Interactive effects of linguistic abstraction and stimulus statistics in the online modulation of neural speech encoding

Joseph C. Y. Lau^{1,2} · Patrick C. M. Wong^{1,2} · Bharath Chandrasekaran³

Published online: 18 December 2018
© The Psychonomic Society, Inc. 2018

Abstract

Speech processing is highly modulated by context. Prior studies examining frequency-following responses (FFRs), an electrophysiological ‘neurophonic’ potential that faithfully reflects phase-locked activity from neural ensembles within the auditory network, have demonstrated that stimulus context modulates the integrity of speech encoding. The extent to which context-dependent encoding reflects general auditory properties or interactivities between statistical and higher-level linguistic processes remains unexplored. Our study examined whether speech encoding, as reflected by FFRs, is modulated by abstract phonological relationships between a stimulus and surrounding contexts. FFRs were elicited to a Mandarin rising-tone syllable (/ji-TR/, ‘second’) randomly presented with other syllables in three contexts from 17 native listeners. In a *contrastive* context, /ji-TR/ occurred with meaning-contrastive high-level-tone syllables (/ji-H/, ‘one’). In an *allotone* context, TR occurred with dipping-tone syllables /ji-D/, a non-meaning-contrastive variant of /ji-TR/. In a *repetitive* context, the same /ji-TR/ occurred with other speech tokens of /ji-TR/. Consistent with prior work, neural tracking of /ji-TR/ pitch contour was more faithful in the *repetitive* condition wherein /ji-TR/ occurred more predictably ($p = 1$) than in the *contrastive* condition ($p = 0.34$). Crucially, in the *allotone* context, neural tracking of /ji-TR/ was more accurate relative to the *contrastive* context, despite both having an identical transitional probability ($p = 0.34$). Mechanistically, the non-meaning-contrastive relationship may have augmented the probability to /ji-TR/ occurrence in the *allotone* context. Results indicate online interactions between bottom-up and top-down mechanisms, which facilitate speech perception. Such interactivities may predictively fine-tune incoming speech encoding using linguistic and statistical information from prior context.

Keywords Neural speech encoding · Frequency-following response (FFR) · Lexical tone · Context-dependent plasticity · Linguistic abstraction · Allotones

Introduction

Background

The General Auditory approach to speech perception posits that domain-general properties of the auditory system evolved in human and non-human species to handle various environmental sounds are essential to speech perception (e.g.,

Diehl, 1987; Holt, Lotto, & Kluender, 1998; Lotto, 2000; Diehl, Lotto, & Holt, 2004). Speech processing in everyday auditory environments requires the neural auditory system to fine-tune and reorganize sensory signal on the fly based on immediate auditory context (Hickok, 2012; Chandrasekaran, Skoe, & Kraus, 2014). One general property of the auditory system that contributes to context-dependent processing is its ability to compute and extract statistical relationships of objects in stimuli, from which expectancies of future sounds can be built and continuously tested (Friston, 2005; Lupyan & Clark, 2015; Denham & Winkler, 2017). Various cortical mechanisms related to context-dependent modulation such as a fronto-temporal network that detects changes and regularity deviance in the stimuli (Giard, Perrin, Pernier, & Bouchet, 1990; Escera, Yago, Corral, Corbera, & Nuñez, 2003), and adaptation mechanisms of neurons

✉ Patrick C. M. Wong
p.wong@cuhk.edu.hk

Bharath Chandrasekaran
b.chandra@pitt.edu

in the auditory cortex that suppress repetitive stimuli while enhancing unexpected stimuli have been identified (Jääskeläinen et al. 2004). Yet, since the subcortical auditory system had been traditionally regarded as passive relay stations of sensory input (Hickok & Peoppel, 2007; Rauschecker & Scott, 2009), not until the past decade has much attention been directed to the extent to which statistics-dependent modulation of speech processing also pertain in the subcortical auditory system (Krishnan & Gandour, 2009; Chandrasekaran & Kraus, 2010a; Chandrasekaran et al., 2014). To this end, scalp-recorded frequency-following response (FFR), which reflects pre-attentive phase-locked neural responses dominantly generated by neuronal ensembles within the auditory brainstem and midbrain (Chandrasekaran & Kraus, 2010a) with potential contributions from the thalamus and auditory cortex (Bidelman, 2015, 2018), has proved to be a useful metric that provides a high-fidelity ‘snapshot’ reflecting the efficiency of auditory processing in the neural auditory system comprehensively (Kraus et al., 2017). A growing body of studies examining FFRs have investigated the extent to which speech encoding is influenced by stimulus statistics (Chandrasekaran, Hornickel, Skoe, Nicol, & Kraus, 2009; Parbery-Clark, Strait, & Kraus, 2011; Strait, Hornickel, & Kraus, 2011; Slabu, Grimm, & Escera, 2012; Skoe, Chandrasekaran, Spitzer, Wong, & Kraus, 2014; Lau, Wong, & Chandrasekaran, 2017; Xie, Reetzke, & Chandrasekaran, 2018). Converging results from these FFR studies have shown that representations of stimulus features (e.g., formant transitions and linguistic pitch patterns) in FFRs are enhanced (i.e., higher FFR integrity) when the stimulus was highly predictable (Chandrasekaran et al., 2009; Parbery-Clark et al., 2011; Strait et al., 2011; Slabu et al., 2012; Skoe et al., 2014; Lau et al., 2017). These results have largely been interpreted as reflecting general auditory properties sensitive to stimulus statistics in early sensory encoding to speech information, likely also at the subcortical level (Chandrasekaran et al., 2014; Lau et al., 2017; Xie et al., 2018).

