Brainstem Encoding of Speech and Music Sounds in Humans

Nina Kraus and Trent Nicol

The Oxford Handbook of the Auditory Brainstem
Edited by Karl Kandler

Subject: Neuroscience, Sensory and Motor Systems Online Publication Date: Sep 2018
DOI: 10.1093/oxfordhb/9780190849061.013.26

Abstract and Keywords

The encoding of speech and music in the auditory brainstem is available at the human scalp via the auditory-evoked frequency following response. The FFR, primarily reflecting activity in the inferior colliculus, may be evoked by speech or music stimulation and represents the combined activity of sensorimotor, cognitive, and reward centers in the brain. Its response properties, like the inferior colliculus itself, are influenced by long-term experience with sound. The transparency, individual-level reliability, and ability to gauge neural plasticity provide the researcher and clinician a powerful probe of auditory processing in the human brainstem. With it, we have learned a great deal about how mechanisms of decline, deprivation, and enrichment affect the processing of complex signals such as music and speech in the human brainstem.

Keywords: speech, music, auditory processing, inferior colliculus, frequency following response, encoding, neural plasticity

The auditory brainstem, as detailed in earlier chapters of this book, comprises a number of structures that are specialized for certain sound-processing functions and, naturally, these structures are involved in the encoding of speech and music, as they are for any sound. In practice, however, when speaking of the brainstem encoding of speech and music in humans, we are largely limited in our investigation to the volume-conducted electrical potentials that can be picked up noninvasively from the scalp. This means that, for the most part, we are speaking of the rostral brainstem, including lateral lemniscus and inferior colliculus (Batra, Kuwada, & Maher, 1986; King, Hopkins, & Plack, 2016), the volume-conducted activity of which dominates the scalp-recorded sound-evoked response (Chandrasekaran & Kraus, 2010).

While this may seem limiting, it really is not, as the inferior colliculus is a major crossroads of afferent and efferent activity. If any single brainstem nucleus must stand as a representative of all, the IC is the worthiest to fulfill that role. In addition to receiving enervation from the lower brainstem nuclei, its response properties are profoundly impacted by learning via cortical, limbic and cerebellar feedback (Winer, 2006) and so likewise are the scalp-recorded responses (Kraus & White-Schwoch, 2015).

Speech, as anyone who has struggled to learn reading spectrograms in a speech-acoustics or phonetics class, comprises many components that overlap in time and frequency. Up to 72 phonemes are present in casually-spoken English alone (Mines, Hanson, & Shoup, 1978), and it goes without saying that for these phonemes to be accurately decoded by the listener, they must be accurately encoded by the auditory brain as discrete, unique, and discriminable auditory events. The subtleties present in music—the ever-changing timbres and textures—likewise are faithfully processed in the auditory brainstem.

A surprisingly faithful representation of both speech and music processing is available to the researcher or clinician at the human scalp. When considering volume-conducted voltage fluctuations emanating from a tiny structure deep below the scalp, we are extremely fortunate that the fidelity of complex sounds like speech or music is well maintained at the recording electrode. Although something of a parlor trick, it is nevertheless fascinating that a brain response that is sonified (played back) through a speaker is identifiable as the speech token that evoked it (Weiss & Bidelman, 2015). This maintenance of fidelity is not possible with blunt cortical evoked potentials, nor is it revealed in the brainstem response to a simple stimulus such as the click or tone pip that is familiar to the audiologist. In this chapter we will provide a very brief tutorial on the *frequency-following response* (FFR), the single best measurement of brainstem encoding of speech and music in humans, and then go on to harness its protean nature toward the description of the neural encoding of speech and music.

This is not to say that the IC in particular or the brainstem as a whole is the only neural structure where speech characteristics can be evaluated by human electrophysiology. There is a vast literature of cortical evoked responses to speech sounds, and the FFR itself is not entirely free of cortical input. However, the classic long-latency cortical evoked response is overwhelmingly blunt, signaling the detection of a speech sound or the detection of a change in a speech sound. However, there is a rather large chasm between *detection* and *encoding*. While the acoustic change complex (to name one cortical evoked-response) is able to *detect* very tiny changes in (for example) the pitch of a vowel (Martinez, Eisenberg, & Boothroyd, 2013), there is no way that the identity of the vowel or the pitch at which it was uttered could be ascertained from the ACC. In contrast, the *encoding* of speech or music by the brainstem permits one, via acoustic analysis of the FFR, to determine both the pitch and, within certain limitations, the identity of the evoking speech sound.

After a description of the FFR, we will unpack how selected classes of speech characteristics that combine to produce the 72 phonemes are encoded and subsequently recorded at the scalp as an FFR. (We will not cover music encoding to such a degree, but we will point out parallels as they arise.) Along the way, we will examine some pathologies that lead to poor brainstem encoding and some circumstances that lead to exceptional encoding.

The Frequency-Following Response

The best probe of speech and music encoding in the human brainstem is the frequency following response (FFR). First described in the 1960s via recordings in various auditory nuclei and species (Boudreau, 1965; Marsh, Worden, & Smith, 1970; Moushegian, Rupert, & Whitcomb, 1964; Worden & Marsh, 1968), it was mostly investigated with simple stimuli such as pure tones and found utility only as an adjunct to audiometry until the 1990s when Galbraith and colleagues (Galbraith, Arbagey, Branski, Comerci, & Rector, 1995; Galbraith, Bhuta, Choate, Kitahara, & Mullen, 1998; Galbraith, Jhaveri, & Kuo, 1997) began investigating the FFR to complex stimuli including speech. They and others discovered that the properties of an individual's brainstem encoding to complex sounds bears a relationship with more than peripheral hearing. Among these are attention, literacy, language, and musical experience.

Although not particularly more technically difficult to collect than a click-evoked auditory brainstem response (ABR), the analysis of the FFR is a bit more involved. Like the ABR, the timing of response peaks is an informative metric; but, because of the richness of the stimulus and, consequently, the response, the FFR lends itself to a variety of signal processing techniques that are beyond the scope of what most clinical ABR-focused evoked-response systems are able to provide. To a speech or music stimulus, the following FFR measures, among others, can be used to comprehensively investigate its encoding.

Timing.

The latencies, referenced either to stimulus onset or to one another are measured. These can be either transient peaks in the response, such as those occurring to stimulus onset and offset, or tonic peaks representing the periodicity of the stimulus during a voiced period of speech, such as a vowel or a voiced formant transition, or during a musical note.

