: Modulation frequency discrimination and speech intonation recognition," Hear. Res. **235**(1-2), 143–156.

Chatterjee, M., Zion, D. J., Deroche, M. L., Burianek, B. A., Limb, C. J., Goren, A. P., Kulkarni, A. M., and Christensen, J. A. (**2015**). "Voice emotion recognition by cochlear-implanted children and their normally-hearing peers," Hear. Res. **322**, 151–162.

Cheung, C., Hamilton, L. S., Johnson, K., and Chang, E. F. (**2016**). "The auditory representation of speech sounds in human motor cortex," eLife **5**, e12577.

Churchill, T. H., Kan, A., Goupell, M. J., Ihlefeld, A., and Litovsky, R. Y. (**2014**). "Speech perception in noise with a harmonic complex excited vocoder," J. Assoc. Res. Otolaryngol. **15**(2), 265–278.

Crew, J. D., Galvin, J. J., and Fu, Q.-J. (**2012**). "Channel interaction limits melodic pitch perception in simulated cochlear implants," J. Acoust. Soc. Am. **132**(5), EL429–EL435.

Croghan, N. B. H., Duran, S. I., and Smith, Z. M. (**2017**). "Re-examining the relationship between number of cochlear implant channels and maximal speech intelligibility," J. Acoust. Soc. Am. **142**(6), EL537–EL543.

Cychosz, M., Xu, K., and Fu, Q.-J. (**2023**). "Effects of spectral smearing on speech understanding and masking release in simulated bilateral cochlear implants," PLoS One **18**, e0287728.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., and McGettigan, C. (**2005**). "Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences," J. Exp. Psychol.: General **134**(2), 222–241.

DeRoy Milvae, K., 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.

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.

Dillon, M. T., Canfarotta, M. W., Buss, E., Hopfinger, J., and O'Connell, B. P. (**2021**). "Effectiveness of place-based mapping in electric-acoustic stimulation devices," Otol. Neurotol. **42**(1), 197–202.

Dillon, M. T., O'Connell, B. P., Canfarotta, M. W., Buss, E., and Hopfinger, J. (**2022**). "Effect of place-based versus default mapping procedures on masked speech recognition: Simulations of cochlear implant alone and electric-acoustic stimulation," Am. J. Audiol. **31**(2), 322–337.

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.

Dorman, M., Loizou, P., and Rainey, D. (**1997a**). "Simulating the effect of cochlear-implant electrode insertion depth on speech understanding," J. Acoust. Soc. Am. **102**, 2993–2996.

Dorman, M. F., Loizou, P. C., Kirk, K. I., and Svirsky, M. A. (**1998**). "Channels, children and the Multisyllabic Lexical Neighborhood Test (MLNT)," NIH Neural Prosthesis Workshop, Bethesda, MD, October 1998.

Dorman, M. F., Loizou, P. C., and Rainey, D. (**1997b**). "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs," J. Acoust. Soc. Am. **102**(4), 2403–2411.

Dorman, M. F., Natale, S. C., Baxter, L., Zeitler, D. M., Carlson, M. L., Lorens, A., Skarzynski, H., Peters, J. P. M., Torres, J. H., and Noble, J. H. (**2020**). "Approximations to the voice of a cochlear implant: Explorations with single-sided deaf listeners," Trends Hear. **24**, 233121652092007.

Drullman, R. (**1995**). "Temporal envelope and fine structure cues for speech intelligibility," J. Acoust. Soc. Am. **97**(1), 585–592.

Dudley, H. (**1939**). "Remaking speech," J. Acoust. Soc. Am. **11**(2), 169–177.

Eisenberg, L. S., Martinez, A. S., Holowecky, S. R., and Pogorelsky, S. (**2002**). "Recognition of lexically controlled words and sentences by children with normal hearing and children with cochlear implants," Ear Hear. **23**(5), 450–462.