In addition to stimulus statistics, the ambient speech environment contains other contextual information. Psycholinguistic and speech perception models have demonstrated that abstract information from larger language context can dynamically influence speech processing (Ganong, 1980). Models such as the adaptive processing theory posit that listeners adapt online to talker information (e.g., representations to the acoustic characteristics of the talker’s vocal organ) to calibrate incoming stimuli to overcome inter- and intra-talker variability present in the acoustic signal (Nusbaum & Magnuson, 1997). Connectionist speech perception models such as the TRACE model posit that the perception of highly overlapping, co-articulated, and degraded acoustic signals of speech is facilitated by lexical-semantic

and phonemic traces activated by prior speech context in a bi-directional, interactive manner (McClelland & Elman, 1986). Neurolinguistics studies in previous decades have shown that the predictability of abstract phonemic (e.g., syllable onset and rimes), syntactic (e.g., syntactic categories) and lexico-semantic properties (e.g., lexical meaning) of speech stimuli in a prior sentence context may modulate ERP components such as the phonological mismatch negativity (PMN) (Connolly & Phillips, 1994; Diaz & Swaab, 2007), early left anterior negativity (ELAN), and N400 (Hagoort & Brown, 2000; Friederici, 2002). A more recent line of research started by the seminal study by Eulitz and Lahiri (2004) has provided converging evidence that abstract matrices of phonological features of phoneme categories can be activated by speech contexts. Such activated memory traces of phonological feature matrices are then mapped onto the eventual speech stimuli to modulate the magnitude of mismatch negativity (MMN) responses. The fact that these electrophysiological components are modulated by linguistic abstractions from prior contexts likely reflects modulatory influences via language-related cortical networks (Friederici, 2002; Lau, Phillips, & Poeppel, 2008; Näätänen, Paavilainen, Rinne, & Alho, 2007). However, the extent to which the encoding of speech at more fundamental sensory levels which precede conscious language processing can be modulated by linguistic abstraction remains elusive. Specifically, whether early sensory encoding of speech (e.g., at the subcortical auditory system) is a mere reflection of general auditory properties (such as the sensitivity to stimulus statistics) or interactivities between both statistical and higher-level speech perception processes that tap into abstract linguistic representations of sounds is an open question the current study aims to address.

The current study

In the current study, we tested the extent to which the encoding of pitch patterns in speech, indexed by FFRs, is modulated by abstract linguistic relationship of the preceding listening context beyond the statistical information available in the context. The type of abstract linguistic property investigated is *allotony* of Mandarin lexical tones. Lexical tones are categories of pitch patterns which function to distinguish lexical meaning in tone languages. In tone languages like Mandarin, a given syllable, when carrying different pitch patterns, can cue changes to word meaning. For example, the syllable “*shi*” means ‘poetry’ when produced with a high-level pitch pattern (Tone 1), but means ‘history’ when associated with a dipping-pitch pattern (Tone 3). The main acoustic correlates of pitch patterns of lexical tones are time-varying *f*₀ contours that fall within FFRs’ phase-locking range (Chandrasekaran & Kraus, 2010a). *Allotones* are

abstract and intricate linguistic sub-categories within a lexical tone category. The third lexical tone category in Mandarin (T3) is a representative example which involves allotony. Mandarin T3 (in the standard variety Putonghua or varieties around the Beijing area) has multiple lexical tone variants (i.e. allotones). These lexical tone variants occur in neighboring lexical tone and morphological environments. A T3 is realized as a rising tone (**TR**) when it precedes another T3 (a process known as tone sandhi), but as a dipping tone (**TD**) elsewhere (Table 1), yet without a change in lexical meaning. ERP evidence has suggested that in language production, the different realizations of T3 are achieved by abstract combinatorial and selection processes that determine which allotone should surface based on morphological combinations, instead of fossilized sequences of pitch patterns chunks stored in the lexicon (Zhang, Xia, & Peng, 2015).

