HHS Public Access

Author manuscript

Brain Res. Author manuscript; available in PMC 2020 January 01.

Published in final edited form as:

Brain Res. 2019 July 01; 1714: 182–192. doi:10.1016/j.brainres.2019.02.025.

Brainstem correlates of concurrent speech identification in adverse listening conditions

Anusha Yellamsetty[a,b,*], **Gavin M. Bidelman**[a,c,d]

[a]School of Communication Sciences & Disorders, University of Memphis, Memphis, TN, USA

[b]Department of Communication Sciences & Disorders, University of South Florida, USA

[c]Institute for Intelligent Systems, University of Memphis, Memphis, TN, USA

[d]University of Tennessee Health Sciences Center, Department of Anatomy and Neurobiology, Memphis, TN, USA

Abstract

When two voices compete, listeners can segregate and identify concurrent speech sounds using pitch (fundamental frequency, F0) and timbre (harmonic) cues. Speech perception is also hindered by the signal-to-noise ratio (SNR). How clear and degraded concurrent speech sounds are represented at early, pre-attentive stages of the auditory system is not well understood. To this end, we measured scalp-recorded frequency-following responses (FFR) from the EEG while human listeners heard two concurrently presented, steady-state (time-invariant) vowels whose F0 differed by zero or four semitones (ST) presented diotically in either clean (no noise) or noise-degraded (+ 5dB SNR) conditions. Listeners also performed a speeded double vowel identification task in which they were required to identify both vowels correctly. Behavioral results showed that speech identification accuracy increased with F0 differences between vowels, and this perceptual F0 benefit was larger for clean compared to noise degraded (+ 5dB SNR) stimuli. Neurophysiological data demonstrated more robust FFR F0 amplitudes for single compared to double vowels and considerably weaker responses in noise. F0 amplitudes showed speech-on-speech masking effects, along with a non-linear constructive interference at 0ST, and suppression effects at 4ST. Correlations showed that FFR F0 amplitudes failed to predict listeners' identification accuracy. In contrast, FFR F1 amplitudes were associated with faster reaction times, although this correlation was limited to noise conditions. The limited number of brain-behavior associations suggests subcortical activity mainly reflects exogenous processing rather than perceptual correlates of concurrent speech perception. Collectively, our results demonstrate that

*Corresponding author at: Department of Communication Sciences & Disorders, University of South Florida, 4202 E. Fowler Ave, PCD 1017, Tampa, FL 32620, USA. yellamsettya@usf.edu (A. Yellamsetty), gmbdlman@memphis.edu (G.M. Bidelman).

Reprints and requests for materials should be directed to G.M.B. [gmbdlman@memphis.edu].

[1]While single polarity stimulus presentation does not entirely preclude the possibility cochlear microphonic (CM) pickup in our recordings, such preneural contributions are likely minimal here since FFRs show a characteristics delay (< 10 ms; see Fig. 3) whereas CM is coincident with the stimulus (i.e., 0 ms latency) (Chimento and Schreiner, 1990). More importantly, fixed presentation allowed us to record FFRs coding both the envelope and fine-structure of speech, which would be lost using alternating polarity (Aiken and Picton, 2008).

Conflict of interest

The authors have no financial or commercial relationships based on the research reported in this paper.

FFRs reflect pre-attentive coding of concurrent auditory stimuli that only weakly predict the success of identifying concurrent speech.

Keywords

FFR; Double-vowel identification; Speech-in-noise perception

1. Introduction

A fundamental phenomenon in human hearing is the ability to parse co-occurring auditory objects (e.g., different voices) to extract the intended message of a target signal. Psychophysical and neurophysiological studies have shown that listeners can use multiple cues to distinguish simultaneous sounds. The segregation of a complex auditory mixture is thought to involve a multistage hierarchy of processing, whereby initial pre-attentive processes that partition the sound waveform into distinct acoustic features (e.g., pitch, harmonicity) are followed by later, post-perceptual principles (Koffka, 1935) (e.g., grouping by physical similarity, temporal proximity, good continuity (Bregman, 1990) and phonetic template matching (Alain et al., 2005a; Meddis and Hewitt, 1992). Psychophysical research from the past several decades confirms that human listeners exploit fundamental frequency (F0) differences (i.e., pitch) to segregate concurrent speech (Arehart et al., 1997; Assmann and Summerfield, 1989; Assmann and Summerfield, 1990; Assmann and Summerfield, 1994; Chintanpalli et al., 2016; de Cheveigne et al., 1997). For example, when two steady-state (time-invariant) synthetic vowels are presented simultaneously to the same ear, listeners' identification accuracy increases when a difference of four semitones(STs) is introduced between vowel F0s (Assmann and Summerfield, 1989; Assmann and Summerfield, 1990; Assmann and Summerfield, 1994; Culling, 1990; McKeown, 1992; Scheffers, 1983; Zwicker, 1984). This improvement is referred to as the "F0-benefit" (Arehart et al., 1997; Bidelman and Yellamsetty, 2017; Chintanpalli et al., 2014; Chintanpalli and Heinz, 2013; Chintanpalli et al., 2016; Yellamsetty and Bidelman, 2018).

To understand the time course of neural processing underlying concurrent speech segregation most investigations have quantified how various acoustic cues including harmonics, spatial location, and onset asynchrony affect perceptual segregation (Alain, 2007b; Carlyon, 2004). However, most neuroimaging studies have been concerned with the *cortical* representations/correlates of concurrent speech perception (Alain et al., 2005b; Bidelman, 2015a; Bidelman and Yellamsetty, 2017; Dyson and Alain, 2004; Yellamsetty and Bidelman, 2018). In contrast, the *subcortical* neural underpinnings have been studied only in animals (Jane and Young, 2000; Palmer and Winter, 1992; Reale and Geisler, 1980; Sinex et al., 2002a; Sinex et al., 2002b; Sinex et al., 2005; Sinex, 2008; Tan and Carney, 2005). Studies that directly examined the F0 representations of concurrent complex tones in auditory nerve (AN) and cochlear nucleus (CN) neurons showed the temporal discharge pattern and spatial distribution of responses contain sufficient information to identify both F0s (Jane and Young, 2000; Keilson et al., 1997; Palmer, 1990; Palmer and Winter, 1992; Sinex, 2008; Tan and Carney, 2005). The same is observed for double vowel speech stimuli (Keilson et al., 1997; Palmer, 1990; Palmer and Winter, 1992). In addition, AN single-unit

population studies have shown neural phase-locking is a primary basis for encoding the tonal features (e.g., F0) of vowels (Reale and Geisler, 1980; Tan and Carney, 2005) and that different sets of neurons are involved in encoding the first and second formants of speech (Miller et al., 1997). Whereas at the level of the inferior colliculus (IC), responses are tuned to low-frequency amplitude fluctuations (Bidelman and Alain 2015; Sinex et al., 2002a; Sinex et al., 2002b; Sinex et al., 2005; Sinex, 2008), providing a robust neural code for both F0 periodicity and the spectral peaks (i.e., formants) that listeners use to separate and identify vowels (Carney et al., 2015; Henry et al., 2017). These temporal discharge patterns are closely related to the auto-correlation model of pitch extraction (Meddis and Hewitt, 1992b) that accounts for the encoding of single and multiple F0s as early as the level of AN (Cariani and Delgutte, 1996; Cedolin and Delgutte, 2005; Meddis and Hewitt, 1992b). It appears that stimulus harmonicity/periodicity (F0) are coded very early in the auditory system and remain largely untransformed in the phase-locked activity of the rostral brainstem (Bidelman and Alain, 2015). Thus, evoked potentials, which measure phase-locked brainstem activity, could offer a window into how sub-cortical regions of the *human* brain encode concurrent sounds, including those based on F0-segregation (i.e., double-vowel mixtures).

In the present study, we used the scalp-recorded human frequency-following response (FFR), which reflects sustained phase-locked activity dominantly from the rostral brainstem (Bidelman, 2018; Glaser et al., 1976; Marsh et al., 1974; Smith et al., 1975; Worden and Marsh, 1968), to measure concurrent sound processing. FFRs can reproduce frequencies of periodic acoustic stimuli below approximately 1500 Hz (Bidelman and Powers, 2018; Gardi et al., 1979; Stillman et al., 1978) and code important properties of speech stimuli such as voice F0 (Bidelman et al., 2011; Krishnan et al., 2010) and several lower speech harmonics/formants (Bidelman, 2015b; Chandrasekaran and Kraus, 2010; Krishnan, 1999; Krishnan and Agrawal, 2010; Krishnan, 2002). FFRs allowed us to estimate how salient properties of speech spectra (e.g., F0s or formants of concurrent vowels) are transcribed by the human auditory nervous system at early, pre-attentive stages of the processing hierarchy.

In addition, FFRs have provided critical insight toward understanding the neurobiological encoding of degraded speech from a subcortical perspective (Anderson et al., 2010a; Bidelman and Krishnan, 2010; Bidelman, 2017; Parbery-Clark et al., 2009b; Song et al., 2011). Speech perception in noise is related to the subcortical encoding of F0 and timbre (Bidelman and Krishnan, 2010; Bidelman, 2016; Song et al., 2011) as well as the effectiveness of the nervous system to extract regularities in speech sounds related to vocal pitch (Chandrasekaran et al., 2009; Xie et al., 2017). Resilience of the FFR at F0 (but not its higher harmonics or onset) in the presence of noise has been noted by a number of investigators (Bidelman and Krishnan, 2010; Li and Jeng, 2011; Prévost et al., 2013; Russo et al., 2004) and suggests that neural synchronization at the fundamental F0 periodicity is relatively robust to acoustic interference [for review, see (Bidelman, 2017)]—at least for *single* speech tokens presented in isolation.

Given its high spectro-temporal fidelity, we reasoned that neural correlates relevant to double vowel identification may be substantiated in nascent signal processing along the auditory pathway, even earlier than documented in cerebral cortex (Alain et al., 2005a; Alain

et al., 2017; Bidelman and Yellamsetty, 2017; Yellamsetty and Bidelman, 2018). We aimed to test this hypothesis by analyzing the spectral response patterns of the single and double vowel FFRs when speech sounds did and did not contain F0 cues (0ST vs. 4ST). Additionally, we examined concurrent vowel processing in different levels of noise interference (quiet vs. +5 dB SNR) to evaluate how the neural encoding of spectro-temporal cues is affected by noise at a subcortical level. Despite ample FFR studies using isolated speech sounds (e.g., vowels, stop consonants) (Al Osman et al., 2017; Anderson and Kraus, 2010; Bidelman and Krishnan, 2010; Hornickel et al., 2009; Krishnan, 2002; Parbery-Clark et al., 2009b), to our knowledge, this is the first to examine brainstem encoding of concurrent *speech mixtures* in human auditory system using FFRs.