Eisenberg, L. S., Shannon, R. V., Schaefer Martinez, A., Wygonski, J., and Boothroyd, A. (**2000**). "Speech recognition with reduced spectral cues as a function of age," J. Acoust. Soc. Am. **107**(5), 2704–2710.

Faulkner, A., Rosen, S., and Smith, C. (**2000**). "Effects of the salience of pitch and periodicity information on the intelligibility of four-channel vocoded speech: Implications for cochlear implants," J. Acoust. Soc. Am. **108**(4), 1877–1887.

Fitzgerald, M., Sagi, E., Morbiwala, T. A., Tan, C.-T., and Svirsky, M. A. (**2013**). "Feasibility of real-time selection of frequency tables in an acoustic simulation of a cochlear implant," Ear Hear. **34**(6), 763–772.

Fogerty, D., Xu, J., and Gibbs, B. E. (**2016**). "Modulation masking and glimpsing of natural and vocoded speech during single-talker modulated noise: Effect of the modulation spectrum," J. Acoust. Soc. Am. **140**(3), 1800–1816.

Fourakis, M., Hawks, J. W., Holden, L. K., Skinner, M. W., and Holden, T. A. (**2004**). "Effect of frequency boundary assignment on vowel recognition with the Nucleus 24 ACE speech coding strategy," J. Am. Acad. Audiol. **15**, 281–299.

Friesen, L. M., Shannon, R. V., Baskent, 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. (**2012**). "AngelSim: Cochlear implant and hearing loss simulation," http://angelsim.emilyfufoundation.org/angelsim_about.html.

Fu, Q.-J., Chinchilla, S., and Galvin, J. J. (**2004**). "The role of spectral and temporal cues in voice gender discrimination by normal-hearing listeners and cochlear implant users," J. Assoc. Res. Otolaryngol. **5**(3), 253–260.

Fu, Q.-J., and Nogaki, G. (**2005**). "Noise susceptibility of cochlear implant users: The role of spectral resolution and smearing," J. Assoc. Res. Otolaryngol. **6**(1), 19–27.

Fu, Q.-J., and Shannon, R. V. (**1999**). "Recognition of spectrally degraded and frequency-shifted vowels in acoustic and electric hearing," J. Acoust. Soc. Am. **105**(3), 1889–1900.

Fu, Q.-J., Shannon, R. V., and Galvin, J. J. (**2002**). "Perceptual learning following changes in the frequency-to-electrode assignment with the Nucleus-22 cochlear implant," J. Acoust. Soc. Am. **112**(4), 1664–1674.

Gaudrain, E. (**2016**). "Vocoder, v1.0," https://github.com/egaudrain/vocoder (Last viewed May 15, 2023).

Gaudrain, E., and Başkent, D. (**2015**). "Factors limiting vocal-tract length discrimination in cochlear implant simulations," J. Acoust. Soc. Am. **137**(3), 1298–1308.

Gibbs, B. E., Bernstein, J. G. W., Brungart, D. S., and Goupell, M. J. (**2022**). "Effects of better-ear glimpsing, binaural unmasking, and spectral resolution on spatial release from masking in cochlear-implant users," J. Acoust. Soc. Am. **152**(2), 1230–1246.

Golinkoff, R. M., Ma, W., Song, L., and Hirsh-Pasek, K. (**2013**). "Twenty-five years using the intermodal preferential looking paradigm to study language acquisition: What have we learned?," Perspect. Psychol. Sci. **8**(3), 316–339.

Gonzalez, J., and Oliver, J. C. (**2005**). "Gender and speaker identification as a function of the number of channels in spectrally reduced speech," J. Acoust. Soc. Am. **118**(1), 461–470.

Goupell, M. J., Gaskins, C. R., Shader, M. J., Walter, E. P., Anderson, S., and Gordon-Salant, S. (**2017**). "Age-related differences in the processing of temporal envelope and spectral cues in a speech segment," Ear Hear. **38**(6), e335–e342.