In the current study, we leverage Mandarin T3 allotony (namely allotones vs. non-allotone relations) to test how statistical (namely transitional probabilities) and linguistic abstractions interact to modulate speech encoding, as reflected by the FFR. A relevant note here is that the neural source of the FFRs is currently under scrutiny. Previously assumed to reflect subcortical processes, recent work suggest significant cortical contribution to the tracking of low-frequency information in speech (Chandrasekaran, Sampath, & Wong, 2010b; Coffey et al., 2016; Coffey, Musacchia, & Zatorre, 2017; Bidelman, 2015; Bidelman, 2018). We examined the fidelity of FFRs to the same token of TR (i.e., the target tone) in three types of listening contexts. The three contexts contained stimuli with different statistical and linguistic properties relative to the target tone. The three contexts were namely: (1) a *Contrastive* context, wherein, the target tone occurred randomly with T1 (which belongs to a separate lexical tone category from the target tone) at a 34% probability; (2) an *allotone* context, wherein, the target tone occurred randomly with TD (which is an allotone to TR) at a 34% probability; (3) a *repetitive* context, wherein, the target tone was presented at a 100% probability.

Previous FFR studies have shown that neural encoding of lexical tones reflected in FFRs is modulated by prior listening contexts. When the target tone is *more predictable* within the listening context (e.g., having a

higher transitional probability of occurrence), integrity of FFRs is higher (i.e., more faithful representation of the stimulus f0 contour in the FFRs) relative to when the same tone is *less predictable* within another listening context (Skoe et al., 2014; Lau et al., 2017; Xie et al., 2018). These results suggest that *general auditory properties* of the neural auditory system such as its sensitivity to stimulus statistics may modulate how sensory signal is encoded (Chandrasekaran et al., 2014; Lau et al., 2017; Xie et al., 2018). As such, a general auditory property-based hypothesis would predict that the FFR to TR would have the highest integrity in the *TR repetitive* context, wherein the transitional probability of a TR occurrence was 100%. Integrity would be lower, and equally so, in the *T1 contrast* and *TD allotone* contexts given a lower transitional probability of a TR occurrence in both conditions.

An intriguing possibility is that *in addition to* general auditory properties, *language-specific* abstract linguistic representations of sounds may modulate early speech encoding. Neurolinguistic and theoretical linguistic works have suggested that the underlying mental representation of Mandarin T3 consist of properties of both TR and TD, e.g., exemplars of both TR and TD (Politzer-Ahles & Zhang, 2012; Li & Chen, 2015), or a set of phonological features shared by TR and TD (Yip, 2002). As such, a repetitive TD context in the *allotone* condition may elicit *some* properties of TR, which could potentially (dynamically) interact with the stimulus statistics to augment the probability of an upcoming TR occurrence above the transitional probability. Since higher stimulus probability of occurrence promotes FFR integrity, FFRs to target TR would be predicted to have higher integrity in the *allotone* context than in the *contrastive* context.

Methods

Participants

Seventeen native speakers of Mandarin (eight male; age: M: 22.53 years, SD 2.35) were selected to join the current experiment. All participants were born and raised in northern areas of Mainland China and self-reported to exclusively speak the Putonghua variety of Mandarin as

Table 1 Mandarin T3 allotony

Neighboring tone context	Example	Surface tone pattern
T1+/T3/	<i>zhōng shǐ</i> ‘Chinese history ’	T1+[TD]
/T3+/T3	<i>shǐ dǎng</i> ‘ historical record’	[TR]+T3

T1: Mandarin Tone 1, high-level tone. Romanization of Mandarin is in *Hanyu Pinyin*. Phonemic categories are marked within slashes, e.g., /T3/; actual realizations are marked within square brackets, e.g., [TR]

their native language. All participants self-reported normal hearing in both ears. In addition, all participants have pure-tone air conduction thresholds of 25 dB or better at frequencies of 500, 1000, 2000, and 4000 Hz. Informed consent approved by The Joint Chinese University of Hong Kong - New Territories East Cluster Clinical Research Ethics Committee was obtained from each participant before any experimental procedure. All participants were compensated for their time.