Here, we sought to determine (1) how concurrent vowels are encoded at pre-attentive, subcortical levels of the auditory system; (2) characterize effects of noise on the neural encoding of voice pitch and timbre (i.e., formant) cues in concurrent speech; and (3) assess the relation between passively evoked (pre-attentive) brainstem neural activity and behavioral concurrent vowel identification in quiet and degraded listening conditions. To this end, we recorded neuroelectric responses as listeners passively heard double-vowel pair and single vowel stimuli (Fig. 1). Stimulus manipulations were designed to promote (increase) or deny (reduce) successful identification (i.e., changes in F0 separation of vowels; with/without noise masking). We expected the spectral components of FFRs to reflect the encoding of non-linear interactions between the two concurrent vowels, such that responses would differ with and without pitch cues in a constructive and suppressive manner. Additionally, we hypothesized FFRs would show reduced amplitudes with noise and correlate with behavioral identification scores, offering an objective, subcortical correlates of concurrent speech perception.

2. Results

2.1. Behavioral data

Behavioral speech identification accuracy and RTs for double-vowel identification are shown in Fig. 2. Listeners obtained near-ceiling performance (97.9 ± 1.4%) when identifying single vowels. In contrast, double-vowel identification was considerably more challenging; listeners' accuracy ranged from ~45 to 70% depending on the presence of noise and pitch cues (Fig. 2A). An ANOVA conducted on behavioral accuracy confirmed a significant SNR × F0 interaction [$F_{1, 45} = 5.65$, p = 0.0218], indicating that successful double-vowel identification depended on both noise and F0 pitch cues. Performance increased ~ 30% across the board with greater F0 separations (i.e., 4ST > 0ST). F0-benefit was larger for clean relative to +5dB SNR speech [$t_{15} = -6.49$, p < 0.0001 (one-tailed)], suggesting they were more successful using pitch cues when segregating clean compared to noisy speech.

Analysis of reaction times (RTs) revealed a significant effect of SNR [$F_{1, 45} = 16.23$, $p = 0.0002$] and ST [$F_{1, 45} = 7.48$, $p = 0.0089$]; listeners tended to be slower identifying clean compared to noisy speech (Fig. 2B). The slowing of RTs coupled with better %-identification for clean compared to noise-degraded speech indicates a time-accuracy

tradeoff in speech perception (Bidelman and Yellamsetty, 2017; Yellamsetty and Bidelman, 2018).

2.2. FFR responses to single and double vowels

Grand average FFR waveforms and spectra are shown for each vowel type (single, double vowels), SNRs (clean, noise), and semitones (0 ST, 4 ST) conditions in Fig. 3A and B. FFRs showed phase-locked energy corresponding to the periodicities of the acoustic speech signals. Comparisons across conditions suggested more robust encoding of single and double vowels in the 0ST condition. Responses were weaker for conditions with 4ST and in noise. Response spectra contained energy at the F0 and the integer-related multiples up to the upper limit of the brainstem phase locking (~1100 Hz) (Liu et al., 2006). Strong FFRs at the F0s were consistent with the stimulus autocorrelation functions of single and double vowels which similarly showed a peak at the delay (time lag) corresponding to the F0 periodicities of each vowel (see Fig. 1).

Quantification of FFR F0 (pitch) and F1 (timbre) coding of single and double vowels at 0 $ST_{(/a+ɛ/150)}$ and 4 $ST_{(/a/150, /ɛ/190)}$ are shown in Fig. 3C. We first evaluated the effects of having multiple vs. single vowels and the effects of noise on FFRs. A two-way mixed model ANOVA with stimulus type (2 levels: single and double vowel) and SNR (2 levels: clean and +5dB SNR) as fixed factors (subjects = random effect) revealed that F0 amplitudes of the single-vowels were more robust than in double-vowels (single > double) [$F_{1,141} = 16.02$, $p < 0.0001$]. Responses were also stronger for double-vowels without pitch cues (i.e., 0 ST > 4 ST) revealing a super-additive effect at F0 (i.e., common F0 between vowels sum constructively in the FFR). This additive effect was less than doubling of acoustic energy, suggesting non-linearity of the response. Noise-related reductions in F1 amplitudes were larger for double compared to single-vowels [$F_{1,141} = 89.11$, $p < 0.0001$].

Next, we evaluated the impact of noise and pitch cues on *doublevowel* FFRs. Both additive and masking effects were observed at 4 ST. An ANOVA conducted on F0 amplitudes showed significant effects of SNR [$F_{1,77} = 31.66$; $p < 0.0001$] and ST [$F_{1,77} = 5.67$; $p = 0.0198$] with an interaction of SNR × ST [$F_{1,77} = 10.39$; $p = 0.0019$]. In contrast, for the neural encoding of F1, we found significant effects of ST [$F_{1,77} = 138.15$; $p < 0.0001$] and SNR [$F_{1,77} = 15.09$; $p = 0.0002$] but no interaction [$F_{1,77} = 1.42$; $p = 0.236$]. Noise-related changes in F1 were greater at 0 ST compared to 4 ST.

To quantify speech-on-speech masking effects in the FFR from having two vs. one vowel we assessed differences between responses to actual double vowel mixtures (i.e., $0ST_{(/a + ɛ/150)}$ and 4 $ST_{(/a/150+/ɛ/190)}$) and those evoked by the summed responses to the individual vowel constituents [e.g., is $FFR_{/a+ɛ/} \geq FFR_{/a/+/ɛ/}$] (Fig. 4). The rationale of this analysis is that when multiple speech components fall within the same auditory filter band (e.g., 0ST condition), this can result in speech-on-speech masking. The amplitude difference reflects the degree of speech-on-speech masking or mutual suppression from having two vowels in double vowel pairs. Speech-on-speech masking effects were observed in both clean ($t_{15} = 2.81$; $p = 0.0132$) and noise ($t_{15} = 3.46$, $p = 0.0035$) conditions. Suppression-like effects were observed in 4ST (in addition to speech-on-speech masking) resulting in further reduction in amplitude in both clean ($t_{15} = -3.97$; $p = 0.001$) and noise ($t_{15} = -2.36$; $p =$

0.0325). These effects were not observed at F1 ($ps \gg 0.05$). The effect of speech-to-noise (i.e., FFR amplitudes of clean vs. noise) was greater than the speech-on-speech masking (single vs. double) at F0 and F1 [$F_{1,140} = 30.85$; $p < 0.0001$; $F_{1,140} = 275.31$; $p < 0.0001$]. These differences indicate that FFRs to concurrent speech stimuli were systematically different than their single vowel counterparts, which also varied as a function of frequency component (i.e., F0, F1) and SNR.

2.3. Brain-behavior relationships

2.3.1. Regression analyses—Linear regressions between FFR F0 amplitudes and behavioral accuracy (%)—aggregating both ST conditions—are shown in Fig. 5A for the clean and noise conditions. Correlations between FFR F1 and behavioral RTs are shown in Fig. 5B. We chose these analyses based on previous literature showing robust correlations between (i) FFR F0 and accuracy (Anderson et al., 2010a; Anderson et al., 2012; Bidelman and Krishnan, 2010; Coffey et al., 2017; Du et al., 2011) and (ii) FFR F1 and RTs (Bidelman et al., 2014a; Bidelman et al., 2014b) in various speech perception tasks. These analyses revealed F1 amplitude was associated with RTs in the noise condition ($R^2 = 0.10$, p = 0.0277). No other correlations reached significance.

2.3.2. Vowel dominancy analysis—As an alternate approach to investigate possible relations between subcortical coding and behavioral identification of concurrent vowels we assessed whether listeners' tendency to report one or another vowel in a speech mixture depended on their FFR. We reasoned that the relative strength of each single vowel in their double-vowel response might drive which vowel was more perceptually dominant. To quantify the relative weighting of each vowel in the FFR we carried out response-to-response Pearson's correlations between each listener's (individual) single-vowel FFR spectra (FFR_a, FFR_ε) and their double-vowel response spectrum ($FFR_{a+\varepsilon}$). We restricted this analysis to the 4 ST clean condition, as this reflected the best behavioral identification (see Fig. 2). This analysis therefore assessed the degree to which listeners' FFR to a double-vowel mixture more closely resembled a response to either /a/ or /ɛ/.

Listeners were then median split based on the counts of the highest and lowest 50% of the cohort reporting /a/ in the behavioral identification task. Similarly, we determined the highest and lowest /ɛ/ reporters who dominantly heard /ɛ/ in /a + ɛ/mixtures. We then conducted a two-way ANOVA on response-to-response correlations with factors group vs. vowel. Fig. 6 shows the response-to-response correlations with the sample split by their behavioral bias. Comparing the relative strength of response-to-response correlations, double-vowel FFRs showed better correspondence to /ɛ/ than /a/ overall. We also found a vowel × group interaction ($F_{1,14} = 4.81$; $p = 0.0457$). Even though there was a significant difference in reporting /a/ vs. /ɛ/ vowels ($F_{1,14} = 42.89$; $p < 0.0001$) in /a + ɛ/ mixtures, FFRs more closely resembled the /ɛ/ response regardless of listeners' behavior, counter to our hypothesis.

3. Discussion

The present study measured FFRs to double vowel stimuli that varied in their voice pitch (F0 separation) and noise level (SNR). Our results showed three primary findings: (i)

behaviorally, listeners exploit F0-differences between vowels to identify speech, and the perceptual F0 benefits degrade with noise; (ii) FFRs amplitudes for dual speech stimuli are altered in a systematic manner from their single vowel counterparts as a function of frequency component (i.e., F0, F1) and noise (SNR); (iii) FFRs predict perceptual speed but not the accuracy of double vowel identification, but only in noisy listening conditions.

3.1. Effects of SNR and F0 cues on behavioral concurrent vowel identification

The effects of F0 on concurrent vowel identification were comparable and consistent with previous data (Arehart et al., 1997; Bidelman and Yellamsetty, 2017; Chintanpalli and Heinz, 2013; Chintanpalli et al., 2016; Reinke et al., 2003; Yellamsetty and Bidelman, 2018), listeners were better at perceptually identifying speech mixtures when vowels contained pitch cues. However, we also showed that this perceptual F0-benefit was larger for the clean than the noise degraded (+5 dB SNR) conditions. Additive noise tends to obscure the salient audible cues that are normally exploited by listeners for comprehension of speech (Bidelman, 2016; Shannon et al., 1995; Swaminathan and Heinz, 2012). Our results indeed showed F0-benefit was weaker for double vowel identification in noise compared to clean listening condition (clean > noise). The identification of both the vowels improved from ~ 40% to 70% from 0 to 4 ST (Fig. 2A), consistent with previous studies (Meddis and Hewitt, 1992a). We also found that RTs for identifying both vowels were faster in noise but these speeds were accompanied by lower accuracy. Longer duration RTs and more accurate identification in the clean condition suggests listeners experienced a time-accuracy-tradeoff (i.e., more accurate identification at the expense of slower decision times) during double vowel perception (Bidelman and Yellamsetty, 2017; Yellamsetty and Bidelman, 2018).