Magnitude.

The size of the omnibus response is measured in the time domain by root-mean-square amplitude over a region of interest (e.g., the consonant-transition period of a response to a consonant-vowel syllable). Magnitude of individual peaks is usually not considered.

Frequency content and magnitude.

Frequency content of the response is obtained by Fourier transform. Key frequencies in the spectrum, such as the fundamental frequency (f0) of the stimulus and its harmonics, are identified and the size is measured.

Frequency tracking.

In the case of a response to a stimulus with a roving f0, the instantaneous frequency of both the stimulus and response are derived over the course of the waveforms by one of several techniques including autocorrelation, Fourier transform, cepstrum analysis, and more (Gerhard, 2003). Once pitches are established for both stimulus and response, the size of phase-locking at those frequencies can be determined, as well as the extent to which the response pitch hews to the stimulus pitch, e.g., by determining root-mean-square error between them.

Phase consistency.

The strength of phaselocking at each time-frequency point is identified for each trial and normalized to a unit vector; the resultant vector length reflects the consistency of coding of time-frequency regions of interest such as the f0 and its harmonics.

Intrinsic Factors.

Two assessments of response “quality,” contra its encoding fidelity or strength, can be derived independent of timing, size, or other acoustically-driven characteristics. 1) Stability of response morphology on a trial-by-trial basis, assessed by inter-response correlation. 2) Magnitude of background noise level, i.e., an averaged response to silence, usually the time between successive stimulus presentations.

Distinction between responses.

Any of these FFR measures can be compared when evoked by the same stimulus under multiple conditions, such as a change in magnitude between monaural and binaural presentation or a change in timing between quiet and noise-masked presentation. Alternatively, the same measure may be assessed in contrasting speech or music sounds. In the latter case, the appropriate comparison is driven by the relevant acoustic contrasts of the stimuli, and may be in the time-, frequency- or time-frequency domains. For example, (1) a comparison of spectra might be made between responses elicited to the vowels /u/ and /e/. (2) Likewise, an investigation of consonant voicing-time between a /bo/ and a /po/ would rely on differences in the timing of transients in the response. (3) A cross-phase technique could be used to compare neural timing in response to contrasting consonant-vowel pairs in the time-frequency domain. The relative phase of the responses at frequencies corresponding to known acoustic differences between stop-consonant pairs is used to quantify the extent to which the nervous system discriminates them.

A quick word about techniques and terminology is in order before moving on. Briefly, the frequency following response to a pure tone is phase dependent. That is, the timing of the response is dependent on the phase of the stimulus. However, other components in the response, such as transient peaks, are phase independent. The FFR to more complex stimuli such as speech and music contains a mix of phase-dependent and phase-independent components. The phase-dependent and phase-independent components map onto fine-structure and envelope components of the stimulus, respectively. The latter also contains nonlinearities such as difference tones produced in the auditory system. The most widely used methodological technique to tease these two response components apart is to present the stimulus one-half of the time with inverted (i.e., pi, or 180 degrees out-of-phase) polarity. In this way, the envelope response, FFR_{env}, is constructed offline by adding the responses to the two stimulus types together; the temporal fine structure response, FFR_{tfs}, is a difference of the two response types. For a thorough treatment of the distinction in response properties of FFR_{env} and FFR_{tfs}, see (Aiken & Picton, 2008). Some researchers who focus on the envelope component alone refer to the FFR_{env} as the envelope following response, or EFR. Others, ourselves included, in the past, termed it cABR (auditory brainstem response to complex stimulation). All are the frequency following response.

Encoding of Speech Features by the Auditory Brainstem

The inferior colliculus has an upper limit in its frequency coding ability that prevents the FFR from being a perfect representation of the evoking stimulus. This speed limit in the IC has been reported to be anywhere from 1 to 4 kHz (Liu, Palmer, & Wallace, 2006; Ping, Li, Galbraith, Wu, & Li, 2008; Warrier, Abrams, Nicol, & Kraus, 2011). Nevertheless, the gamut of acoustic-phonetic events—the flapping of the vocal folds, the plosive bursts, the stops, and so forth—results in corresponding, if imperfect, copies in the form of neural firing patterns in the IC. In this section, we will address particular classes of the speech sound, how they are reflected in the FFR, and what we have learned about this coding with respect to external manifestations, in other words real-life correlates, of auditory processing ability.

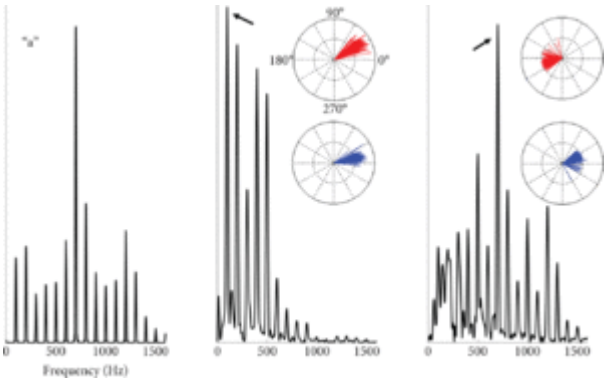
Vowels

The vowel sounds of speech are periodic in nature. Periodicity is imparted by vocal cord vibration, the rate of which is expressed perceptually as pitch in both global and local manifestations. A voice can be globally high or low on average, while roving higher or lower locally as emotion and intent vary over the course of running speech. Regardless of its pitch, this voicing sets up a harmonic series consisting of a fundamental frequency (f0) and integer harmonics, whereupon the shape and position of articulators (mouth, palate, tongue, nasal tract, etc.) cause selective filtering of the harmonic series. The resonances of the vocal tract, which appear as local peaks in the harmonic series, define the identity of the vowel. For example, if the first two energy concentrations (the first and second formants) are widely spaced at around 300 and 2500 Hz, the sound will typically be identifiable as an /i/ as in “beet.” More closely spaced formants, perhaps at 800 and 1500 Hz, would produce an /ae/ as in “bat.” These formant values hold for both high- and low-pitched voices; the approximate values given earlier will sound like /i/ and /ae/ whether spoken by a young child or a deep-voiced man. Both the identity of a vowel and the pitch at which it is spoken are coded in the FFR.