Stimuli

Speech stimuli used for electrophysiological testing consisted of three Mandarin lexical tone categories, namely a high-level pitch pattern (Tone 1, henceforth T1), a high-rising pitch pattern (henceforth TR), and a dipping pitch pattern (Tone 3 citation form, henceforth TD). It should be noted that TR can be the manifestation of two tone categories, namely Tone 2 (T2), and the allotone of Tone 3 triggered by tone sandhi. The three tones had the same syllable /ji/, which in combination with the lexical tones, lead to three different Mandarin words: *yī* /ji1/ (T1, ‘doctor’), *yí* /ji2/ (T2, ‘aunt’), and *yǐ* /ji3/ (T3, ‘second (the ordinal number)’). The syllable /ji/, which in combination with the TR, could also be the sandhi form of the word *yǐ* /ji3/. To induce acoustic variability, we used multiple tokens of stimuli to represent each category. The use of resynthesized stimuli rather than natural tokens allowed maximal acoustic control across the categories.

A male native speaker of Beijing Mandarin produced the three syllables (*yī*, *yí*, and *yǐ*), which were then resynthesized in Praat (Boersma & Weenink, 2014). All syllables were first segmented, and normalized for duration (175 ms) and intensity (74 dB SPL). Then for each syllable, the f0 (fundamental frequency) values for 14 points (10-ms intervals starting from 22.5 ms) along the 175-ms syllable were estimated using an autocorrelation-based method built-in in Praat. The 14 points of f0 values for T1 and TD were then adjusted (with the overall shapes of the f0 contours maintained) such that the averaged Euclidean distance of the 14 points between T1 and TR was identical to that between TD and TR (0.933 ERB). Then, based on these acoustic-distance matched f0 contours (i.e., the ‘anchor’ contours), four additional 14-point f0 contours were each created for the three lexical tone categories. For each lexical tone category, the four additional f0 contours had a Euclidean distance of +0.1 ERB, +0.2 ERB, -0.1 ERB, and -0.2 ERB respectively for each of the 14 points of the f0 contour relative to the ‘anchor’ contour. With this design, acoustic variability was induced by the five f0 contours in each lexical tone category, while acoustic distance could

also be maintained (0.933 ERB on average) between the T1/TR (*contrast*) vs. TD/TR (*allotone*) distinctions. The resulting 15 f0 contours were presented in Fig. 1.

Each of the 15 f0 contours was then superimposed on a /ji1/ syllable and resynthesized with the overlap-add method (Moulines & Charpentier, 1990) in Praat. As such, the f0 contour was the main acoustic feature that differed across the stimuli. Native speakers of Mandarin confirmed all tokens of stimuli to be natural exemplars of their respective lexical tone categories.

Design

FFRs to the ‘anchor’ TR syllable were elicited in three context conditions. The three contexts were namely a TR context, a T1 context, and a TD context (Fig. 2).

In the TR context condition, 1530 sweeps of the ‘anchor’ TR were presented pseudorandomly in the context of 2970 sweeps of ‘non-anchor’ TR syllables. In the T1 context condition, 1530 sweeps of the ‘anchor’ TR were pseudorandomly presented in the context of 2970 sweeps of ‘non-anchor’ T1 syllables. Likewise, in the TD context condition, 1530 sweeps of the ‘anchor’ TR were presented pseudorandomly in the context of 2970 sweeps of ‘non-anchor’ TD syllables. For each condition, the 2970 sweeps of ‘non-anchor’ tones were comprised of 765 sweeps each of +0.2 ERB and -0.2 ERB ‘non-anchor’ tones (relative to the ‘anchor’ tone for each tone category), and 720 sweeps each of +0.1 and -0.1 ERB ‘non-anchor’ tones. As such, acoustically, the ‘anchor’ TR stimulus was presented in all three context conditions at a probability of 34%. However,

Fig. 1 Stimuli characteristics: F0 parameters (transformed into Hz) of all stimuli across fourteen time points along the 175 ms of the stimuli. Each category [Tone 1 (T1, high-level tone); rising tone (TR), and dipping tone (TD)] was represented by five stimuli. The black f0 contours denote the ‘anchor’ contour for each category

Fig. 2 Experimental design: event-matched paradigm: Rising tone (TR) stimuli were presented pseudorandomly in *TR* (repetitive) (top), *Tone 1* (T1, Contrastive) (middle), and *dipping tone* (TD, allotone) (bottom) contexts. To control for presentation order, electrophysiological responses to TRs were event-matched across all three contexts (dotted lines), achieved by using the same pseudorandom order in the presentation of all three conditions. A: ‘anchor’ stimulus; NA: ‘non-anchor’ stimuli

in terms of lexical tone category, the occurrence of a TR category was 100% in the *TR* context condition, whereas the occurrence of a TR category in *T1* and *TD* context conditions remained at 34%. The same pseudorandom order was used for all three context conditions for each participant, such that the relative location of the ‘anchor’ TR trials within the stream of all stimuli could be identical across conditions (Fig. 2) (Chandrasekaran et al., 2009).