3.2. Subcortical encoding of single vs. double vowels

FFRs to single vowels showed more robust encoding than double vowels. For concurrent stimuli that do not have pitch cues (i.e., 0ST conditions with common F0s) the information for identifying the vowels is carried only by the F1s. The improvement in the identification with pitch cues is presumably due to the more distinct timbral representations between vowels with the additional F0 separation. The pattern of nonlinear harmonic interactions in double vowels with the same F0s (0 ST) likely differs from when the vowels are at 4 ST. This is seen in the stimulus autocorrelation function (ACF): at 0ST the ACF showed a large peak at a 150 Hz time lag compared to the flanking autocorrelation peaks, whereas the 4ST ACF showed weaker more distributed peaks at 150 and 190 Hz time lags, respectively. At 0 ST, harmonics of both vowels fall within the same auditory filter channel and thus can add in a constructive manner. However, these within channel interactions also produce simultaneous speech-on-speech masking that results in reduced F0 amplitude for double compared to single vowels (Fig. 4A). At 4 ST, vowel harmonics fall in different auditory filters resulting in energy being spread between channels leading to a further reduction in amplitudes (Fig. 4B). Mechanistically, this additional amplitude reduction could reflect the nonlinear phenomena of suppression (Ruggero et al., 1992; Sachs and Kiang, 1968). Indeed, the ratio of our F0s at 4 ST is 1.26 (190 Hz/150 Hz), a frequency separation known to produce optimal suppression effects (Houtgast, 1974; Shannon, 1976). The spread of synchrony within/across channels most likely reflects nonlinear signal processing that helps in the identification of both vowels. In addition to non-linearity at F0, the acoustic structure

of vowels and formant-based synchrony (Delgutte and Kiang, 1984; Palmer, 1990; Sinex and Geisler, 1983; Young and Sachs, 1979) to harmonics near the formant (Carney et al., 2015; Miller et al., 1997; Tan and Carney, 2005; Young and Sachs, 1979) can further sharpen the temporal representation of spectral shape in neural responses (Young and Sachs, 1979).

Noise tends to obscure amplitude modulations in speech that are essential for its comprehension (Bidelman, 2016; Shannon et al., 1995; Swaminathan and Heinz, 2012). In contrast, in cases of speech-on-speech masking, listeners can better utilize spectral dips for perception, resulting in less effective masking than continuous noise (Peters et al., 1998; Shetty, 2016). FFR changes related to speech-on-speech masking and SNR were evident in both the time and frequency domain results, consistent with previous studies (Bidelman and Krishnan, 2010; Bidelman, 2016; Hornickel et al., 2011; Song et al., 2011; Tierney et al., 2011). Both F0 and higher spectral components (e.g., formant-related harmonics) were systematically degraded with noise, paralleling their deterioration behaviorally (Liu and Kewley-Port, 2004). This reduction in amplitude probably also reflects reduced temporal synchrony and thus worse performance. Studies that have instead showed invariant or larger F0s in noise may reflect stochastic resonance (Prévost et al., 2013; Russo et al., 2004; Smalt et al., 2012) and/or engagement of low-frequency tails of basal, high frequency neurons at high intensity (Kiang and Moxon, 1974).

3.3. Subcortical correlates of double vowel perception

Our study showed only weak links between subcortical neural activity and behavioral percepts in the double vowel paradigm. FFRs failed to predict listeners' identification accuracy. In contrast, FFR F1 amplitudes were associated with faster RT speeds, although this correlation was limited to the noise condition (Fig. 5B). These results replicate previous FFR studies which have shown correlations between F1 coding and behavioral RTs for speech perception (Bidelman et al., 2014a; Bidelman et al., 2014b). Yet, the F0 results contrast a large literature that has shown robust correlations between FFR F0 and degraded speech perception accuracy (Anderson et al., 2012; Anderson et al., 2010; Coffey et al., 2017; Du et al., 2011; Parbery-Clark et al., 2009b). However, one important difference between this and previous work is that all speech-FFR studies to date have used single, isolated speech tokens (e.g., vowels, CVs) rather than the more complex double-vowel mixtures used here. Additionally, our stimuli were designed to have relatively high F0s (150 Hz), compared to other FFR studies where tokens predominantly had voice pitches of ~100 Hz. This is an important distinction as recent studies have shown that FFRs can sometimes have cortical contributions (Coffey et al., 2016) when the F0 of the stimulus is low enough to elicit phase-locking from cortical neurons (<100 Hz). Above the F0s used here (150 Hz), only subcortical (brainstem) sources contribute to the FFR (Bidelman, 2018). It is possible that at least some of the correlations between spectral properties of the FFR (e.g., F0) and various aspects of speech perception reported in earlier studies (Anderson et al., 2012; Anderson et al., 2010; Bidelman and Krishnan, 2010; Coffey et al., 2017; Du et al., 2011; Parbery-Clark et al., 2009b) may be cortical, rather than *subcortical,* in origin. The lack of robust links between the FFR and concurrent speech perception in the present study may be due to the fact that our FFRs reflect more pre-attentive, exogenous neural encoding of the brainstem, which does not always covary with perceptual measures (Bidelman et al., 2013;

Gockel et al., 2011). While our data do not provide strong evidence that perceptual correlates of concurrent vowel processing exist in FFRs, brainstem signal processing is no doubt critical in feeding later decision-based mechanisms at a cortical level. Neural encoding in brainstem might ultimately enhance segregation and perception by higher-order cognitive processes (Bidelman and Alain 2015; Bidelman et al., 2018). Concurrent recordings of FFR (brainstem) and cortical event-related potentials (ERPs) at low (<100 Hz) and high F0s (> 100 Hz) could test this possibility.

Relationships between perceptual and brainstem auditory coding, where they do exist, can be viewed within the framework of corticofugal (top-down) tuning of sensory function. Corticofugal neural pathways, that project back to peripheral structures (Suga et al., 2000; Zhang and Suga, 2005) may control and enhance subcortical encoding of the F0 (voice pitch)-and formant (vowel identity) related information of the stimulus that are necessary for speech-in-noise perception. Of the brain-behavior correlates we did observe, F1 was associated with behavioral RTs, particularly in noise. The higher variability in F1 responses may be due to greater individual differences in the encoding of these higher spectral cues in this more challenging listening condition, producing a larger spread in the data that subsequently allows for correlations. Alternatively, this variability may also be related to corticofugal tuning of sensory FFR encoding that enhances acoustic features of target speech subcortically (Anderson and Kraus, 2013; Reetzke et al., 2018). In background noise, corticofugal mechanisms might search for sensory features that allow the listener to extract and enhance pertinent speech information. This notion is consistent with previous neural data (Cunningham et al., 2001; Parbery-Clark et al., 2009a; Parbery-Clark et al., 2011) and perceptual models showing changes in the weighting of perceptual dimensions because of feedback (Amitay, 2009; Nosofsky, 1987). Online corticofugal activity may adapt rapidly especially in challenging environments (e.g., noise) (Atiani et al., 2009; Elhilali et al., 2009).

Still, why corticofugal effects would be present at F1 but not F0 is unclear. Corticofugal activity may be related to the change in the power of ongoing theta-band rhythms in noise. Indeed, our previous work showed correspondence of theta-band activity with behavioral RTs in noise (Yellamsetty and Bidelman, 2018). Speculatively, lower oscillatory theta-rhythms at a cortical level may act to modulate the encoding of spectral features at a subcortical level, especially in noise. Still, our results are probably not due corticofugal mechanisms as we used a passive listening task whereas cortico-collicular efferent are recruited mainly in tasks requiring goal-directed attention (Slee and David, 2015; Vollmer et al., 2017). Nevertheless, it would be interesting to see how the variable weighting of FFR F0/F1 coding changes with simultaneous changes in oscillatory rhythms (e.g., theta-band) during an active listening task. Attention (theta rhythms) might act to bias and enhance incoming acoustic speech relevant information and suppress noise cf. (cf.Suga, 2012).

A handful of studies have shown certain vowels dominate perception among different vowel pair combinations (Assmann and Summerfield 1990; Assmann and Summerfield, 2004; Chintanpalli et al., 2014; Chintanpalli and Heinz, 2013; Chintanpalli et al., 2016; Meddis and Hewitt, 1992a), reminiscent of our vowel dominancy data (Fig. 6). At 0ST, listeners can take advantage of the relative differences in the levels of spectral peaks between two vowels and one vowel is identified dominantly over the other; whereas identification of both the

vowels is better at 4 ST. Our stimuli did contain a level difference between the F1 spectral peaks of the two vowels; acoustically, /ɛ/ was slightly stronger (2 dB) than /a/ in acoustic power. This level difference was captured in FFR amplitudes (Fig. 3C). In addition, the amplitude of F1 was larger for /ɛ/ than /a/, and for 4ST than 0ST (/ɛ/ 190 > /ɛ/ 150) (Fig. 3C). This effect could be due to the harmonic peaks falling in the F1 region being lower in frequency for the /ɛ/ vowel, indicating more precise phase locking at lower frequencies. Indeed, when FFRs were split by listeners' behavior, double-vowel responses showed closer correspondence to the single /ɛ/ vowel (Fig. 6). Thus, FFRs were largely independent of behavior bias and instead showed a stimulus (rather than perceptual) dominancy.

In sum, we find that FFRs reflect neuro-acoustic representations of peripheral nonlinearities that are carried forward to brainstem processing. Spectro-temporal changes observed in FFRs with pitch cues and noise and the weak behavioral correlations suggest that brainstem responses mainly reflect exogenous stimulus properties of concurrent speech mixtures. Nevertheless, correlations between F1 and behavioral RTs in noisy listening conditions suggest possible corticofugal involvement in enhancing speech relevant representations in the brainstem during more difficult task and/or in challenging listening conditions. Our results show that FFRs reflect pre-attentive mechanisms and concurrent stimulus interactions that can, under certain conditions, predict the successful identification of complex speech mixtures.

4. Methods

4.1. Participants

Sixteen young adults (age $M \pm SD$: 24 ± 2.25 years; 10 females, 6 males) participated in the experiment. All the participants had obtained a similar level of formal education (18.18 ± 2.16 years), were right handed (> 43.2% laterality) (Oldfield, 1971), had normal pure-tone audiometric thresholds (i.e., ≤ 25 dB HL air conduction thresholds) at octave frequencies between 250 and 8000 Hz, and reported no history of neuropsychiatric disorders. Each gave written informed consent in compliance with a protocol (#2370) approved by the University of Memphis Institutional Review Board.