In music, there is a fairly direct parallel to the pitch and harmonic features in a vowel. Pitch has a one-to-one mapping in the two domains; the pitch of a vowel and the pitch of a note have an identical underpinning—the periodicity of the acoustic waveform. The shape of the harmonic series in music defines the identity of the instrument. Much like an /i/ and an /ae/ spoken by the same talker are readily distinguished, the same note played on an oboe and a flute are distinguishable—they have the same pitch with different emphases in the harmonic series.

Vowel pitch/fundamental frequency (f0)

The FFR to the fundamental frequency (or perhaps more accurately fundamental periodicity) of speech is an envelope component and, in fact, this aspect of the FFR is sometimes referred to as the "envelope following response." The fundamental frequency, by definition, is the lowest harmonic in a speech spectrum, but the f0 is often very small or even missing. In this respect the speech signal is analogous to an amplitude modulated signal, with the modulator being the f0 and the higher-frequency harmonics (and so, the formants) serving as the carrier. Modulation components are essentially phase invariant regardless of stimulus polarity and consequently, the f0 response, along with other non-spectral components, is best viewed as a summation of responses evoked by stimuli of opposing polarities (FFR_{env}) as reviewed earlier. Figure 1 shows the spectrum of a speech utterance /a/ and its FFR_{env} and FFR_{tfs}. In this example, the fundamental frequency of the syllable, at 100 Hz, is rather small in the speech signal (left) and essentially absent from the FFR_{tfs} (right). In contrast, the f0 dominates the FFR_{env} (center), along with some activity up to about 500 Hz that are distortion products of the first and second formants (i.e., 200 Hz = 2f1–f2; 500 Hz = f2–f1). The compass plots in the inset of the middle panel illustrate the relative phase *invariance* at 100 Hz in the FFR_{env}. In contrast, the phase *variance* of the harmonics of the fundamental is illustrated by the compass plot of 700 Hz activity in the FFR_{tfs} (inset, right).



Click to view larger

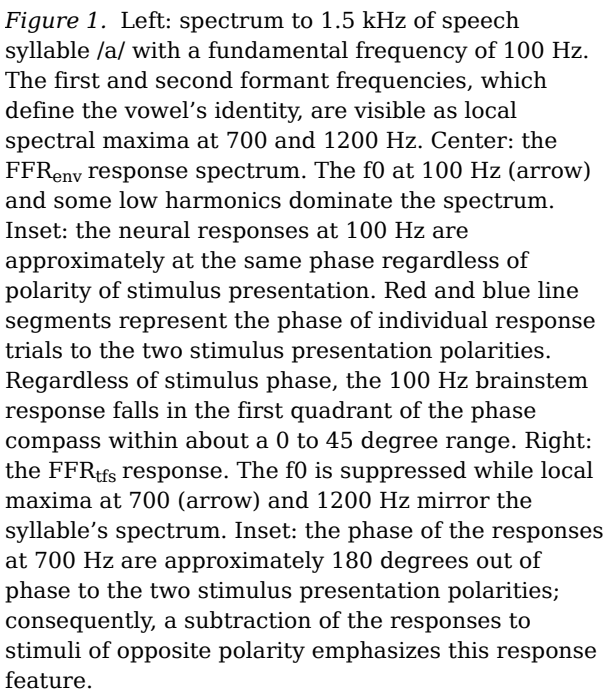


Figure 1. Left: spectrum to 1.5 kHz of speech syllable /a/ with a fundamental frequency of 100 Hz. The first and second formant frequencies, which define the vowel's identity, are visible as local spectral maxima at 700 and 1200 Hz. Center: the FFR_{env} response spectrum. The f0 at 100 Hz (arrow) and some low harmonics dominate the spectrum. Inset: the neural responses at 100 Hz are approximately at the same phase regardless of polarity of stimulus presentation. Red and blue line segments represent the phase of individual response trials to the two stimulus presentation polarities. Regardless of stimulus phase, the 100 Hz brainstem response falls in the first quadrant of the phase compass within about a 0 to 45 degree range. Right: the FFR_{tfs} response. The f0 is suppressed while local maxima at 700 (arrow) and 1200 Hz mirror the syllable's spectrum. Inset: the phase of the responses at 700 Hz are approximately 180 degrees out of phase to the two stimulus presentation polarities; consequently, a subtraction of the responses to stimuli of opposite polarity emphasizes this response feature.

Varieties of manipulated speech can be used in FFR research. For example, the encoding of a "missing fundamental" stimulus in the brainstem can be evaluated. A missing fundamental stimulus is one in which the spacing of the harmonics implies a particular fundamental frequency, but that frequency is absent from the stimulus. For example, a stimulus that contains spectral energy at 400, 500, 600, 700, … Hz has a percept of a 100 Hz pitch and the FFR will have a prominent 100 Hz peak—in fact it may be larger than the f0 response to an f0-only stimulus. Although investigations of missing-f0 with speech stimuli are relatively uncommon (Jeng, Costilow, Stangherlin, & Lin, 2011), tonal missing-f0 FFR studies (Galbraith, 1994; Greenberg, Marsh, Brown, & Smith, 1987) remain relevant to speech encoding because these stimuli are analogous to the spectral characteristics of speech signals, wherein the fundamental is often much smaller than the peak amplitudes of the lower formants. Another speech manipulation is vocoded speech. Often studied in normal listeners to simulate cochlear implant processing strategies, vocoded speech is devoid of fine spectral content, consisting of modulated noise bands. The encoding of vocoded speech has been investigated with the FFR (Ananthakrishnan, Luo, & Krishnan, 2017) showing an agreement between its encoding and its perception.

A consistent finding in FFR literature is that the encoding of the speech f0 is indicative of the ability to hear in noise. In normal hearing listeners of all ages, variability in the aptitude of understanding speech when masked by noise is mirrored in the amplitude of the f0 components in speech (Anderson, Parbery-Clark, Yi, & Kraus, 2011; Anderson, Skoe, Chandrasekaran, Zecker, & Kraus, 2010; Coffey, Chepesiuk, Herholz, Baillet, & Zatorre, 2017; J. Song, Skoe, Banai, & Kraus, 2011). Relatedly, bilingual speakers have particularly strong f0 encoding to speech (Krizman, Marian, Shook, Skoe, & Kraus, 2012), and the strength of phase-locking to the f0 in the FFR to a speech sound of a foreign language relates to the ability to learn that language (Omote, Jasmin, & Tierney, 2017). This form of linguistic enrichment hones attention and memory—skills that improve the sort of auditory object formation that is essential for understanding speech in a noisy environment—so it follows that a bilingual speaker would have this characteristic enhancement in the brainstem response. Another f0 finding has emerged wherein an

increased FFR amplitude is seen in older adults with mild cognitive impairment, suggesting an overcompensation in sensory processing to offset cognitive processing inefficiencies (Bidelman, Lowther, Tak, & Alain, 2017).