Electrophysiological recording procedures

Electrophysiological recording took place in an acoustically and electromagnetically shielded booth. During recording, participants were told to ignore the stimuli and to rest or sleep in a reclining chair, consistent with prior FFR recording protocols (Krishnan, Xu, Gandour, & Cariani, 2004; Skoe & Kraus, 2010). Stimuli were presented in a single polarity to the participant’s right ear through electromagnetically shielded insert earphones (ER-3A, Etymotic Research, Elk Grove Village, IL, USA) at 80 dB SPL. Stimuli in all conditions were presented with a 74–114-ms inter-stimulus interval (ISI). Stimuli were presented via the presentation software Neuroscan Stim2 (Compumedics, El Paso, TX, USA). The total duration of the testing including preparation time lasted approximately 70 minutes including preparation for each participant.

Electrophysiological responses were recorded using a SynAmps2 Neuroscan system (Compumedics, El Paso, TX, USA) with Ag-AgCl scalp electrodes, and digitized at a sampling rate of 20,000 Hz using CURRY Scan 7 Neuroimaging Suite (Compumedics, El Paso, TX, USA). A vertical electrode montage (Skoe & Kraus, 2010) that differentially recorded electrophysiological responses from the vertex (Cz, active) to bilateral linked mastoids (M1+M2,

references), with the forehead (aFz) as ground was used. Contact impedance was less than 2 k Ω for all electrodes.

Preprocessing procedures

Filtering, artifact rejection, and averaging were performed offline using CURRY 7 (Compumedics, El Paso, TX, USA). Responses were bandpass filtered from 80 to 2500 Hz (12 dB/octave), consistent with prior FFR analysis protocols (Krishnan et al., 2004; Skoe & Kraus, 2010). Trials with activities greater than $\pm 35 \mu\text{V}$ were considered artifacts and rejected. Responses to the TR stimulus were averaged with a 275-ms epoching window encompassing -50 ms before stimulus onset, the 175 ms of the stimulus, and 50 ms after stimulus offset. Responses to TR in the *TR* context condition were averaged according to their occurrence relative to the order of presentation in the *T1* and *TD* context conditions. The average number of accepted trials in the *T1* ($M = 1269.22$, $SD = 273.891$), *TR* ($M = 1189.06$, $SD = 433.666$), and *TD* ($M = 1285$, $SD = 357.633$) context conditions did not differ, as revealed by a one-way repeated-measures analysis of variance (ANOVA) with Greenhouse–Geisser correction [$F(1.872, 31.83) = 1.231$, $p = 0.303$].

Data Analysis

FFR data were further analyzed using customized MATLAB (The MathWorks, Natick, MA, USA) scripts adapted from the Brainstem Toolbox (Skoe & Kraus, 2010). Before analysis, the stimulus was down-sampled to 20,000 Hz (from 44,100 Hz) to match the sampling rate of the response. For each FFR, computation began with an estimate of the FFR’s onset delay relative to the stimulus presentation time (*neural lag*) due to neural conduction of

the auditory pathway. This *neural lag* value was computed using a cross-correlation technique that slid the response waveform (the portion of FFR wave from 0–175 ms) and the stimulus waveform in time with respect to one another (Liu et al., 2014). The *neural lag* value (in ms) was taken as the time point in which maximum positive correlation was achieved between 6 and 12 ms, the expected latency of the onset component of the auditory brainstem response, with the transmission delay of the insert earphones also taken into account (Bidelman, Gandour, & Krishnan, 2011; Strait, Parbery-Clark, Hittner, & Kraus, 2012). Then, the *f0* contour of each FFR was estimated using a fast Fourier transform-based (FFT) (Wong, Skoe, Russo, Dees, & Kraus, 2007; Liu et al., 2014) procedure. To estimate how *f0* values changed through the waveform, the post-stimulus portion of the FFR waveform (shifted by *neural lag* in the FFR) was first divided into Hanning-windowed bins in the frequency domain, each 50 ms (49 ms overlap between adjacent time bins). Then, a narrow-band spectrogram was calculated for each bin by applying the FFT. Before the FFT, each bin was zero-padded to 1 s to interpolate missing frequencies. For each bin, the spectral peak in the spectrogram that was closest to the expected *f0* (from the stimulus) was determined as the response *f0* value of that bin. The resulting *f0* values from each bin formed an *f0* contour. The *f0* contour of the stimulus was also derived separately using the same procedure, but the analysis window of the waveform was not shifted by the *neural lag*.