Goupell, M. J., Laback, B., Majdak, P., and Baumgartner, W.-D. (**2008**). "Effects of upper-frequency boundary and spectral warping on speech intelligibility in electrical stimulation," J. Acoust. Soc. Am. **123**(4), 2295–2309.

Goupell, M. J., Majdak, P., and Laback, B. (**2010**). "Median-plane sound localization as a function of the number of spectral channels using a channel vocoder," J. Acoust. Soc. Am. **127**(2), 990–1001.

Goupell, M. J., Stoelb, C., Kan, A., and Litovsky, R. Y. (**2013**). "Effect of mismatched place-of-stimulation on the salience of binaural cues in conditions that simulate bilateral cochlear-implant listening," J. Acoust. Soc. Am. **133**(4), 2272–2287.

Grange, J. A., Culling, J. F., Harris, N. S. L., and Bergfeld, S. (**2017**). "Cochlear implant simulator with independent representation of the full spiral ganglion," J. Acoust. Soc. Am. **142**(5), EL484–EL489.

Greenwood, D. D. (**1990**). "A cochlear frequency-position function for several species—29 years later," J. Acoust. Soc. Am. **87**(6), 2592–2605.

Grieco-Calub, T. M., Simeon, K. M., Snyder, H. E., and Lew-Williams, C. (**2017**). "Word segmentation from noise-band vocoded speech," Lang. Cogn. Neurosci. **32**(10), 1344–1356.

Hartmann, W. M., and Pumplin, J. (**1988**). "Noise power fluctuations and the masking of sine signals," J. Acoust. Soc. Am. **83**(6), 2277–2289.

Hazan, V., and Barrett, S. (**2000**). "The development of phonemic categorization in children aged 6–12," J. Phon. **28**(4), 377–396.

Henry, B. A., McKay, C. M., McDermott, H. J., and Clark, G. M. (**2000**). "The relationship between speech perception and electrode discrimination in cochlear implantees," J. Acoust. Soc. Am. **108**(3), 1269–1280.

Hervais-Adelman, A. G., Davis, M. H., Johnsrude, I. S., Taylor, K. J., and Carlyon, R. P. (**2011**). "Generalization of perceptual learning of vocoded speech," J. Exp. Psychol.: Hum. Percept. Perform. **37**(1), 283–295.

Hughes, M. L., Vander Werff, K. R., Brown, C. J., Abbas, P. J., Kelsay, D. M. R., Teagle, H. F. B., and Lowder, M. W. (**2001**). "A longitudinal study of electrode impedance, the electrically evoked compound action potential, and behavioral measures in nucleus 24 cochlear implant users," Ear Hear. **22**(6), 471–486.

Hughes, S. E., Hutchings, H. A., Rapport, F. L., McMahon, C. M., and Boisvert, I. (**2018**). "Social connectedness and perceived listening effort in adult cochlear implant users: A grounded theory to establish content validity for a new patient-reported outcome measure," Ear Hear. **39**(5), 922–934.

Jaekel, B. N., Newman, R. S., and Goupell, M. J. (**2018**). "Age effects on perceptual restoration of degraded interrupted sentences," J. Acoust. Soc. Am. **143**(1), 84–97.

Jahn, K. N., DiNino, M., and Arenberg, J. G. (**2019**). "Reducing simulated channel interaction reveals differences in phoneme identification between children and adults with normal hearing," Ear Hear. **40**(2), 295–311.

Johnson, K. (**2011**). *Acoustic and Auditory Phonetics*, 3rd ed. (Wiley-Blackwell, Chichester, UK).

Kasturi, K., Loizou, P. C., Dorman, M., and Spahr, T. (**2002**). "The intelligibility of speech with 'holes' in the spectrum," J. Acoust. Soc. Am. **112**(3), 1102–1111.

Klein, K. E., Walker, E. A., and McMurray, B. (**2023**). "Delayed lexical access and cascading effects on spreading semantic activation during spoken word recognition in children with hearing aids and cochlear implants: Evidence from eye-tracking," Ear Hear. **44**(2), 338–357.