Training explicitly designed to bolster listening in noise also has a positive influence on the f0 response (Song, Skoe, Banai, & Kraus, 2012). These behavior-physiology relationships are maintained in the overall morphological similarity between evoking stimulus and FFR (Anderson et al., 2011; Cunningham, Nicol, Zecker, Bradlow, & Kraus, 2001; Kraus & White-Schwoch, 2016)—a metric that is influenced to a great extent by f0 encoding.

The physiological-behavioral relationship between f0 encoding and listening in noise is mirrored in individuals in clinical populations that are characterized by difficulty following speech in noise. An example is traumatic brain injury. People who have undergone sports-related concussions often complain of difficulty with sound processing (Gallun et al., 2012; Thompson et al., 2018) and their speech-evoked f0 is indeed diminished, even after the concussion has been judged resolved (Kraus et al., 2017; Kraus et al., 2016).

Vowels with changing pitch

There are constant changes in the pitch of natural speech. These changes can signal intention, attitude, emphasis, or emotion in all languages and, in some, they have linguistic and grammatical meaning as well. These speech attributes are known as intention and tone, respectively, and both can be measured in the brainstem response to speech. Using standard signal-processing techniques for pitch extraction, the contour of the fundamental frequency in the response can be compared to its counterpart in the evoking speech utterance.

Experience with language has a bearing on the accuracy of neural pitch tracking. Speakers of tonal languages (in contrast to speakers of languages where varying pitch signals intention only) have more accurate responses in terms hewing to the pitch of the stimulus (Krishnan, Xu, Gandour, & Cariani, 2005; Wong, Skoe, Russo, Dees, & Kraus, 2007; Yu & Zhang, 2018). Another form of auditory expertise—musicianship—also produces stronger pitch-tracking responses (Wong et al., 2007). On the other hand, even within tone-language speakers, a decline in FFR pitch tracking is seen in older adults (Wang et al., 2016). Russo et al. (Russo et al., 2008) noted weaker pitch tracking in school-age children on the autism spectrum compared to typically developing peers, evocative of the characteristic flat affect in the speech of some autistic individuals and the diminished ability to perceive emotional valence (Matsumoto et al., 2016). Imperfect perception of the contours of speech pitch may have a bearing on aberrant speech production.

Vowel identity (harmonics)

Unlike the FFR to the f0 of speech or music, the response to harmonic content of the stimulus is phase-dependent. So, as noted earlier, an appropriate manipulation must be made, the 180-degree inversion of the responses elicited by 180-degree out of phase stimuli, resulting in the FFR_{tfs}. An example spectrum of an FFR_{tfs} response to the vowel /a/ can be seen in the right panel of Figure 1. The spectrum matches well that of the evoking stimulus, especially past about 500 Hz, where clear local maxima at 700 and 1200 Hz, represent the first two formants of this particular /a/ vowel. Brainstem activity to contrasting vowels can be distinguished by spectral peaks in the FFRs (Aiken & Picton, 2008; Krishnan, 2002) and, when sonified (played audibly), responses to contrasting vowels can be distinguished (Weiss & Bidelman, 2015).

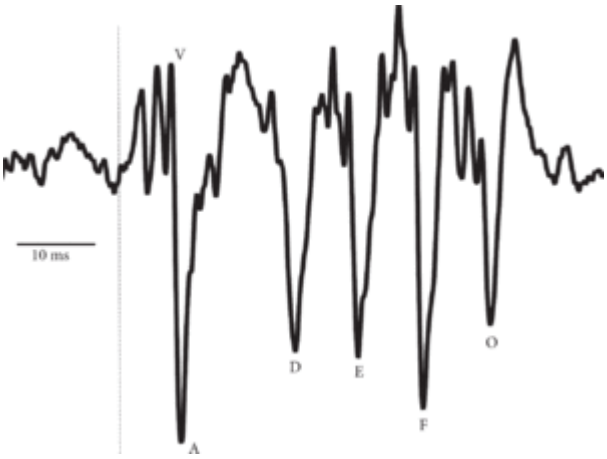
Of particular interest, from a clinical or diagnostic point of view, is not the ability to discern in the response the particular identity of the stimulus, but rather the strength of encoding of harmonics that are relevant to the identity of the syllable. There is a pattern that emerges in poor readers. Among other signs of response degradation, covered later in the chapter, individuals with dyslexia or who perform poorly in standardized measures of reading exhibit reduced harmonic encoding at frequencies corresponding to formant frequencies (Banai et al., 2009). Conversely, auditory enrichment via musical training boosts the harmonic representation of speech formants in the FFR (Strait & Kraus, 2014).

Consonants

Consonants are sometimes defined by what they are not—that is, they are the parts of speech that are not voiced like vowels, or at least not voiced with a stable spectrum. To be slightly more precise, they are partial or full obstructions of the breath that produces voicing in speech. Therefore, consonants are somewhat at odds with the nature of the frequency following response—the very name of which implies that there is a periodic component of the signal that is to be neurally followed. A response to a consonant in isolation (to the extent that such a thing exists) is a neural transient, in many respects like the brainstem response to a click that audiologists are familiar with. Nevertheless, it is possible to investigate consonant encoding in the brainstem using the FFR, even if in some cases we must loosen our definition of consonant to include its transition either to or from a conjoined fully voiced vowel.

Timing/transients.

Much speech FFR work, for better or worse, has focused on the consonant-vowel syllable "da." One in particular, a 40-ms variant, was chosen and created in our lab because it contains key acoustic elements important for speech perception, contains phonemes that are present in most world languages, and is short enough to be amenable to evoked response recording. This stimulus ships with many major clinical auditory evoked potential systems and has been widely used in research and, increasingly, in clinical settings. Norms are available for both the timing of well-characterized discrete peaks and for frequency-domain representation of features of the stimulus (Skoe, Krizman, Anderson, & Kraus, 2015). Although a very short utterance, during its 40-ms duration, the da progresses from a (relatively) broadband release burst, to a fully-voiced transition to a vowel. Figure 2 depicts a typical FFR_{env} to this short syllable. The response to the release burst (V and A), along with a response to the *offset* of the utterance (O), together can be viewed as qualitatively different than the intermediate peaks, D, E, and F, which arise from the periodicity of the stimulus f0 and will be discussed further in the next section.