4.2. Stimulus and behavioral task

4.2.1. Double vowel stimuli—Speech sounds were modeled after stimuli from previous studies on concurrent double-vowel segregation (Alain, 2007a; Assmann and Summerfield, 1989; Assmann and Summerfield 1990; Bidelman and Yellamsetty, 2017; Yellamsetty and Bidelman, 2018). Synthetic, steady-state (time-invariant) vowel tokens (/a/, /ɛ/, and /u/) were created using a Klatt synthesizer (Klatt, 1980) implemented in MATLAB® 2014 (The MathWorks, Inc.). Each token was 200 ms in duration including 10-ms $\cos^2$ onset/offset ramping. F0 was either 150 or 190 Hz and formant frequencies (F1, F2) were 766 Hz, 1299 Hz; 542 Hz, 1780 Hz and 329 Hz, 810 Hz for /a/ /ɛ/ and /u/, respectively (Fig. 1). These F0s were selected since they are above the frequencies of observable FFRs in cortex (Bidelman, 2018; Brugge et al., 2009), and thus ensured responses would be of brainstem origin (Bidelman, 2018). Double-vowel stimuli were then created by superimposing single-vowels at 0ST and 4 ST, as shown in Fig. 1. Each vowel pair had either identical (0ST) or different

F0s (4ST). That is, one vowel F0 was set at 150 Hz while the other had an F0 of 150 or 190 Hz so as to produce double-vowels with an F0 separation of either 0 or 4 STs, resulting in two double-vowel pair (1 pair × 2 F0 combinations). Fig. 1 shows the time waveforms, spectra, and the autocorrelation functions (ACFs) of single and double vowel stimuli.

Given time constraints on recording brainstem potentials (i.e., several thousand trials are needed per stimulus condition), the vowels /a/ and /ɛ/ were used to record FFRs. FFRs were recorded in a passive listening paradigm (no behavior task) consistent with previous studies on the relation between FFRs and speech perception (Anderson et al., 2010a; Anderson and Kraus, 2013; Bidelman and Alain 2015; Bidelman, 2016; Bidelman, 2017; Song et al., 2011). For the behavioral identification task (described below), pairs of the vowels /a/, /ɛ/, and /u/ were used, replicating the double-vowel task of our previous reports (Bidelman and Yellamsetty, 2017; Yellamsetty and Bidelman, 2018).

For FFR recordings, both single and double-vowels were presented in clean and noise conditions (separate blocks). The noise was a continuous backdrop of multi-talker noise babble (+ 5 dB SNR) (e.g., Bidelman and Howell, 2016; Nilsson et al., 1994). SNR was manipulated by changing the level of the masker rather than the signal to ensure that SNR was not positively correlated with overall sound level (Bidelman and Howell, 2016; Binder et al., 2004). Babble was presented continuously to avoid it time-locking with stimulus presentation. We chose continuous babble over other forms of acoustic inference (e.g., white noise) because it more closely mimics real-world listening situations and tends to have a larger effect on the auditory evoked potentials (Kozou et al., 2005). Examining FFR responses to both single and double-vowels speech sounds allowed us to assess potential speech-on-speech-masking effects and additivity of speech encoding at the brainstem level.

4.2.2. Behavioral double-vowel identification task.—Participants were presented with double-vowel combination of synthetic steady-state vowel tokens (/a/, /ɛ/, and /u/) as in our previous studies (Bidelman and Yellamsetty, 2017; Yellamsetty and Bidelman, 2018). Double-vowels were presented in two separate blocks of clean and noise (+5 dB SNR) conditions. During each block, listeners heard 50 exemplars of each double vowel combination and were asked to identify both vowels as quickly and accurately as possible by pressing two keys on the keyboard. The inter-stimulus interval was jittered randomly between 800 and 1000 ms to avoid listeners anticipating subsequent trials. The next trial commenced following the listener's behavioral response. Order of vowel pairs was randomized within and across participants and clean and noise conditions were run in separate blocks. Feedback was not provided and listeners were told ahead of time that every trial would contain two unique vowels. For additional details of the stimuli and task, see (Bidelman and Yellamsetty, 2017; Yellamsetty and Bidelman, 2018).

Prior to the experiment proper, we required participants be able to identify single vowels (/a/, /ɛ/, and /u/) in a practice run with > 90% accuracy (e.g., Alain et al., 2007). This ensured task performance would be mediated by *concurrent* sound segregation skills rather than isolated identification, *per se*.

4.3. FFR data recording and preprocessing

For the FFR recordings, participants reclined comfortably in an IAC electro-acoustically shield booth. Participants were instructed to relax and refrain from extraneous body movements while they watched a muted subtitled movie (i.e., passive listening task). EEGs were recorded differentially between Ag/AgCl disk electrodes placed on the scalp at the high forehead (~Fpz) referenced to link mastoids A1/A2) and forehead electrode as ground. Interelectrode impedances were maintained < 2 kΩ. Stimulus presentation was controlled by MATLAB routed to a TDT RP2 interface (Tucker-Davis Technologies). Speech stimuli were delivered binaurally using fixed (rarefaction) polarity at an intensity of 81 dB SPL through shielded ER-2 insert earphones (Etymotic Research).[1] Control runs confirmed the absence of artifacts in response waveforms. The order of single and double vowel stimuli was randomized within and across participants; clean and noise conditions were run in separate blocks. The inter-stimulus interval was 50 ms. In total, there were 2000 trials for each of the individual stimulus conditions.

Neural activity was digitized using a sampling rate of 10 kHz and online filter passband of 0–3500 Hz (SynAmps RT amplifiers; Compumedics Neuroscan). EEGs were then epoched (0–250 ms) and averaged in the time domain to derive FFRs for each condition. Sweeps exceeding ± 50 μV were rejected as artifacts prior to averaging. FFRs were then bandpass filtered (100–3000 Hz) for response visualization and quantification. The entire experimental protocol including behavioral and electrophysiological testing lasted ~ 2.5 h.

4.4. FFR analysis

Fast Fourier transforms (FFTs) were computed from the response time-waveforms (0–250 ms) using Brainstorm (V.3.4) (Tadel et al., 2011). Brainstorm expresses FFT amplitudes as power with a scaling factor of units^2/Hz * 10^{-13}; subsequent measures reflect this scaling. From each FFR spectrum, we measured the F0, harmonics, and F1-formant frequency amplitudes to quantify “pitch” and “timbre” coding for each condition. We estimated the magnitude of the response at F0 and harmonics of the single and double vowels by manually picking the maximum spectral energy within 10 Hz wide bins surrounding the F0 and five harmonics. F1 magnitude was taken as the average spectral energy (on a linear scale) in the frequency ranges between 392 and 692 Hz for /ɛ/$_{150Hz}$ (0ST), 352 and 732 Hz for /ɛ/$_{190Hz}$ (4ST) and 616 and 916 Hz for /a/$_{150Hz}$ vowels. These ranges were determined based on the expected F0/F1 frequencies from the input stimulus. Stimulus-related changes in F0 and F1-formant magnitudes provided an index of how concurrent stimuli and noise interference degrade the brainstem representation of pitch and timbre cues in speech.

4.5. Behavioral data analysis

4.5.1. Identification accuracy and the “F0 benefit”—

Behavioral speech identification accuracy was analyzed as the percent of trials where *both* vowel sounds were correctly identified. Percent correct scores were arcsine transformed to improve homogeneity of variance assumptions necessary for parametric statistics (Studebaker, 1985). Increasing the F0 between two vowels provides a pitch cue which leads to an improvement in accuracy identifying concurrent vowels (Assmann and Summerfield, 1990; Chintanpalli and Heinz, 2013; Meddis and Hewitt, 1992)-an effect referred to as the “F0-benefit”

(Arehart et al., 1997; Bidelman and Yellamsetty, 2017; Chintanpalli and Heinz, 2013; Yellamsetty and Bidelman, 2018). We calculated the F0-benefit for each listener, computed as the difference in performance (%-correct) between the 4ST and 0ST conditions. F0-benefit was computed separately for clean and noise stimuli to compare the magnitude of benefit with and without noise interference.

4.5.2. Reaction time (RTs)—For a given double-vowel condition, behavioral speech labeling speeds [i.e., reaction times (RTs)] were computed separately for each participant as the median response latency across trials. RTs were taken as the time lapse between the onset of the stimulus presentation and listeners' identification of both vowel sounds. RTs shorter than 250 ms or exceeding 6000 ms were discarded as implausibly fast responses and lapses of attention, respectively (e.g., Bidelman and Yellamsetty, 2017; Yellamsetty and Bidelman, 2018).

4.6. Statistical analysis

Unless otherwise noted, two-way, mixed-model ANOVAs were conducted on all dependent variables (GLIMMIX Procedure, SAS® 9.4, SAS Institute, Inc.). Stimulus SNR (2 levels; clean, +5 dB noise) and semitones (2 levels; 0ST, 4ST) functioned as fixed effects; subjects served as a random factor. Tukey-Kramer multiple comparisons-controlled Type I error inflation. An *a priori* significance level was set at $\alpha = 0.05$. To examine the degree to which neural responses predicted behavioral speech perception, we performed weighted least squares regression between listeners' FFR amplitudes and their perceptual identification accuracy (percept correct scores) in the double-vowel task. Robust bisquare fitting was achieved using "fitlm" in MATLAB.

Acknowledgements

This research was supported by the Institute for Intelligent Systems Student Dissertation Grant Program at the University of Memphis (A.Y.) and by the National Institute On Deafness And Other Communication Disorders of the National Institutes of Health under award number R01DC016267 (G.M.B.).

References

Aiken SJ, Picton TW, 2008 Envelope and spectral frequency-following responses to vowel sounds. Hear. Res. 245, 35–47. [PubMed: 18765275]

Al Osman R, Dajani HR, Giguère C, 2017 Self-masking and overlap-masking from reverberation using the speech-evoked auditory brainstem response. J. Acoust. Soc. Am. 142 EL555–EL560.

Alain C, Reinke K, He Y, Wang C, Lobaugh N, 2005a Hearing two things at once: neurophysiological indices of speech segregation and identification. J. Cognit. Neurosci. 17, 811–818. [PubMed: 15904547]

Alain C, Reinke K, McDonald KL, Chau W, Tam F, Pacurar A, Graham S, 2005b Left thalamo-cortical network implicated in successful speech separation and identification. Neuroimage 26, 592–599. [PubMed: 15907316]

Alain C, 2007 Breaking the wave: effects of attention and learning on concurrent sound perception. Hear. Res. 229, 225–236. [PubMed: 17303355]

Alain C, Snyder JS, He Y, Reinke KS, 2007 Changes in auditory cortex parallel rapid perceptual learning. Cereb. Cortex 17, 1074–1084. [PubMed: 16754653]

Alain C, Arsenault JS, Garami L, Bidelman GM, Snyder JS, 2017 Neural correlates of speech segregation based on formant frequencies of adjacent vowels. Sci. Rep. 7, 1–11. [PubMed: 28127051]

Amitay S, 2009 Forward and reverse hierarchies in auditory perceptual learning. Learn. Percept. 1, 59–68.

Anderson S, Kraus N, 2010 Sensory-cognitive interaction in the neural encoding of speech in noise: a review. J. Am. Acad. Audiol. 21, 575–585. [PubMed: 21241645]

Anderson S, Skoe E, Chandrasekaran B, Zecker S, Kraus N, 2010 Brainstem correlates of speech-in-noise perception in children. Hear. Res. 270, 151–157. [PubMed: 20708671]

Anderson S, Parbery-Clark A, White-Schwoch, Kraus, 2012 Aging affects neural precision of speech encoding. J. Neurosci. 32, 14156–14164. [PubMed: 23055485]

Anderson S, Kraus N, 2013 cABR: a neural probe of speech-in-noise processing. In: Proceedings of the International Symposium on Auditory and Audiological Research, Vol. 3, pp. 231–241.