Kohlrausch, A., Fassel, R., and Dau, T. (**2000**). "The influence of carrier level and frequency on modulation and beat-detection thresholds for sinusoidal carriers," J. Acoust. Soc. Am. **108**, 723–734.

Kong, Y.-Y., and Carlyon, R. P. (**2010**). "Temporal pitch perception at high rates in cochlear implants," J. Acoust. Soc. Am. **127**(5), 3114–3123.

Landsberger, D. M., Svrakic, S., Roland, J. T., and Svirsky, M. (**2015**). "The relationship between insertion angles, default frequency allocations, and spiral ganglion place pitch in cochlear implants," Ear Hear. **36**(5), e207–e213.

Lenarz, M., Sönmez, H., Joseph, G., Büchner, A., and Lenarz, T. (**2012**). "Long-term performance of cochlear implants in postlingually deafened adults," Otolaryngol. Head Neck Surg. **147**(1), 112–118.

Li, T., and Fu, Q.-J. (**2010**). "Effects of spectral shifting on speech perception in noise," Hear. Res. **270**(1-2), 81–88.

Litovsky, R. (**2015**). "Development of the auditory system," Handbook Clin. Neurol. **129**, 55–72.

Litvak, L. M., Spahr, A. J., Saoji, A. A., and Fridman, G. Y. (**2007**). "Relationship between perception of spectral ripple and speech recognition in cochlear implant and vocoder listeners," J. Acoust. Soc. Am. **122**(2), 982–991.

Loizou, P. C. (**2006**). "Speech processing in vocoder-centric cochlear implants," in *Advances in Oto-Rhino-Laryngology*, edited by A. Møller (S. Karger AG, Basel, Switzerland), Vol. 64, pp. 109–143.

Loizou, P. C., Dorman, M., and Tu, Z. (**1999**). "On the number of channels needed to understand speech," J. Acoust. Soc. Am. **106**(4), 2097–2103.

Lyons, R. G. (**2004**). *Understanding Digital Signal Processing*, 3rd ed. (Pearson, London, UK).

Macherey, O., and Carlyon, R. P. (**2014**). "Cochlear implants," Curr. Biol. **24**(18), R878–R884.

Martin, I. A., Goupell, M. J., and Huang, Y. T. (**2022**). "Children's syntactic parsing and sentence comprehension with a degraded auditory signal," J. Acoust. Soc. Am. **151**(2), 699–711.

Mehta, A. H., Lu, H., and Oxenham, A. J. (**2020**). "The perception of multiple simultaneous pitches as a function of number of spectral channels and spectral spread in a noise-excited envelope vocoder," J. Assoc. Res. Otolaryngol. **21**(1), 61–72.

Mesnildrey, Q., Hilkhuysen, G., and Macherey, O. (**2016**). "Pulse-spreading harmonic complex as an alternative carrier for vocoder simulations of cochlear implants," J. Acoust. Soc. Am. **139**(2), 986–991.

Moore, B. C. J. (**2001**). "Dead regions in the cochlea: Diagnosis, perceptual consequences, and implications for the fitting of hearing aids," Trends Amplif. **5**(1), 1–34.

Moore, B. C. J., and Glasberg, B. R. (**2004**). "A revised model of loudness perception applied to cochlear hearing loss," Hear. Res. **188**(1-2), 70–88.

Morgan, J. L., and Demuth, K. (**2014**). *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*, 2nd ed. (Taylor & Francis, London, UK).

Newman, R. S., and Chatterjee, M. (**2013**). "Toddlers' recognition of noise-vocoded speech," J. Acoust. Soc. Am. **133**(1), 483–494.

Newman, R. S., Chatterjee, M., Morini, G., and Remez, R. E. (**2015**). "Toddlers' comprehension of degraded signals: Noise-vocoded versus sine-wave analogs," J. Acoust. Soc. Am. **138**(3), EL311–EL317.