Click to view larger

Figure 2. A response to a 40-ms da, with major peaks labeled. The vertical dashed line represents stimulus onset, time zero. Onset peak V (five) is analogous to wave V of the click-evoked auditory brainstem response. A is also considered an onset peak. D, E, and F are "following" the voicing of the syllable, at a period equivalent to the f0. O is an offset response.

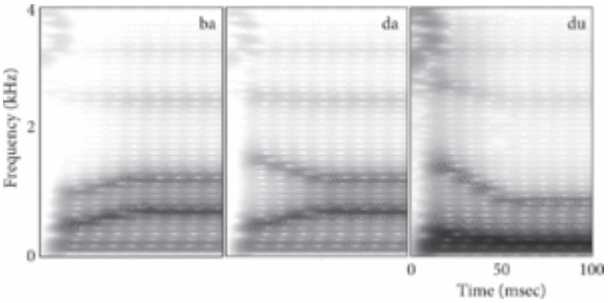
Some of the earliest findings equating the brainstem encoding of speech with language skills involved the timing of the onset transients of da. Cunningham et al. (Cunningham et al., 2001) noted that the timing of the V/A complex to a noise-masked da was delayed in children with a diagnosed reading disability. Later, a larger study relied on standardized tests of reading rather than external diagnoses. It found that discrete peaks elicited by the 40-ms da were consistently delayed in children (of any diagnosis) who scored poorly on reading assessments (Banai et al., 2009). Later investigations using the same stimulus confirmed this finding in children with reading impairment and extended it to auditory processing disorder (Rocha-Muniz, Befi-Lopes, & Schochat, 2012, 2014; White-Schwoch et al., 2015). In adults, on the other hand, the brainstem/reading relationship manifests itself differently, with earlier peak timing (Skoe, Brody, & Theodore, 2017). However, this is

consistent with a more juvenile-like response, as response latencies prolong between childhood and adulthood (Skoe et al., 2015).

An intriguing corollary to the transient timing delays in poor readers is found in musicians. Listening experts, namely musicians whose precisely-driven auditory systems have been tuned by years of focused attention to the emotion-laden auditory input of their instruments (Patel, 2011), have enhancements in brainstem encoding that are evocative of the degraded response properties seen in poor readers. Among other response enhancements, transient timing to speech stimulation is, in fact, earlier in musicians (Kraus & White-Schwoch, 2016; Parbery-Clark, Anderson, & Hittner, 2012; Parbery-Clark, Anderson, Hittner, & Kraus, 2012).

Temporal-spectral (frequency sweeps).

As noted previously, a stop consonant is a near timeless event; a short-duration closure of the articulators (i.e., tongue, lips) that rapidly leads to or follows from a voiced segment of speech. However, this unvoiced-to-voiced transition (or voiced-to-unvoiced in the case of vowel-consonant syllables) is an important phonetic event that is rich in acoustic interest. In the case of consonant-vowel (CV) utterances, such as “go” or “bee,” there is a distinct formant trajectory, visible in a spectrogram, that defines the identity of both the consonant and the vowel. Figure 3 illustrates spectrograms of three synthesized CV syllables, ba, da, and du. The first 2, ba and da conclude with the formants at the same frequencies, particularly well-depicted by dark F1 and F2 lines at about 700 and 1200 Hz, values typical for an “a” vowel. The difference between these two syllables is in the initial 50 ms—that is, the path the formants take to arrive at the shared steady-state vowel. F2 begins at about 900 Hz for ba and 1700 Hz for da, rising and falling, respectively, to 1200 Hz at 50 ms. In contrast, the origin frequency of F2 is 1700 Hz for both da and du, a frequency typical of the consonant “d,” but the syllables end at very different places, with “u” attaining steady-state conclusions of 250 and 870 Hz for F1 and F2, respectively.



Click to view larger

Figure 3. Spectrograms of three consonant-vowels syllables. The first two have a vowel in common, and so conclude with formants 1 and 2 at the same frequencies. The last two have a common consonant, so formants originate at the same frequencies, but conclude at frequencies prototypical of /a/ and /u/, respectively.

Different approaches are available to evaluate encoding of frequency-modulated formant glides. For one, there are transient peaks in the FFR during this period, such as D, E, and F, illustrated in Figure 2; these can be assessed for timing. Second, a spectrum over the formant transition can be produced and frequency encoding can be measured. Third,

spectrograms, as those shown for the stimuli in Figure 3, offer a time-frequency approach. Finally, “cross-phase” approach can evaluate frequency-specific timing differences between the responses to stimulus pairs such as the da and ba in Figure 3.

Using one or more of these techniques, several investigations into reading (dis)ability, reading subskills such as phonological awareness, or the ability to understand speech in noise have revealed deficits in the encoding of consonant-vowel formant transitions in speech (Anderson, Skoe, Chandrasekaran, & Kraus, 2010; Hornickel, Anderson, Skoe, Yi, & Kraus, 2012; Hornickel, Skoe, Nicol, Zecker, & Kraus, 2009; Skoe, Nicol, & Kraus, 2011; White-Schwoch & Kraus, 2013). Musicians, a model for auditory expertise, on the other hand, excel in this aspect of speech encoding (Parbery-Clark, Tierney, Strait, & Kraus, 2012), even after a year-long music training program (Kraus et al., 2014).

Other Speech Features

A remaining feature of speech that to date remains largely unexplored is that of voicing timing. A major acoustical cue that identifies stop consonants is the extent to which they are fully voiced (g, b, d) or unvoiced (k, p, t). The brainstem encoding of voicing timing and its behavioral consequences are ripe for investigation. Also, only beginning to be explored is speech at a more macro level. Ongoing speech, such as full sentences, presents practical problems because of the large number of stimulus presentations typically required for satisfactory FFR collection. Recently, investigations have been undertaken that look at concatenated-word or full-sentence stimulation (Choi, Purcell, Coyne, & Aiken, 2013; Reichenbach, Braiman, Schiff, Hudspeth, & Reichenbach, 2016), demonstrating that even with a reduced presentation count, structural components of the stimulus are maintained in the response, and that an attentional effect can be discerned in the FFR (Forte, Etard, & Reichenbach, 2017).