Subsequent analyses focused on whether and how *neural pitch tracking* varied as a function of the three experimental conditions. Two main metrics previously used to define the fidelity of the neural responses to linguistic pitch patterns were derived from the *f0* contours (Wong et al., 2007; Song, Skoe, Wong, & Kraus, 2008; Skoe et al., 2014; Liu et al., 2014): (1) *Stimulus-to-response correlation*, and (2) *f0 error*. These two metrics have been shown to be stable across different days of data collection, hence demonstrating their reliability as objective metrics to neural pitch encoding fidelity (Xie, Reetzke, & Chandrasekaran, 2017). *Stimulus-to-response correlation* (values between -1 and 1) is the Pearson's correlation coefficient (r) between the stimulus and response *f0* contours. It indicates the similarity between the stimulus and response *f0* contours in terms of the strength and direction of their linear relationship (Wong et al., 2007; Liu et al., 2014). *F0 error* (in Hz) is the mean absolute Euclidean distance between the stimulus and response *f0* contours across the total number of bins in the FFT-based analysis. This metric represents the pitch encoding accuracy of the FFR by reflecting how many Hz the FFR *f0* contour deviates from the stimulus *f0* contour on average (Song et al., 2008; Skoe et al., 2014).

In addition, the *signal-to-noise ratio* (*SNR*) of each FFR was also derived to assess whether the overall magnitude

of neural activation over the entire FFR period (relative to pre-stimulus baseline) (Russo, Nicol, Musacchia, & Kraus, 2004) varied as a function of stimulus context. To derive the *SNR* of each FFR, the root mean square (RMS) amplitudes (the mean absolute value of all sample points of the waveform within the respective time windows, in μV) of the FFR period and the pre-stimulus baseline period of the waveform were first recorded. The quotient of the FFR RMS amplitude and the pre-stimulus RMS amplitude was taken as the *SNR* value (Russo et al., 2004).

Statistical analysis

Before subsequent parametric statistical analyses, *stimulus-to-response correlation* values were first converted into Fisher's z' scores (Wong et al., 2007), as Pearson's correlation coefficients do not comprise a normal distribution. To examine the extent to which FFR pitch encoding and phase-locking varied as a function of the three types of stimulus context (*TR*, *T1*, and *TD* contexts), separate one-way repeated measures ANOVAs on the FFR metrics (*stimulus-to-response correlation*, *f0 error*, and *SNR*) were conducted.

Results

The grand averaged waveforms and spectrograms of the stimulus and the FFRs of the three context conditions are presented in Fig. 3 (panels A and B). Figure 3 (panel C) shows the mean *f0 error*, *stimulus-to-response correlation* and *SNR* of all conditions. Data, stimuli, and MATLAB scripts for data analyses are available from Lau on request.

Repeated measures ANOVA on *stimulus-to-response correlation* with the Greenhouse–Geisser correction revealed significant differences across the three context conditions [$F(1.776, 30.185) = 6.156, p = 0.007$]. Post hoc pairwise comparisons with Bonferroni corrections revealed that *stimulus-to-response correlation* of the *Contrastive* context condition was significantly lower than that of the *allotone* context condition ($p = 0.036$), and that *stimulus-to-response correlation* of the *contrastive* context condition was also significantly lower than that of the *repetitive* context condition ($p = 0.014$). The *stimulus-to-response correlation* differences across the *repetitive* and *allotone* context conditions were not significant ($p = 1.000$).

Repeated measures ANOVA on *f0 error* with the Greenhouse–Geisser correction revealed significant differences across the three context conditions [$F(1.524, 25.909) = 4.012, p = 0.040$]. Post hoc pairwise comparisons with Bonferroni corrections revealed that *f0 error* of the *contrastive* context condition was significantly higher than that of the *allotone* context condition ($p = 0.006$). The

Fig. 3 Results: Event-matched frequency-following responses: **a** waveforms and **b** spectrograms of grand-averaged event-matched frequency-following responses (FFRs) to a rising tone (TR) stimulus (1st row) in TR, T1, and TD context conditions (2nd to 4th rows). **c** Mean f_0 error, stimulus-to-response correlation, and signal-to-noise ratio of event-matched TR FFRs from TR (left bars), T1 (middle bars), and TD (right bars) contexts. Error bars denote ± 1 standard error from the mean. $*p < 0.05$ (in post hoc pairwise comparisons, Bonferroni corrected)

f_0 error differences between the *contrastive* and *allotone* context conditions ($p = 0.376$), and between *allotone* and *repetitive* context conditions ($p = 1.000$) were not significant.

Repeated measures ANOVA on SNR with the Greenhouse–Geisser correction was not significant [$F(1.711, 29.080) = 0.493$, $p = 0.587$].

Discussion

Context-dependent sensory encoding of speech signals

Our results unambiguously demonstrate that the neural representation of linguistic pitch patterns varies as a function of

stimulus statistics. Specifically, we found that integrity of FFR was higher (indexed by higher stimulus-to-response correlation) when the transitional probability of TR stimulus was 100% in the *repetitive* context, relative to in the *contrastive* context in which TR only occurred with a 34% transitional probability.