Newman, R. S., Morini, G., Shroads, E., and Chatterjee, M. (**2020**). "Toddlers' fast-mapping from noise-vocoded speech," J. Acoust. Soc. Am. **147**(4), 2432–2441.

NIDCD (**2021**). "Quick statistics about hearing," Technical report, National Institutes of Health, Washington, DC.

Nilsson, M., Soli, S. D., and Sullivan, J. A. (**1994**). "Development of the Hearing in Noise Test for the measurement of speech recognition thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**(2), 1085–1099.

Nittrouer, S., and Lowenstein, J. H. (**2010**). "Learning to perceptually organize speech signals in native fashion," J. Acoust. Soc. Am. **127**(3), 1624–1635.

Nittrouer, S., Lowenstein, J. H., and Packer, R. (**2009**). "Children discover the spectral skeletons in their native language before the amplitude envelopes," J. Exp. Psychol. Human Percept. Perform. **35**(4), 1245–1253.

Nittrouer, S., Lowenstein, J. H., Wucinich, T., and Tarr, E. (**2014a**). "Benefits of preserving stationary and time-varying formant structure in alternative representations of speech: Implications for cochlear implants," J. Acoust. Soc. Am. **136**(4), 1845–1856.

Nittrouer, S., Tarr, E., Bolster, V., Caldwell-Tarr, A., Moberly, A. C., and Lowenstein, J. H. (**2014b**). "Low-frequency signals support perceptual organization of implant-simulated speech for adults and children," Int. J. Audiol. **53**(4), 270–284.

O'Neill, E. R., Parke, M. N., Kreft, H. A., and Oxenham, A. J. (**2021**). "Role of semantic context and talker variability in speech perception of cochlear-implant users and normal-hearing listeners," J. Acoust. Soc. Am. **149**(2), 1224–1239.

Oxenham, A. J., and Kreft, H. A. (**2014**). "Speech perception in tones and noise via cochlear implants reveals influence of spectral resolution on temporal processing," Trends Hear. **18**, 233121651455378.

Pfingst, B. E., Xu, L., and Thompson, C. S. (**2004**). "Across-site threshold variation in cochlear implants: Relation to speech recognition," Audiol. Neurotol. **9**(6), 341–352.

Pisoni, D. B., Kronenberger, W. G., Harris, M. S., and Moberly, A. C. (**2017**). "Three challenges for future research on cochlear implants," World J. Otorhinolaryngol. Head Neck Surg. **3**(4), 240–254.

Pumplin, J. (**1985**). "Low-noise noise," J. Acoust. Soc. Am. **78**(1), 100–104.

Qin, M. K., and Oxenham, A. J. (**2006**). "Effects of introducing unprocessed low-frequency information on the reception of envelope-vocoder processed speech," J. Acoust. Soc. Am. **119**(4), 2417–2426.

Roman, A. S., Pisoni, D. B., Kronenberger, W. G., and Faulkner, K. F. (**2017**). "Some neurocognitive correlates of noise-vocoded speech perception in children with normal hearing: A replication and extension of Eisenberg *et al.* (2002)," Ear Hear. **38**(3), 344–356.

Rosen, S. (**1992**). "Temporal information in speech: Acoustic, auditory and linguistic aspects," Philos. Trans. R. Soc. London B Biol. Sci. **336**(1278), 367–373.

Rosen, S., Faulkner, A., and Wilkinson, L. (**1999**). "Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants," J. Acoust. Soc. Am. **106**(6), 3629–3636.

Rosen, S., Zhang, Y., and Speers, K. (**2015**). "Spectral density affects the intelligibility of tone-vocoded speech: Implications for cochlear implant simulations," J. Acoust. Soc. Am. **138**(3), EL318–EL323.

Schvartz, K. C., Chatterjee, M., and Gordon-Salant, S. (**2008**). "Recognition of spectrally degraded phonemes by younger, middle-aged, and older normal-hearing listeners," J. Acoust. Soc. Am. **124**(6), 3972–3988.