Conclusion

Our life in sound has a cumulative impact on the biological response to sound in the inferior colliculus of the brainstem. Engagement with sound activates the cognitive, sensorimotor, and reward networks of the brain, and the auditory midbrain represents a crossroads where these networks converge and gel. In the past 20 years, research into the response properties of the auditory midbrain have provided us with insight into the biological legacy of experience, such as music engagement and bilingualism. Further, the brainstem's response to speech and other complex stimulation such as music has a connection to communication skills, such as literacy and listening in noise. This growing recognition of the valuable insights that can come from probing the auditory brainstem have moved the needle on interest in pursuing the measurement of the auditory brainstem. FFR is beginning to emerge as a clinical tool on the strength of its capacity for discerning impairment in auditory processing as well as expertise. As researchers who have been involved in FFR to speech for nearly two decades, we are gratified that the slow process of divorcing the FFR from its cousin the auditory brainstem response (ABR) is progressing. FFR is no longer met with a shrug, viewed only as a modestly valuable alternative to ABR for peripheral auditory screening.

Brainstem processing of speech and music, though still a comparatively new line of inquiry, is gaining traction. Much of the pioneering work has been done in terms of developing best practices for the collection and analysis of speech- and music-evoked frequency following responses. Norms are being established as the developmental time course of the FFR is becoming more understood (Skoe et al., 2015; Van Dyke, Lieberman, Presacco, & Anderson, 2017) and the evaluation of an individual's response component relative to a normative reference is increasing in validity. It is gaining prominence among auditory neuroscientists: its mindshare at conferences is rapidly growing for at least two reasons. One, from a basic science perspective, it is a direct and accessible metric of human communication. We can learn so much more by accessing the auditory circuit's systemic response to a complex sound than can be achieved by probing single neurons in experimental animals with a clicks, pure tones or pulse trains. FFR affords us the luxury of complex stimulation and the direct applicability of accessibility in humans. Second, from a clinical standpoint, is its value in *objectively assessing the encoding of the fast components of speech in individuals*. Fast components of speech elude cortical auditory evoked responses which also suffer from only moderate individual reliability. And imaging techniques, though reliable in individuals, are also severely limited in terms of their ability to probe the fast components of speech. As the best pure measure of objective auditory processing available, FFR is poised for immediate and near-future applications in the clinical realm.

References

Aiken, S. J., & Picton, T. W. (2008). Envelope and spectral frequency-following responses to vowel sounds. *Hearing Research*, *245*(1–2), 35–47. doi:10.1016/j.heares.2008.08.004

Ananthakrishnan, S., Luo, X., & Krishnan, A. (2017). Human frequency following responses to vocoded speech. *Ear and Hearing*, *38*(5), e256–e267. doi:10.1097/AUD.0000000000000432

Anderson, S., Parbery-Clark, A., Yi, H. G., & Kraus, N. (2011). A neural basis of speech-in-noise perception in older adults. *Ear and Hearing*, *32*(6), 750–757. doi:10.1097/AUD.0b013e31822229d3

Anderson, S., Skoe, E., Chandrasekaran, B., & Kraus, N. (2010). Neural timing is linked to speech perception in noise. *Journal of Neuroscience*, *30*(14), 4922–4926. doi:10.1523/JNEUROSCI.0107-10.2010

Anderson, S., Skoe, E., Chandrasekaran, B., Zecker, S., & Kraus, N. (2010). Brainstem correlates of speech-in-noise perception in children. *Hearing Research*, *270*(1–2), 151–157. doi:10.1016/j.heares.2010.08.001

Banai, K., Hornickel, J. M., Skoe, E., Nicol, T., Zecker, S., & Kraus, N. (2009). Reading and subcortical auditory function. *Cerebral Cortex*, *19*(11), 2699–2707. doi:10.1093/cercor/bhp024

Batra, R., Kuwada, S., & Maher, V. L. (1986). The frequency-following response to continuous tones in humans. *Hearing Research*, *21*(2), 167–177.

Bidelman, G. M., Lowther, J. E., Tak, S. H., & Alain, C. (2017). Mild cognitive impairment is characterized by deficient brainstem and cortical representations of speech. *Journal of Neuroscience*, *37*(13), 3610–3620. doi:10.1523/JNEUROSCI.3700-16.2017

Boudreau, J. C. (1965). Neural volleying: Upper frequency limits detectable in the auditory system. *Nature*, *208*(5016), 1237–1238.

Chandrasekaran, B., & Kraus, N. (2010). The scalp-recorded brainstem response to speech: Neural origins and plasticity. *Psychophysiology*, *47*, 236–246. doi:10.1111/j.1469-8986.2009.00928.x

Choi, J. M., Purcell, D. W., Coyne, J. A., & Aiken, S. J. (2013). Envelope following responses elicited by English sentences. *Ear and Hearing*, *34*(5), 637–650. doi:10.1097/AUD.0b013e31828e4dad

Coffey, E. B. J., Chepesiuk, A. M. P., Herholz, S. C., Baillet, S., & Zatorre, R. J. (2017). Neural correlates of early sound encoding and their relationship to speech-in-noise perception. *Frontiers in Neuroscience*, *11*, 479. doi:10.3389/fnins.2017.00479

Cunningham, J., Nicol, T., Zecker, S. G., Bradlow, A., & Kraus, N. (2001). Neurobiologic responses to speech in noise in children with learning problems: Deficits and strategies for improvement. *Clinical Neurophysiology*, *112*, 758–767.

Forte, A. E., Etard, O., & Reichenbach, T. (2017). The human auditory brainstem response to running speech reveals a subcortical mechanism for selective attention. *Elife*, *6*. doi: 10.7554/eLife.27203

Galbraith, G. C. (1994). Two-channel brain-stem frequency-following responses to pure tone and missing fundamental stimuli. *Electroencephalography & Clinical Neurophysiology*, *92*(4), 321–330.

Galbraith, G. C., Arbagey, P. W., Branski, R., Comerci, N., & Rector, P. M. (1995). Intelligible speech encoded in the human brain stem frequency-following response. *Neuroreport*, *6*(17), 2363–2367.