Arehart KH, King CA, McLean-Mudgett KS, 1997 Role of fundamental frequency differences in the perceptual separation of competing vowel sounds by listeners with normal hearing and listeners with hearing loss. J. Speech Language Hearing Res. 40, 1434–1444.

Assmann PF, Summerfield Q, 1990 Modeling the perception of concurrent vowels: vowels with different fundamental frequencies. J. Acoust. Soc. Am. 88, 680–697. [PubMed: 2212292]

Assmann PE, Summerfield Q, 2004 The perception of speech under adverse conditions In: Speech Processing in the Auditory System. Springer, pp. 231–308.

Assmann PF, Summerfield Q, 1989 Modeling the perception of concurrent vowels: vowels with the same fundamental frequency. J. Acoust. Soc. Am. 85, 327–338. [PubMed: 2921415]

Assmann PF, Summerfield Q, 1994 The contribution of waveform interactions to the perception of concurrent vowels. J. Acoust. Soc. Am. 95, 471–484. [PubMed: 8120258]

Atiani S, Elhilali M, David SV, Fritz JB, Shamma SA, 2009 Task difficulty and performance induce diverse adaptive patterns in gain and shape of primary auditory cortical receptive fields. Neuron 61, 467–480. [PubMed: 19217382]

Bidelman GM, Krishnan A, 2010 Effects of reverberation on brainstem representation of speech in musicians and non-musicians. Brain Res. 1355, 112–125. [PubMed: 20691672]

Bidelman GM, Gandour JT, Krishnan A, 2011 Musicians and tone-language speakers share enhanced brainstem encoding but not perceptual benefits for musical pitch. Brain Cogn. 77, 1–10. [PubMed: 21835531]

Bidelman GM, Moreno S, Alain C, 2013 Tracing the emergence of categorical speech perception in the human auditory system. Neuroimage 79, 201–212. [PubMed: 23648960]

Bidelman GM, Villafuerte JW, Moreno S, Alain C, 2014a Age-related changes in the subcortical-cortical encoding and categorical perception of speech. Neurobiol. Aging 35, 2526–2540. [PubMed: 24908166]

Bidelman GM, Weiss MW, Moreno S, Alain C, 2014b Coordinated plasticity in brainstem and auditory cortex contributes to enhanced categorical speech perception in musicians. Eur. J. Neurosci. 40, 2662–2673. [PubMed: 24890664]

Bidelman GM, 2015a Induced neural beta oscillations predict categorical speech perception abilities. Brain Lang. 141, 62–69. [PubMed: 25540857]

Bidelman GM, Alain C, 2015 Hierarchical neurocomputations underlying concurrent sound segregation: connecting periphery to percept. Neuropsychologia 68, 38–50. [PubMed: 25542675]

Bidelman GM, 2016 Relative contribution of envelope and fine structure to the sub-cortical encoding of noise-degraded speech. J. Acoust. Soc. Am. 140 EL358–EL363.

Bidelman GM, Howell M, 2016 Functional changes in inter-and intra-hemispheric auditory cortical processing underlying degraded speech perception. Neuroimage 124, 581–590. [PubMed: 26386346]

Bidelman GM, 2017 Communicating in challenging environments: Noise and reverberation In: The Frequency-Following Response. Springer, pp. 193–224.

Bidelman GM, Davis MK, Pridgen MH, 2018 Brainstem-cortical functional connectivity for speech is differentially challenged by noise and reverberation. Hear. Res.

Bidelman GM, Powers L, 2018 Response properties of the human frequency-following response (FFR) to speech and non-speech sounds: level dependence, adaptation and phase-locking limits. Int. J. Audiol. 1–8.

Bidelman GM, 2015b Multichannel recordings of the human brainstem frequency-following response: scalp topography, source generators, and distinctions from the transient ABR. Hear. Res. 323, 68–80. [PubMed: 25660195]

Bidelman GM, Yellamsetty A, 2017 Noise and pitch interact during the cortical segregation of concurrent speech. Hear. Res. 351, 34–44. [PubMed: 28578876]

Bidelman GM, 2018 Subcortical sources dominate the neuroelectric auditory frequency-following response to speech. Neuroimage 175, 56–69. [PubMed: 29604459]

Binder JR, Liebenthal E, Possing ET, Medler DA, Ward BD, 2004 Neural correlates of sensory and decision processes in auditory object identification. Nat. Neurosci. 7, 295–301. [PubMed: 14966525]

Bregman, 1990 Auditory Scene Analysis: The Perceptual Organization of Sound. MIT Press, Cambridge, MA.

Brugge JF, Nourski KV, Oya H, Reale RA, Kawasaki H, Steinschneider M, Howard MA 3rd, 2009 Coding of repetitive transients by auditory cortex on Heschl's gyrus. J. Neurophysiol. 102, 2358–2374. [PubMed: 19675285]

Cariani PA, Delgutte B, 1996 Neural correlates of the pitch of complex tones. I. Pitch and pitch salience. J. Neurophysiol. 76, 1698–1716. [PubMed: 8890286]

Carlyon RP, 2004 How the brain separates sounds. Trends Cogn. Sci. 8, 465–471. [PubMed: 15450511]

Carney LH, Li T, McDonough JM, 2015 Speech coding in the brain: representation of vowel formants by midbrain neurons tuned to sound fluctuations. Eneuro 2 ENEURO. 0004–15.2015.

Cedolin L, Delgutte B, 2005 Pitch of complex tones: rate-place and interspike interval representations in the auditory nerve. J. Neurophysiol. 94, 347–362. [PubMed: 15788522]

Chandrasekaran B, Hornickel J, Skoe E, Nicol T, Kraus N, 2009 Context-dependent encoding in the human auditory brainstem relates to hearing speech in noise: implications for developmental dyslexia. Neuron 64, 311–319. [PubMed: 19914180]

Chandrasekaran B, Kraus N, 2010 The scalp-recorded brainstem response to speech: neural origins and plasticity. Psychophysiology 47, 236–246. [PubMed: 19824950]

Chimento TC, Schreiner CE, 1990 Selectively eliminating cochelar microphonic contamination from the frequency-following response. Electroencephalogr. Clin. Neurophysiol. 75, 88–96. [PubMed: 1688778]

Chintanpalli A, Ahlstrom JB, Dubno JR, 2014 Computational model predictions of cues for concurrent vowel identification. J. Assoc. Res. Otolaryngol.: JARO 15, 823–837. [PubMed: 25002128]

Chintanpalli A, Heinz MG, 2013 The use of confusion patterns to evaluate the neural basis for concurrent vowel identification. J. Acoust. Soc. Am. 134, 2988–3000. [PubMed: 24116434]

Chintanpalli A, Ahlstrom JB, Dubno JR, 2016 Effects of age and hearing loss on concurrent vowel identification. J. Acoust. Soc. Am. 140, 4142. [PubMed: 28040038]

Coffey ER, Herholz SC, Chepesiuk AM, Baillet S, Zatorre RJ, 2016 Cortical contributions to the auditory frequency-following response revealed by MEG. Nat. Commun. 7, 11070. [PubMed: 27009409]

Coffey ERJ, Chepesiuk AMP, Herholz SC, Baillet S, Zatorre RJ, 2017 Neural correlates of early sound encoding and their relationship to speech-in-noise perception. Front. Neurosci. 11, 479. [PubMed: 28890684]

Culling J, 1990 Exploring the conditions for the perceptual separation of concurrent voices using F0 differences. Proc. Inst. Acoust. 12, 559–566.

Cunningham J, Nicol T, Zecker SG, Bradlow A, Kraus N, 2001 Neurobiologic responses to speech in noise in children with learning problems: deficits and strategies for improvement. Clin. Neurophysiol. 112, 758–767. [PubMed: 11336890]

de Cheveigne A, Kawahara H, Tsuzaki M, Aikawa K, 1997 Concurrent vowel identification. I. Effects of relative amplitude and F0 difference. J. Acoust. Soc. Am. 101, 2839–2847.

Delgutte B, Kiang N, 1984 Speech coding in the auditory nerve: I. Vowel-like sounds. J. Acoust. Soc. Am. 75, 866–878. [PubMed: 6707316]

Du Y, Kong L, Wang Q, Wu X, Li L, 2011 Auditory frequency-following response: a neurophysiological measure for studying the “cocktail-party problem”. Neurosci. Biobehav. Rev. 35, 2046–2057. [PubMed: 21645541]

Dyson BJ, Alain C, 2004 Representation of concurrent acoustic objects in primary auditory cortex. J. Acoust. Soc. Am. 115, 280–288. [PubMed: 14759021]

Elhilali M, Ma L, Micheyl C, Oxenham AJ, Shamma SA, 2009 Temporal coherence in the perceptual organization and cortical representation of auditory scenes. Neuron 61, 317–329. [PubMed: 19186172]

Gardi J, Merzenich M, McKean C, 1979 Origins of the scalp-recorded frequency-following response in the cat. Audiology 18, 353–380.

Glaser EM, Suter CM, Dasheiff R, Goldberg A, 1976 The human frequency-following response: its behavior during continuous tone and tone burst stimulation. Electroencephalogr. Clin. Neurophysiol. 40, 25–32. [PubMed: 55345]

Gockel H, Carlyon P, Mehta A, Plack C, 2011 The frequency following response (FFR) may reflect pitch-bearing information but is not a direct representation of pitch. J. Assoc. Res. Otolaryngol. 12, 767–782. [PubMed: 21826534]

Henry KS, Abrams KS, Forst J, Mender MJ, Neilans EG, Idrobo F, Carney LH, 2017 Midbrain synchrony to envelope structure supports behavioral sensitivity to single-formant vowel-like sounds in noise. J. Assoc. Res. Otolaryngol. 18, 165–181. [PubMed: 27766433]

Hornickel J, Skoe E, Nicol T, Zecker S, Kraus N, 2009 Subcortical differentiation of stop consonants relates to reading and speech-in-noise perception. Proc. Natl. Acad. Sci. 106, 13022–13027. [PubMed: 19617560]

Hornickel J, Chandrasekaran B, Zecker S, Kraus N, 2011 Auditory brainstem measures predict reading and speech-in-noise perception in school-aged children. Behav. Brain Res. 216, 597–605. [PubMed: 20826187]

Houtgast T, 1974 Lateral Suppression in Hearing. Acad. Pers BV, Amsterdam.

Jane JY, Young ED, 2000 Linear and nonlinear pathways of spectral information transmission in the cochlear nucleus. Proc. Natl. Acad. Sci. 97, 11780–11786. [PubMed: 11050209]

Keilson SE, Richards VM, Wyman BT, Young ED, 1997 The representation of concurrent vowels in the cat anesthetized ventral cochlear nucleus: evidence for a periodicity-tagged spectral representation. J. Acoust. Soc. Am. 102, 1056–1071. [PubMed: 9265754]

Kiang N, Moxon E, 1974 Tails of tuning curves of auditory-nerve fibers. J. Acoust. Soc. Am. 55, 620–630. [PubMed: 4819862]

Klatt DH, 1980 Software for a cascade/parallel formant synthesizer. J. Acoust. Soc. Am. 67, 971–995.