Shader, M. J., Yancey, C. M., Gordon-Salant, S., and Goupell, M. J. (**2020**). "Spectral-temporal trade-off in vocoded sentence recognition: Effects of age, hearing thresholds, and working memory," Ear Hear. **41**(5), 1226–1235.

Shannon, R. V. (**1992**). "Temporal modulation transfer functions in patients with cochlear implants," J. Acoust. Soc. Am. **91**, 2156–2164.

Shamma, S., and Lorenzi, C. (**2013**). "On the balance of envelope and temporal fine structure in the encoding of speech in the early auditory system," J. Acoust. Soc. Am. **133**(5), 2818–2833.

Shannon, R. V., Fu, Q.-J., and Galvin, J. J. (**2004**). "The number of spectral channels required for speech recognition depends on the difficulty of the listening situation," Acta Oto-Laryngolog. (Suppl.) **124**, 50–54.

Shannon, R. V., Galvin, J. J., III, and Baskent, D. (**2002**). "Holes in hearing," J. Assoc. Res. Otolaryngol. **3**(2), 185–199.

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.

Shannon, R. V., Zeng, F.-G., and Wygonski, J. (**1998**). "Speech recognition with altered spectral distribution of envelope cues," J. Acoust. Soc. Am. **104**(4), 2467–2476.

Sheldon, S., Pichora-Fuller, M. K., and Schneider, B. A. (**2008**). "Effect of age, presentation method, and learning on identification of noise-vocoded words," J. Acoust. Soc. Am. **123**(1), 476–488.

Smith, M. L., and Winn, M. B. (**2021**). "Individual variability in recalibrating to spectrally shifted speech: Implications for cochlear implants," Ear Hear. **42**, 1412–1427.

Souza, P., and Rosen, S. (**2009**). "Effects of envelope bandwidth on the intelligibility of sine- and noise-vocoded speech," J. Acoust. Soc. Am. **126**(2), 792–805.

Srinivasan, A. G., Shannon, R. V., and Landsberger, D. M. (**2012**). "Improving virtual channel discrimination in a multi-channel context," Hear. Res. **286**(1-2), 19–29.

Stafford, R. C., Stafford, J. W., Wells, J. D., Loizou, P. C., and Keller, M. D. (**2014**). "Vocoder simulations of highly focused cochlear stimulation with limited dynamic range and discriminable steps," Ear Hear. **35**(2), 262–270.

Stakhovskaya, O., Sridhar, D., Bonham, B. H., and Leake, P. A. (**2007**). "Frequency map for the human cochlear spiral ganglion: Implications for cochlear implants," J. Assoc. Res. Otolaryngol. **8**(2), 220–233.

Stilp, C., Donaldson, G., Oh, S., and Kong, Y.-Y. (**2016**). "Influences of noise-interruption and information-bearing acoustic changes on understanding simulated electric-acoustic speech," J. Acoust. Soc. Am. **140**(5), 3971–3979.

Stone, M. A., Füllgrabe, C., and Moore, B. C. J. (**2008**). "Benefit of high-rate envelope cues in vocoder processing: Effect of number of channels and spectral region," J. Acoust. Soc. Am. **124**(4), 2272–2282.

Strydom, T., and Hanekom, J. J. (**2011**). "The performance of different synthesis signals in acoustic models of cochlear implants," J. Acoust. Soc. Am. **129**(2), 920–933.

Sun, X. (**2000**). "Voice quality conversion in TD-PSOLA speech synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, Vol. 2, pp. II953–II956.

Tarr, E. (**2018**). *Hack Audio: An Introduction to Computer Programming and Digital Signal Processing in MATLAB*, 1st ed. (Routledge, New York).

Tinnemore, A. R., Gordon-Salant, S., and Goupell, M. J. (**2020**). "Audiovisual speech recognition with a cochlear implant and increased perceptual and cognitive demands," Trends Hear. **24**, 233121652096060.