Galbraith, G. C., Bhuta, S. M., Choate, A. K., Kitahara, J. M., & Mullen, T. A. (1998). Brain stem frequency-following response to dichotic vowels during attention. *Neuroreport*, *9*(8), 1889–1893.

Galbraith, G. C., Jhaveri, S. P., & Kuo, J. (1997). Speech-evoked brainstem frequency-following responses during verbal transformations due to word repetition. *Electroencephalography & Clinical Neurophysiology*, *102*(1), 46–53.

Gallun, F. J., Diedesch, A. C., Kubli, L. R., Walden, T. C., Folmer, R. L., Lewis, M. S., … Leek, M. R. (2012). Performance on tests of central auditory processing by individuals exposed to high-intensity blasts. *Journal of Rehabilitation Research and Development*, *49*(7), 1005–1025.

Gerhard, D. (2003). *Pitch extraction and fundamental frequency: history and current techniques*. Retrieved from University of Regina Technical Report, Regina, Saskatchewan, Canada: **http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.58.834**

Greenberg, S., Marsh, J. T., Brown, W. S., & Smith, J. C. (1987). Neural temporal coding of low pitch. I. Human frequency-following responses to complex tones. *Hearing Research*, *25*(2–3), 91–114.

Hornickel, J., Anderson, S., Skoe, E., Yi, H. G., & Kraus, N. (2012). Subcortical representation of speech fine structure relates to reading ability. *Neuroreport*, *23*(1), 6–9. doi:10.1097/WNR.0b013e32834d2ffd

Hornickel, J., Skoe, E., Nicol, T., Zecker, S., & Kraus, N. (2009). Subcortical differentiation of stop consonants relates to reading and speech-in-noise perception. *Proceedings of the National Academy of Sciences USA*, *106*(31), 13022–13027. doi:10.1073/pnas.0901123106

Jeng, F. C., Costilow, C. E., Stangherlin, D. P., & Lin, C. D. (2011). Relative power of harmonics in human frequency-following responses associated with voice pitch in American and Chinese adults. *Perceptual and Motor Skills*, *113*(1), 67–86.

King, A., Hopkins, K., & Plack, C. J. (2016). Differential group delay of the frequency following response measured vertically and horizontally. *Journal of the Association for Research in Otolaryngology*, *17*(2), 133–143. doi:10.1007/s10162-016-0556-x

Kraus, N., Lindley, T., Colegrove, D., Krizman, J., Otto-Meyer, S., Thompson, E. C., & White-Schwoch, T. (2017). The neural legacy of a single concussion. *Neuroscience Letters*, *646*, 21–23. doi:10.1016/j.neulet.2017.03.008

Kraus, N., Slater, J., Thompson, E., Hornickel, J., Strait, D., Nicol, T., & White-Schwoch, T. (2014). Auditory learning through active engagement with sound: Biological impact of community music lessons in at-risk children. *Frontiers in Neuroscience*, *8*, 351. doi: 10.3389/fnins.2014.00351

Kraus, N., Thompson, E. C., Krizman, J., Cook, K., White-Schwoch, T., & LaBella, C. R. (2016). Auditory biological marker of concussion in children. *Scientific Reports*, *6*, 39009. doi:10.1038/srep39009

Kraus, N., & White-Schwoch, T. (2015). Unraveling the biology of auditory learning: A cognitive-sensorimotor-reward framework. *Trends in Cognitive Sciences*, *19*(11), 642–654. doi:10.1016/j.tics.2015.08.017

Kraus, N., & White-Schwoch, T. (2016). Neurobiology of everyday communication: What have we learned from music? *Neuroscientist*, *23*(3), 287–298. doi: 10.1177/1073858416653593

Krishnan, A. (2002). Human frequency-following responses: Representation of steady-state synthetic vowels. *Hearing Research*, *166*(1–2), 192–201. doi:10.1016/S0378-5955(02)00327-1

Krishnan, A., Xu, Y. S., Gandour, J., & Cariani, P. (2005). Encoding of pitch in the human brainstem is sensitive to language experience. *Cognitive Brain Research*, *25*(1), 161–168.

Krizman, J., Marian, V., Shook, A., Skoe, E., & Kraus, N. (2012). Subcortical encoding of sound is enhanced in bilinguals and relates to executive function advantages. *Proceedings of the National Academy of Sciences USA*, *109*(20), 7877–7881. doi:10.1073/pnas.1201575109

Liu, L. F., Palmer, A. R., & Wallace, M. N. (2006). Phase-locked responses to pure tones in the inferior colliculus. *Journal of Neurophysiology*, *95*(3), 1926–1935. doi:10.1152/jn.00497.2005

Marsh, J. T., Worden, F. G., & Smith, J. C. (1970). Auditory frequency-following response: Neural or artifact? *Science*, *169*(3951), 1222–1223.

Martinez, A. S., Eisenberg, L. S., & Boothroyd, A. (2013). The acoustic change complex in young children with hearing loss: A preliminary study. *Seminars in Hearing*, *34*(4), 278–287. doi:10.1055/s-0033-1356640

Matsumoto, K., Sugiyama, T., Saito, C., Kato, S., Kuriyama, K., Kanemoto, K., & Nakamura, A. (2016). Behavioral study on emotional voice perception in children with autism spectrum disorder. *Journal of Pediatric Neuropsychology*, *2*(3–4), 108–118. doi: 10.1007/s40817-016-0021-0

Mines, M. A., Hanson, B. F., & Shoup, J. E. (1978). Frequency of occurrence of phonemes in conversational English. *Language and Speech*, *21*(3), 221–241. doi: 10.1177/002383097802100302

Moushegian, G., Rupert, A., & Whitcomb, M. A. (1964). Brain-stem neuronal response patterns to monaural and binaural tones. *Journal of Neurophysiology*, *27*, 1174–1191.