Koffka K, 1935 Principles of Gestalt Psychology International Library of Psychology, Philosophy and Scientific Method. Harcourt Brace, New York.

Kozou H, Kujala T, Shtyrov Y, Toppila E, Starck J, Alku P, Naatanen R, 2005 The effect of different noise types on the speech and non-speech elicited mismatch negativity. Hear. Res. 199, 31–39. [PubMed: 15574298]

Krishnan A, 1999 Human frequency-following responses to two-tone approximations of steady-state vowels. Audiol. Neurotol. 4, 95–103.

Krishnan A, Agrawal S, 2010 Human frequency-following response to speech-like sounds: correlates of off-frequency masking. Audiol. Neurotol. 15, 221–228.

Krishnan A, Bidelman GM, Gandour JH, 2010 Neural representation of pitch salience in the human brainstem revealed by psychophysical and electrophysiological indices. Hear. Res. 268, 60–66. [PubMed: 20457239]

Krishnan A, 2002 Human frequency-following responses: representation of steady-state synthetic vowels. Hear. Res. 166, 192–201. [PubMed: 12062771]

Li X, Jeng FC, 2011 Noise tolerance in human frequency-following responses to voice pitch. J. Acoust. Soc. Am. 129, EL21–EL26.

Liu C, Kewley-Port D, 2004 Formant discrimination in noise for isolated vowels. J. Acoust. Soc. Am. 116, 3119–3129. [PubMed: 15603157]

Liu L-F, Palmer AR, Wallace MN, 2006 Phase-locked responses to pure tones in the inferior colliculus. J. Neurophysiol. 95, 1926–1935. [PubMed: 16339005]

Marsh JT, Brown WS, Smith JC, 1974 Differential brainstem pathways for the conduction of auditory frequency-following responses. Electroencephalogr. Clin. Neurophysiol. 36, 415–424. [PubMed: 4140069]

McKeown JD, 1992 Perception of concurrent vowels: the effect of varying their relative level. Speech Commun. 11, 1–13.

Meddis R, Hewitt MJ, 1992 Modeling the identification of concurrent vowels with different fundamental frequencies. J. Acoust. Soc. Am. 91, 233–245. [PubMed: 1737874]

Miller RL, Schilling JR, Franck KR, Young ED, 1997 Effects of acoustic trauma on the representation of the vowel/ɛ/in cat auditory nerve fibers. J. Acoust. Soc. Am. 101, 3602–3616. [PubMed: 9193048]

Nilsson M, Soli SD, Sullivan JA, 1994 Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. J. Acoust. Soc. Am. 95, 1085–1099. [PubMed: 8132902]

Nosofsky RM, 1987 Attention and learning processes in the identification and categorization of integral stimuli. J. Exp. Psychol. Learn. Mem. Cogn. 13, 87. [PubMed: 2949055]

Oldfield RC, 1971 The assessment and analysis of handedness: The Edinburgh inventory. Neuropsychologia 9, 97–113. [PubMed: 5146491]

Palmer AR, 1990 The representation of the spectra and fundamental frequencies of steady-state single- and double-vowel sounds in the temporal discharge patterns of guinea pig cochlear-nerve fibers. J. Acoust. Soc. Am. 88, 1412–1426. [PubMed: 2229676]

Palmer AR, Winter IM, 1992 Cochlear nerve and cochlear nucleus responses to the fundamental frequency of voiced speech sounds and harmonic complex tones. Auditory Physiol. Percept. 83, 231–239.

Parbery-Clark A, Skoe E, Lam C, Kraus N, 2009a Musician enhancement for speech-in-noise. Ear Hear. 30, 653–661. [PubMed: 19734788]

Parbery-Clark A, Skoe E, Kraus N, 2009b Musical experience limits the degradative effects of background noise on the neural processing of sound. J. Neurosci. 29, 14100–14107. [PubMed: 19906958]

Parbery-Clark A, Marmel F, Bair J, Kraus N, 2011 What subcortical-cortical relationships tell us about processing speech in noise. Eur. J. Neurosci. 33, 549–557. [PubMed: 21255123]

Peters RW, Moore BCJ, Baer T, 1998 Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people. J. Acoust. Soc. Am. 103, 577–587. [PubMed: 9440343]

Prévost F, Laroche M, Marcoux A, Dajani H, 2013 Objective measurement of physiological signal-to-noise gain in the brainstem response to a synthetic vowel. Clin. Neurophysiol. 124, 52–60. [PubMed: 22688081]

Reale RA, Geisler CD, 1980 Auditory-nerve fiber encoding of two-tone approximations to steady-state vowels. J. Acoust. Soc. Am. 67, 891–902. [PubMed: 7358914]

Reetzke R, Xie Z, Llanos F, Chandrasekaran B, 2018 Tracing the trajectory of sensory plasticity across different stages of speech learning in adulthood. Curr. Biol.

Reinke K, He Y, Wang C, Alain C, 2003 Perceptual learning modulates sensory evoked response during vowel segregation. Cogn. Brain Res. 17, 781–791.

Ruggero MA, Robles L, Rich NC, 1992 Two-tone suppression in the basilar membrane of the cochlea: mechanical basis of auditory-nerve rate suppression. J. Neurophysiol. 68, 1087–1099. [PubMed: 1432070]

Russo N, Nicol T, Musacchia G, Kraus N, 2004 Brainstem responses to speech syllables. Clin. Neurophysiol. 115, 2021–2030. [PubMed: 15294204]

Sachs MB, Kiang NY, 1968 Two-tone inhibition in auditory-nerve fibers. J. Acoust. Soc. Am. 43, 1120–1128. [PubMed: 5648103]

Scheffers MTM, 1983 Sifting Vowels: Auditory Pitch Analysis and Sound Segregation. Rijksuniversiteit te Groningen.

Shannon RV, 1976 Two-tone unmasking and suppression in a forward-masking situation. J. Acoustical Soc. Am. 59, 1460–1470.

Shannon RV, Zeng FG, Kamath V, Wygonski J, Ekelid M, 1995 Speech recognition with primarily temporal cues. Science 270, 303–304. [PubMed: 7569981]

Shetty HN, 2016 Temporal cues and the effect of their enhancement on speech perception in older adults-a scoping review. J. Otol. 11, 95–101. [PubMed: 29937817]

Sinex GD, Geisler CD, 1983 Responses of auditory-nerve fibers to consonant-vowel syllables. J. Acoust. Soc. Am. 73, 602–615. [PubMed: 6841800]

Sinex DG, Henderson J, Li H, Chen G-D, 2002a Responses of chinchilla inferior colliculus neurons to amplitude-modulated tones with different envelopes. JARO-J. Assoc. Res. Otolaryngol. 3, 390–402.

Sinex DG, Sabes JH, Li H, 2002b Responses of inferior colliculus neurons to harmonic and mistuned complex tones. Hear. Res. 168, 150–162. [PubMed: 12117517]

Sinex DG, Li H, Velenovsky DS, 2005 Prevalence of stereotypical responses to mistuned complex tones in the inferior colliculus. J. Neurophysiol. 94, 3523–3537. [PubMed: 16079190]

Sinex DG, 2008 Responses of cochlear nucleus neurons to harmonic and mistuned complex tones. Hear. Res. 238, 39–48. [PubMed: 18078726]

Slee SJ, David SV, 2015 Rapid task-related plasticity of spectrotemporal receptive fields in the auditory midbrain. J. Neurosci. 35, 13090–13102. [PubMed: 26400939]

Smalt CJ, Krishnan A, Bidelman GM, Ananthakrishnan S, Gandour JT, 2012 Distortion products and their influence on representation of pitch-relevant information in the human brainstem for unresolved harmonic complex tones. Hear. Res. 292, 26–34. [PubMed: 22910032]

Smith, Marsh, Brown, 1975 Far-field recorded frequency-following responses: evidence for the locus of brainstem sources. Electroencephalogr. Clin. Neurophysiol. 39, 465–472. [PubMed: 52439]

Song JH, Skoe E, Banai K, Kraus N, 2011 Perception of speech in noise: neural correlates. J. Cogn. Neurosci. 23, 2268–2279. [PubMed: 20681749]

Stillman RD, Crow G, Moushegian G, 1978 Components of the frequency-following potential in man. Electroencephalogr. Clin. Neurophysiol. 44, 438–446. [PubMed: 76552]

Studebaker GA, 1985 A “rationalized” arcsine transform. J. Speech Lang. Hear. Res. 28, 455–462.

Suga N, Gao E, Zhang Y, Ma X, Olsen JF, 2000 The corticofugal system for hearing: recent progress. Proc. Natl. Acad. Sci. 97, 11807–11814. [PubMed: 11050213]

Suga N, 2012 Tuning shifts of the auditory system by corticocortical and corticofugal projections and conditioning. Neurosci. Biobehav. Rev. 36, 969–988. [PubMed: 22155273]

Swaminathan J, Heinz MG, 2012 Psychophysiological analyses demonstrate the importance of neural envelope coding for speech perception in noise. J. Neurosci. 32, 1747–1756. [PubMed: 22302814]

Tadel F, Baillet S, Mosher JC, Pantazis D, Leahy RM, 2011 Brainstorm: a user-friendly application for MEG/EEG analysis. Comput. Intell. Neurosci. 2011, 8.

Tan Q, Carney LH, 2005 Encoding of vowel-like sounds in the auditory nerve: model predictions of discrimination performance. J. Acoustical Soc. Am. 117, 1210–1222.

Tierney A, Parbery-Clark A, Skoe E, Kraus N, 2011 Frequency-dependent effects of background noise on subcortical response timing. Hear. Res. 282, 145–150. [PubMed: 21907782]

Vollmer M, Beitel RE, Schreiner CE, Leake PA, 2017 Passive stimulation and behavioral training differentially transform temporal processing in the inferior colliculus and primary auditory cortex. J. Neurophysiol. 117, 47–64. [PubMed: 27733594]

Worden FG, Marsh JT, 1968 Frequency-following (microphonic-like) neural responses evoked by sound. Electroencephalogr. Clin. Neurophysiol. 25, 42–52. [PubMed: 4174782]

Xie Z, Reetzke R, Chandrasekaran B, 2017 Stability and plasticity in neural encoding of linguistically relevant pitch patterns. J. Neurophysiol. 117, 1409–1424.

Yellamsetty A, Bidelman GM, 2018 Low-and high-frequency cortical brain oscillations reflect dissociable mechanisms of concurrent speech segregation in noise. Hear. Res. 361, 92–102. [PubMed: 29398142]

Young ED, Sachs, 1979 Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. J. Acoust. Soc. Am. 66, 1381–1403. [PubMed: 500976]

Zhang Suga, 2005 Corticofugal feedback for collicular plasticity evoked by electric stimulation of the inferior colliculus. J. Neurophysiol. 94, 2676–2682. [PubMed: 16000518]

Zwicker, 1984 Auditory recognition of diotic and dichotic vowel pairs. Speech Commun. 3, 265–277.