Tinnemore, A. R., Montero, L., Gordon-Salant, S., and Goupell, M. J. (**2022**). "The recognition of time-compressed speech as a function of age in listeners with cochlear implants or normal hearing," Front. Aging Neurosci. **14**, 887581.

Tinnemore, A. R., Zion, D. J., Kulkarni, A. M., and Chatterjee, M. (**2018**). "Children's recognition of emotional prosody in spectrally-degraded speech is predicted by their age and cognitive status," Ear Hear. **39**(5), 874–880.

Turner, C. W., Chi, S.-L., and Flock, S. (**1999**). "Limiting spectral resolution in speech for listeners with sensorineural hearing loss," J. Speech. Lang. Hear. Res. **42**(4), 773–784.

Verschooten, E., Shamma, S., Oxenham, A. J., Moore, B. C., Joris, P. X., Heinz, M. G., and Plack, C. J. (**2019**). "The upper frequency limit for the use of phase locking to code temporal fine structure in humans: A compilation of viewpoints," Hear. Res. **377**, 109–121.

Waddington, E., Jaekel, B. N., Tinnemore, A. R., Gordon-Salant, S., and Goupell, M. J. (**2020**). "Recognition of accented speech by cochlear-implant listeners: Benefit of audiovisual cues," Ear Hear. **41**(5), 1236–1250.

Waked, A., Dougherty, S., and Goupell, M. J. (**2017**). "Vocoded speech perception with simulated shallow insertion depths in adults and children," J. Acoust. Soc. Am. **141**(1), EL45–EL50.

Warner-Czyz, A. D., Roland, J. T., Thomas, D., Uhler, K., and Zombek, L. (**2022**). "American cochlear implant alliance task force guidelines for determining cochlear implant candidacy in children," Ear Hear. **43**(2), 268–282.

Whitmal, N. A., Poissant, S. F., Freyman, R. L., and Helfer, K. S. (**2007**). "Speech intelligibility in cochlear implant simulations: Effects of carrier type, interfering noise, and subject experience," J. Acoust. Soc. Am. **122**(4), 2376–2388.

Wilson, B. S., and Dorman, M. F. (**2008**). "Cochlear implants: A remarkable past and a brilliant future," Hear. Res. **242**(1-2), 3–21.

Winn, M. B. (**2020**). "Accommodation of gender-related phonetic differences by listeners with cochlear implants and in a variety of vocoder simulations," J. Acoust. Soc. Am. **147**(1), 174–190.

Winn, M. B. (**2024**). "Praat vocoder," https://github.com/ListenLab/Vocoder (Last viewed November 15, 2023).

Winn, M. B., Edwards, J. R., and Litovsky, R. Y. (**2015**). "The impact of auditory spectral resolution on listening effort revealed by pupil dilation," Ear Hear. **36**(4), e153–e165.

Winn, M. B., and O'Brien, G. (**2022**). "Distortion of spectral ripples through cochlear implants has major implications for interpreting performance scores," Ear Hear. **43**(3), 764–772.

Won, J. H., Jones, G. L., Moon, I. J., and Rubinstein, J. T. (**2015**). "Spectral and temporal analysis of simulated dead regions in cochlear implants," J. Assoc. Res. Otolaryngol. **16**(2), 285–307.

Xu, J., Ariyaeeinia, A., Sotudeh, R., and Ahmad, Z. (**2005a**). "Pre-processing speech signals in FPGAs," in *2005 6th International Conference on ASIC*, Vol. 2, pp. 778–782.

Xu, L., and Pfingst, B. E. (**2003**). "Relative importance of temporal envelope and fine structure in lexical-tone perception (L)," J. Acoust. Soc. Am. **114**(6), 3024–3027.

Xu, L., Thompson, C. S., and Pfingst, B. E. (**2005b**). "Relative contributions of spectral and temporal cues for phoneme recognition," J. Acoust. Soc. Am. **117**(5), 3255–3267.

Zeng, F.-G. (**2022**). "Celebrating the one millionth cochlear implant," JASA Express Lett. **2**(7), 077201.