Omote, A., Jasmin, K., & Tierney, A. (2017). Successful non-native speech perception is linked to frequency following response phase consistency. *Cortex*, *93*, 146–154. doi: 10.1016/j.cortex.2017.05.005

Parbery-Clark, A., Anderson, S., & Hittner, E. (2012). Musical experience offsets age-related delays in neural timing. *Neurobiology of Aging*, *33*(7), 1483.e1–1483.e4. doi: 10.1016/j.neurobiolaging.2011.12.015

Parbery-Clark, A., Anderson, S., Hittner, E., & Kraus, N. (2012). Musical experience strengthens the neural representation of sounds important for communication in middle-aged adults. *Frontiers in Aging Neuroscience*, *4*(30), 1–12. doi:10.3389/fnagi.2012.00030

Parbery-Clark, A., Tierney, A., Strait, D., & Kraus, N. (2012). Musicians have fine-tuned neural distinction of speech syllables. *Neuroscience*, *219*, 111–119. doi:10.1016/j.neuroscience.2012.05.042

Patel, A. D. (2011). Why would musical training benefit the neural encoding of speech? The OPERA hypothesis. *Frontiers in Psychology*, *2*, 142. doi:10.3389/fpsyg.2011.00142

Ping, J., Li, N., Galbraith, G. C., Wu, X., & Li, L. (2008). Auditory frequency-following responses in rat ipsilateral inferior colliculus. *Neuroreport*, *19*(14), 1377–1380. doi: 10.1097/WNR.0b013e32830c1cfa

Reichenbach, C. S., Braiman, C., Schiff, N. D., Hudspeth, A. J., & Reichenbach, T. (2016). The auditory-brainstem response to continuous, non-repetitive speech is modulated by the speech envelope and reflects speech processing. *Frontiers in Computational Neuroscience*, *10*, 47. doi:10.3389/fncom.2016.00047

Rocha-Muniz, C. N., Befi-Lopes, D. M., & Schochat, E. (2012). Investigation of auditory processing disorder and language impairment using the speech-evoked auditory

brainstem response. *Hearing Research*, *294*(1), 143–152. doi:10.1016/j.heares.2012.08.008

Rocha-Muniz, C. N., Befi-Lopes, D. M., & Schochat, E. (2014). Sensitivity, specificity and efficiency of speech-evoked ABR. *Hearing Research*, *317*, 15–22. doi:10.1016/j.heares.2014.09.004

Russo, N. M., Skoe, E., Trommer, B., Nicol, T., Zecker, S., Bradlow, A., & Kraus, N. (2008). Deficient brainstem encoding of pitch in children with autism spectrum disorders. *Clinical Neurophysiology*, *119*(8), 1720–1731. doi:10.1016/j.clinph.2008.01.108

Skoe, E., Brody, L., & Theodore, R. M. (2017). Reading ability reflects individual differences in auditory brainstem function, even into adulthood. *Brain and Language*, *164*, 25–31. doi:10.1016/j.bandl.2016.09.003

Skoe, E., Krizman, J., Anderson, S., & Kraus, N. (2015). Stability and plasticity of auditory brainstem function across the lifespan. *Cerebral Cortex*, *25*(6), 1415–1426. doi:10.1093/cercor/bht311

Skoe, E., Nicol, T., & Kraus, N. (2011). Cross-phaseogram: Objective neural index of speech sound differentiation. *Journal of Neuroscience Methods*, *196*(2), 308–317. doi:10.1016/j.jneumeth.2011.01.020

Song, J., Skoe, E., Banai, K., & Kraus, N. (2011). Perception of speech in noise: Neural correlates. *Journal of Cognitive Neuroscience*, *23*(9), 2268–2279. doi:10.1162/jocn.2010.21556

Song, J. H., Skoe, E., Banai, K., & Kraus, N. (2012). Training to improve hearing speech in noise: Biological mechanisms. *Cerebral Cortex*, *22*(5), 1180–1190. doi:10.1093/cercor/bhr196

Strait, D. L., & Kraus, N. (2014). Biological impact of auditory expertise across the life span: Musicians as a model of auditory learning. *Hearing Research*, *308*, 109–121. doi:10.1016/j.heares.2013.08.004

Thompson, E. C., Krizman, J., White-Schwoch, T., Nicol, T., LaBella, C. R., & Kraus, N. (2018). Difficulty hearing in noise: A sequela of concussion in children. *Brain Injury*, *32*(6), 763–769. doi:10.1080/02699052.2018.1447686

Van Dyke, K. B., Lieberman, R., Presacco, A., & Anderson, S. (2017). Development of phase locking and frequency representation in the infant frequency-following response. *Journal of Speech, Language, and Hearing Research*, *60*, 2740–2751. doi:10.1044/2017_JSLHR-H-16-0263

Wang, S., Hu, J., Dong, R., Liu, D., Chen, J., Musacchia, G., & Liu, B. (2016). Voice pitch elicited frequency following response in Chinese elderlies. *Frontiers in Aging Neuroscience*, *8*, 286. doi:10.3389/fnagi.2016.00286

Warrier, C. M., Abrams, D. A., Nicol, T. G., & Kraus, N. (2011). Inferior colliculus contributions to phase encoding of stop consonants in an animal model. *Hearing Research*, *282*(1–2), 108–118. doi:10.1016/j.heares.2011.09.001

Weiss, M. W., & Bidelman, G. M. (2015). Listening to the brainstem: Musicianship enhances intelligibility of subcortical representations for speech. *Journal of Neuroscience*, *35*(4), 1687–1691. doi:10.1523/jneurosci.3680-14.2015

White-Schwoch, T., & Kraus, N. (2013). Physiologic discrimination of stop consonants relates to phonological skills in pre-readers: A biomarker for subsequent reading ability? *Frontiers in Human Neuroscience*, *7*, 899. doi:10.3389/fnhum.2013.00899

White-Schwoch, T., Woodruff Carr, K., Thompson, E. C., Anderson, S., Nicol, T., Bradlow, A. R., … Kraus, N. (2015). Auditory processing in noise: A preschool biomarker for literacy. *PLoS Biology*, *13*(7), e1002196. doi:10.1371/journal.pbio.1002196

Winer, J. A. (2006). Decoding the auditory corticofugal systems. *Hearing Research*, *212*(1–2), 1–8.

Wong, P. C. M., Skoe, E., Russo, N. M., Dees, T., & Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nature Neuroscience*, *10*(4), 420–422.

Worden, F. G., & Marsh, J. T. (1968). Frequency-following (microphonic-like) neural responses evoked by sound. *Electroencephalography & Clinical Neurophysiology*, *25*(1), 42–52.

Yu, L., & Zhang, Y. (2018). Testing native language neural commitment at the brainstem level: A cross-linguistic investigation of the association between frequency-following response and speech perception. *Neuropsychologia*, *109*, 140–148. doi:10.1016/j.neuropsychologia.2017.12.022

Nina Kraus

Nina Kraus, Northwestern University

Trent Nicol

Trent Nicol, Northwestern University