This result replicates a series of prior studies (Chandrasekaran et al., 2009; Parbery-Clark et al., 2011; Strait et al., 2011; Slabu et al., 2012; Skoe et al., 2014; Lau et al., 2017; Xie et al., 2018) which found that the integrity of FFRs to speech stimuli was higher when transitional probability of the stimuli was higher in prior auditory contexts. The current results converge with prior findings to provide critical evidence for online auditory plasticity, i.e., the malleability of auditory processing to listening environment, in speech encoding. Consistent with prior studies, we

collected FFRs using a passive listening paradigm, i.e., participants did not pay overt attention to the stimulus stream. Our findings of stimulus probability-related effect are therefore likely to be underlain by highly automatic mechanisms which operate even without overt attention or explicit goal-directed behavior to modulate speech processing online.

Previous studies have suggested that various mechanisms may be at play in modulating FFR in different stimulus contexts. One neural mechanism known to contribute to context-dependent modulation in FFR is stimulus-specific adaptation (SSA) (Lau et al., 2017; Xie et al., 2017). SSA is a fundamental novelty-detection mechanism which *attenuates* repetitive sensory presentation, while *enhancing* the encoding of novel stimuli (Natan et al., 2015). Animal models have suggested that neurons in the inferior colliculus (IC) of the midbrain demonstrate SSA to commonly recurring auditory stimuli (Pérez-González, Malmierca, & Covey, 2005; Malmierca, Cristaudo, Pérez-González, & Covey, 2009), and that SSA at the level of the IC is likely to be a local process largely unaffected by the cortex (Anderson & Malmierca, 2013). One FFR study has found that integrity of FFRs to a lexical tone [Cantonese Tone 4 (T4)] was reduced when it was presented *repetitively* (T4T4T4T4T4T4...) relative to when presented with two other tones in a *patterned* context (T1T2T4T1T2T4...), while transitional probability were both 100% in the two conditions (Lau et al., 2017). The reduced FFR integrity in the *repetitive* condition was interpreted as indexing local SSA processes at the IC which had attenuated the more repetitive T4 in the *repetitive* condition than in the *patterned* condition wherein transitional probability was held constant. However, the current result demonstrates that FFR integrity was *enhanced*, but not *attenuated* (i.e., reduced integrity, as an SSA-based account would predict) when the target tone was presented in a *repetitive* context relative to when the target tone was presented in the context of another tone. As such, the current results are unlikely to be solely underlain by SSA.

In light of the recent findings of potential contributions from phase-locked activities of the primary auditory cortex to FFRs (Coffey et al., 2016, 2017; Bidelman 2018), one may also consider the possibility of attention-related cortical mechanisms known to *inhibit* phase-locking at the auditory cortex to be a factor in modulating the context-dependent FFRs.¹ Recent electrocorticography evidence has suggested that phase-locked neural responses at human posteromedial Heschl's gyrus (HG) is more robust during anesthesia (Nourski et al., 2017). The interpretation to this result was that without anesthesia, simultaneous non-phase-locked synaptic events initiated from other cortical

regions may project to posteromedial HG to inhibit phase-locking synaptic events therein. Anesthesia had likely suppressed simultaneous non-phase-locked synaptic events initiated from other cortical regions (e.g., from the attention networks) that would inhibit phase-locking at the auditory cortex. As such, one may postulate that not only anesthesia, but also other factors that affects attention may in fact modulate phase-locking at the auditory cortex, hence the integrity of FFRs. In the current study, the more variable *contrastive* context may have been deemed more interesting, and hence elevated simultaneous non-phase-locked activities projected to the auditory cortex (e.g., stimulating more overt attention). Such elevation in non-phase-locked activities may thus attenuate the phase-locking activities at the auditory cortex reflected in FFRs. However, external evidence suggests that such cortical activities that inhibit phase-locking may not be the crucial factor that determines context-dependent modulation. As mentioned previously, one study found that with transitional probability controlled, FFRs to a lexical tone elicited in a *patterned* context (i.e., target tone presented with two other tones) had higher integrity than FFRs elicited in a *repetitive* context (Lau et al., 2017). This result suggests that potential effects of cortical phase-locking inhibition that would have inhibited FFRs at the *patterned* context (due to its more variable, hence more attention stimulating nature), even if at play, have at least been overridden by effects of SSA which have attenuated sensory representations of the more repetitive stimuli.