Further reading

Aiken SJ, Picton TW, 2006 Envelope following responses to natural vowels. Audiol. Neurotol. 11, 213–232.

HIGHLIGHTS

- Examined subcortical encoding of concurrent speech identification via FFR.
- Varied pitch (F0) and noise (SNR) in double-vowel mixtures.
- FFRs for double vowels altered in a systematic manner from their single vowel counterparts.
- Pre-attentive subcortical encoding could predict perceptual speed but not accuracy.

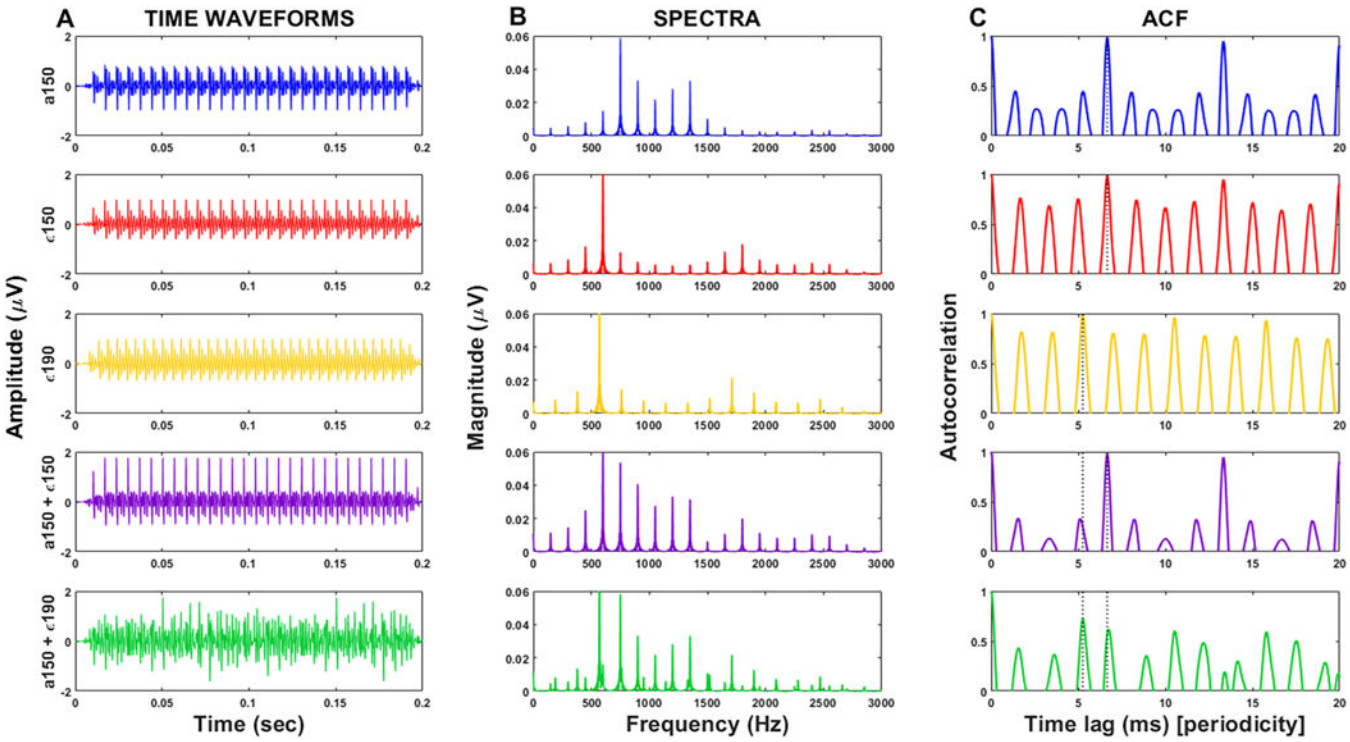

Fig. 1.

Time-waveforms, spectra, and autocorrelation functions (ACFs) of single and double vowel speech stimuli. 150 and 190 represent the two F0s (in Hz) used to form concurrent vowel pairs with a 0 and 4 ST pitch difference. Autocorrelation functions show the strength of stimulus periodicity across time-lags (i.e., 1/frequency). Dotted lines mark the F0 periodicities of the two vowels (i.e., 150 and 190 Hz).

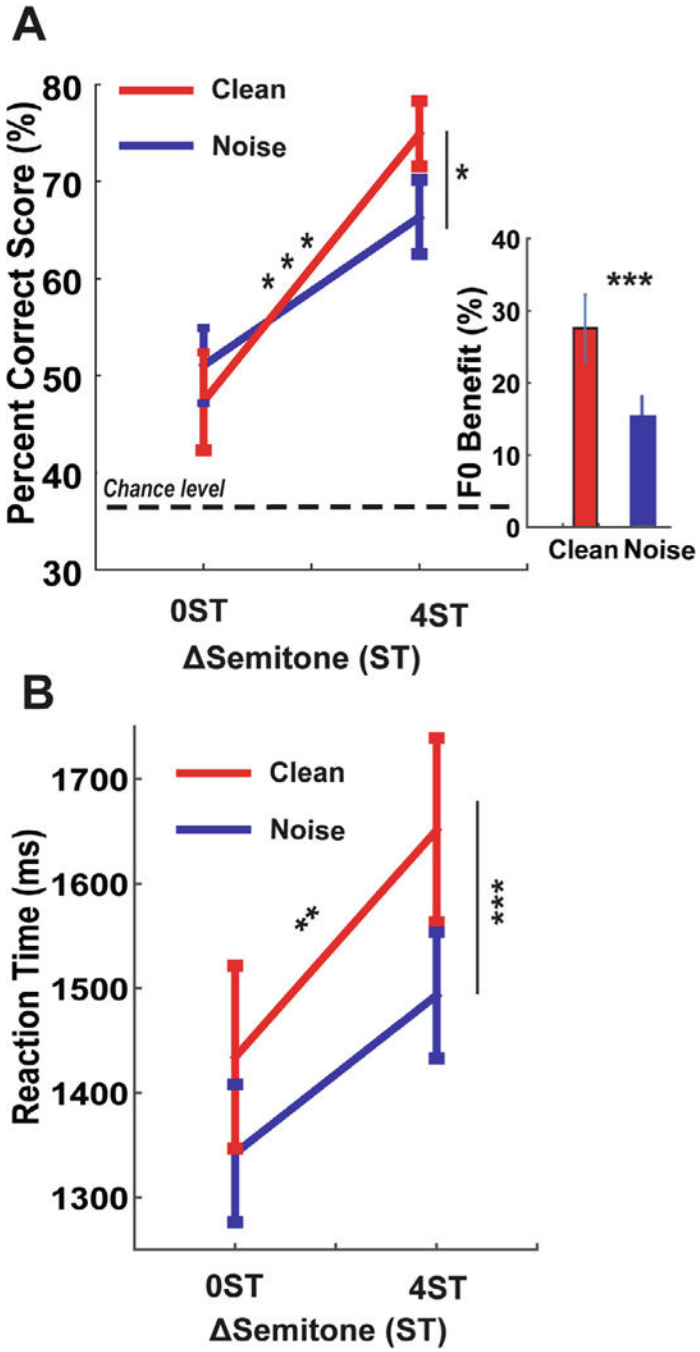

Fig. 2.

Behavioral responses for double-vowel stimuli. (A) Accuracy for identifying both tokens of a two-vowel mixture. Performance is poorer when concurrent speech sounds contain the same F0 (0ST) and improve ~30% when vowels contain differing F0s (4ST). (*Inset*) Behavioral FO-benefit, defined as the improvement in %-accuracy from 0ST to 4ST, indexes the benefit of pitch cues to speech identification. F0-benefit is stronger for clean vs. noisy (+5 dB SNR) speech indicating that listeners are poorer at exploiting pitch cues when segregating acoustically-degraded signals. (B) Speed (i.e., RTs) for double-vowel

identification. Listeners are marginally faster at identifying speech in noise. However, faster RTs at the expense of poorer accuracy (panel A) suggests a time-accuracy tradeoff in double-vowel identification. Error bars = ± 1 s.e.m. *$p < 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$.

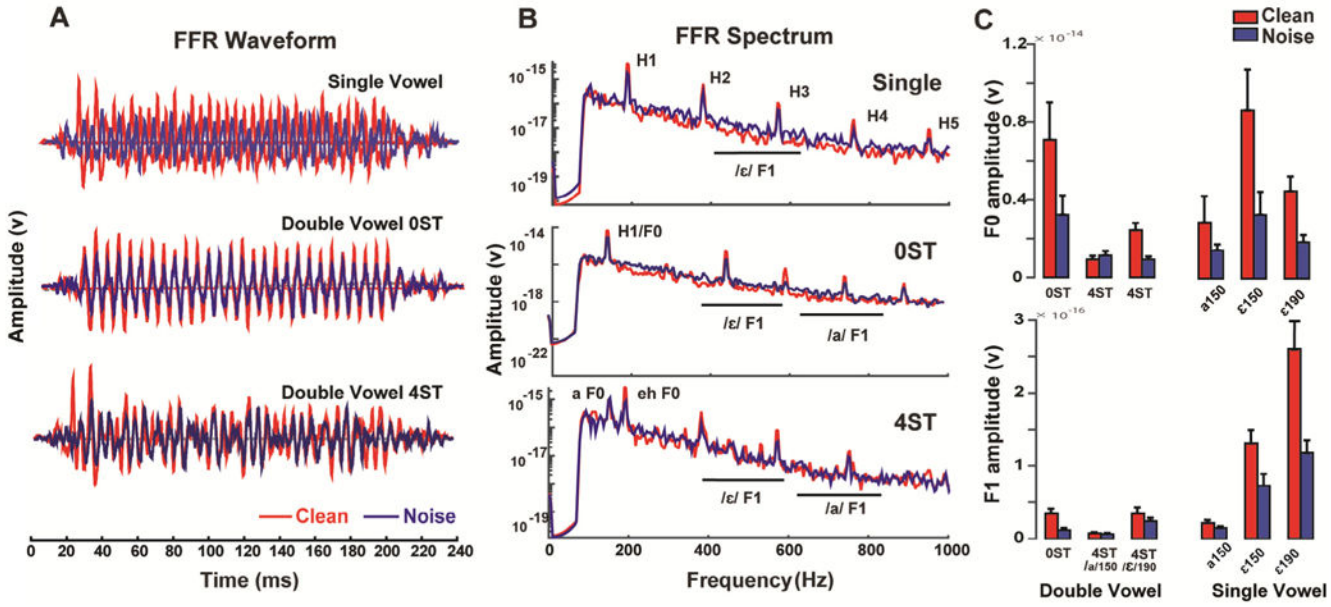

Fig. 3.

Brainstem FFR to double vowel mixtures. (A) FFR waveforms (B) spectra. Neural responses reveal energy at the voice fundamental (F0) and integer-related harmonics (H1-H5). F1, first formant range. (C) Brainstem encoding of the pitch (F0) and timbre (F1) as a function of the vowel count (i.e., single vs. double) and SNR. FFRs are more robust for (i) single than double vowels (single > double) and (ii) at 0ST vs. 4ST (0ST > 4ST). Responses also deteriorate with noise. Error bars = ± 1 s.e.m.

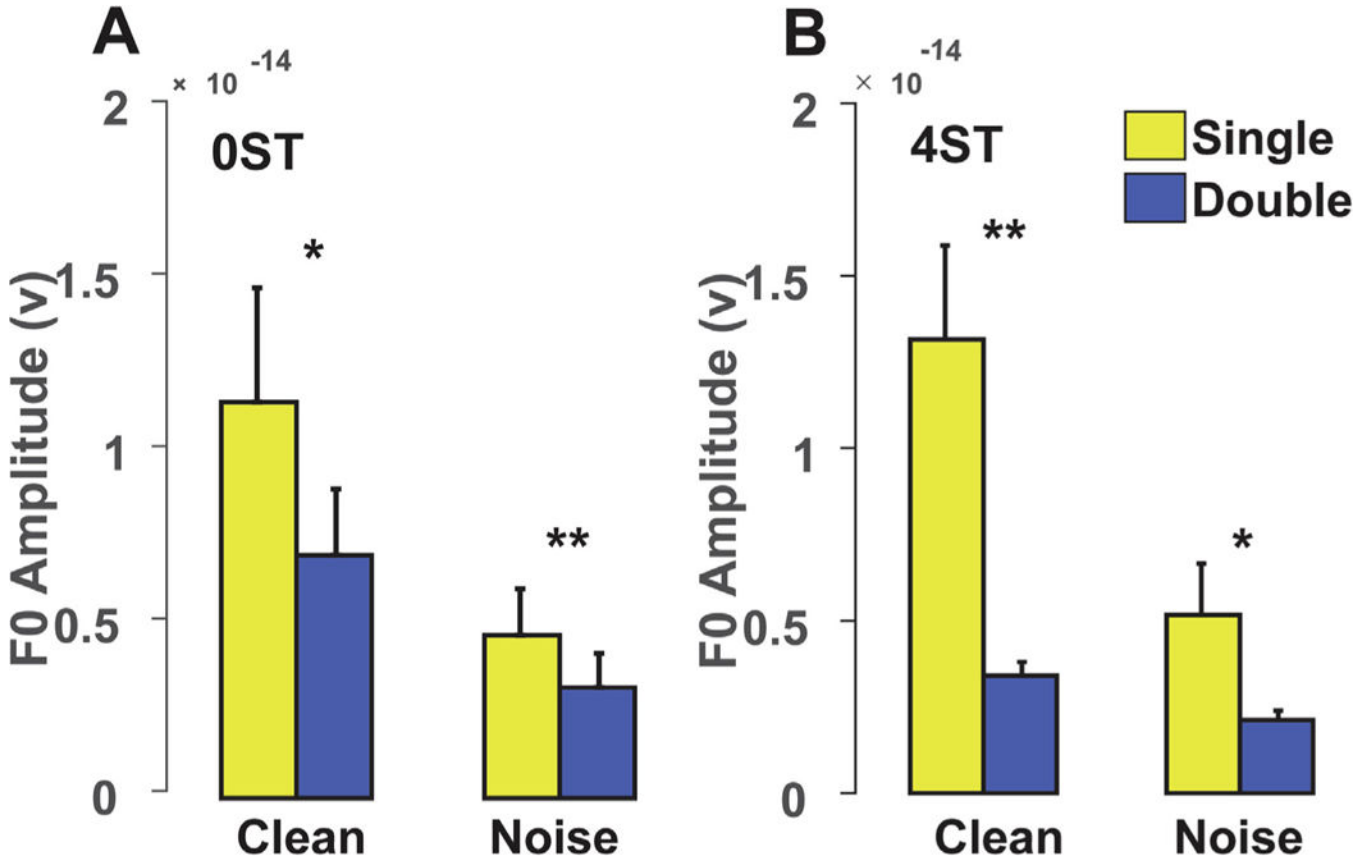

Fig. 4.

Additive noise vs. speech-on-speech masking effects at 0ST (A) and 4ST (B) measured at F0. (A) 0ST and (B) 4 ST mixtures. At 0 ST, (within channel) responses reflect constructive interference (additive effect) due to the same F0s and speech-on-speech masking between vowels. At 4ST (across channel), additional suppression is observed along with the speech-on-speech masking resulting in further reduction in amplitude in both clean and noise conditions. The masking of babble noise on speech (cf.clean vs. noise) was greater than the speech-on-speech masking (cf.double vs. single vowel) at both 0ST and 4ST. Error bars = ± 1 s.e.m. *$p < 0.05$, **$p \leq 0.01$.

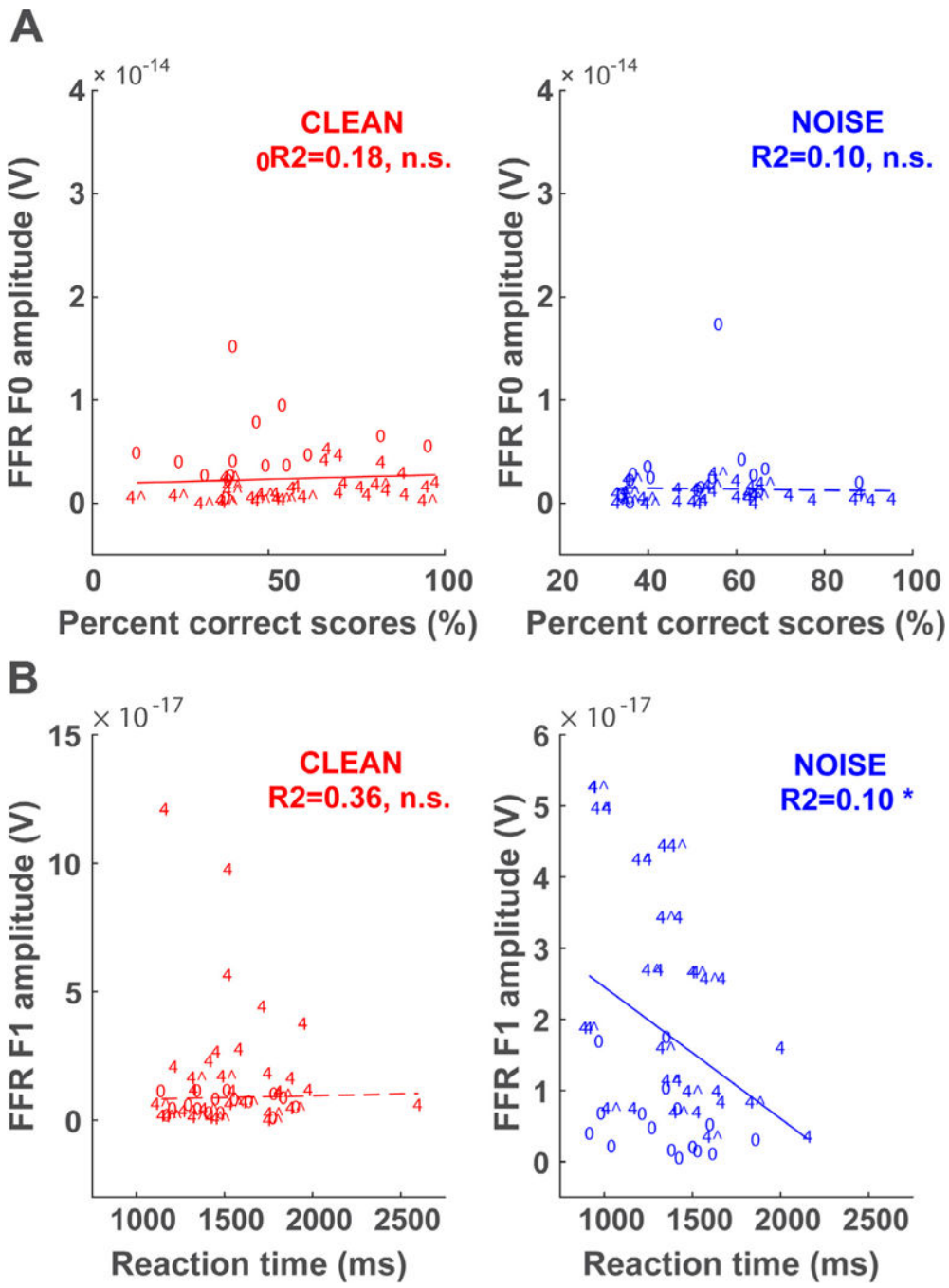

Fig. 5.

Brain-behavior correlations underlying double-vowel perception. Scatter plots and linear regression functions showing the relationship between (A) FFR F0 amplitudes and behavioral accuracy and (B) FFR F1 amplitudes and behavioral RTs for clean and noise-degraded speech. Data points are labeled according to each condition ('0' = 0ST; '4^' = 4ST @ 150 Hz; '4' = 4ST @ 190 Hz). *$p < 0.05$, n.s. − non-significant.

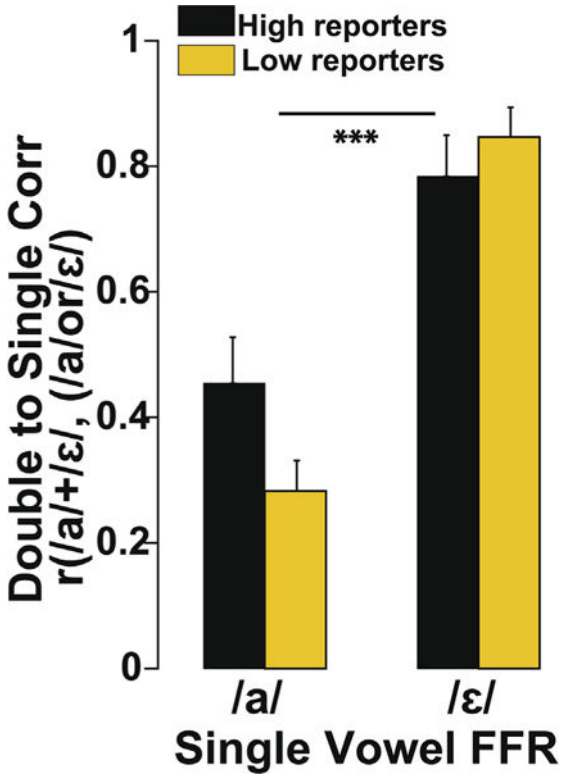

Fig. 6.

FFRs are modulated by stimulus salience rather than perceptual dominancy. Response-to-response Pearson's correlations between each listeners' (individual) single-vowel FFR spectra (FFR_a, $FFR_ε$) and their double-vowel response spectrum ($FFR_{a+ε}$). Shown here are the clean, 4 ST responses. The group split is based on the median highest and lowest 50% of listeners reporting /a/ (or /ɛ/) in the behavioral identification. Regardless of listeners' perceptual bias, FFRs showed better correspondence to the /ɛ/ vowel stimulus than /a/. Error bars = ± 1 s.e.m. ***$p \leq 0.0001$.