Instead, we posit that the converging stimulus statistics-related online context-modulation effects in FFRs are mainly underlain by neural mechanisms that enhance sensory representation of stimuli presented with higher statistical probability. Following prior studies (Lau et al., 2017; Xie et al., 2018), we interpret such stimulus statistics-related online context-modulation effects as a reflection of the underpinning of a *predictive tuning* mechanism (Chandrasekaran et al., 2014). The *predictive tuning* model postulates the auditory system automatically fine-tunes the representation of stimulus features that matches *top-down* expectation. As such, more predictable bottom-up sensory input, including speech sounds, would be enhanced when the stimuli are more predictable in prior context (e.g., with a higher transitional probability) relative to less predictable ones. This mechanism is partly subserved by a cortico-fugal feedback network that spans the auditory pathway from the auditory midbrain, thalamus, to the auditory cortex (Malmierca, Anderson, & Antunes, 2015). Besides the ascending auditory pathway that relays sensory signal from subcortical hubs to the auditory cortex, this feedback network is crucially supported by neural pathways that back-project from auditory cortical regions onto subcortical structures like the IC in a feedback loop fashion

¹We thank an anonymous reviewer for suggesting this possible interpretation.

(Winer, Larue, Diehl, & Hefti, 1998). This cortico-fugal feedback loop allows auditory encoding as early as at the subcortical level to be dynamically modified by *top-down* feedback computed by the cortex, as evidenced by animal models (Suga, 2008). As such, the predictability-enhancement effect found in speech FFRs may reflect the continuous computation and updating of expectations of upcoming speech signal at the cortex. Such expectations were then back-projected top-down through the cortico-fugal pathways to the subcortical auditory system to fine-tune eventual bottom-up speech encoding as reflected in the FFRs (Krishnan & Gandour, 2009; Chandrasekaran et al., 2014).

The *predictive tuning* mechanism is likely to be in a constant push-pull with other neural mechanisms such as SSA local to the IC (Lau et al., 2017) as well as cortical inhibition at the HG (Nourski et al., 2017) in mediating sensory encoding in different auditory contexts, as indexed by the FFR. Indeed, a recent FFR study found that the predictability enhancement could be reversed when FFRs were elicited in an irrelevant visual task with high processing load (Xie et al., 2018). The high processing load in the irrelevant task presumably took away cortical resources needed for computations involving *predictive tuning*. With *predictive tuning* through the cortico-fugal feedback loop attenuated, SSA processes local to the IC (which are not affected by cognitive load presumably at the cortical level) likely persisted: FFRs presented in a repetitive context were attenuated relative to in a variable context. On the other hand, when FFRs were elicited in passive listening paradigms without any orienting task, the additive effect of SSA on top of *predictive tuning* was apparent when the level of repetitiveness varied while stimulus statistics was held constant (Lau et al., 2017). In the current study in which stimulus statistics varied, the predictability enhancement effect in FFRs replicated was likely contributed by *predictive tuning*, which had overridden the effects of SSA.

It is worth mentioning that an emerging view is that FFRs, although dominantly contributed by activities by the IC (Chandrasekaran & Kraus, 2010a; Bidelman, 2015, 2018), reflect an integrated and dynamic interplay between cortical and subcortical circuitry (Kraus & White-Schwoch, 2015), as FFRs are also partly contributed by the medial geniculate body of the thalamus and the auditory cortex (Coffey et al., 2016, 2017; Bidelman, 2018). As such, one may argue that the neural mechanism that enhances FFRs presented in the more predictable repetitive context may in fact solely be cortical in origin. Such potential predictability-enhancing cortical mechanism, and possibly the aforementioned cortical phase-locking inhibition mechanisms, may have induced additive effects which have even overridden SSA effects local to the IC to enhance FFRs in

repetitive context despite the fact that FFR signals are dominantly subcortical (Bidelman, 2018). To gain a more definite understanding to this issue, future studies can employ high density EEG recordings and source localization techniques to delineate the unique contribution of the subcortical auditory system (e.g., IC) as well as the cortex (e.g., the auditory cortex) to FFRs which are modulated by stimulus-statistics.

Nevertheless, in a broader perspective, the demonstration of context modulation effect in speech encoding also lends support to the working hypothesis of the General Auditory approach to speech processing. The General Auditory approach to speech perception posits that speech sounds are perceived using domain-general mechanisms of audition evolved in humans as well as other species to handle various environmental sounds (e.g. Diehl, 1987; Holt et al., 1998; Lotto, 2000; Diehl et al., 2004). Animals models have suggested that contextual modulation in the auditory system is pervasive in other non-human mammalian species such as rats (Pérez-González et al., 2005; Malmierca et al., 2009) and cats (Rubin, Ulanovsky, Nelken, & Tishby, 2016). From an evolutionary perspective, contextual modulation to sensory encoding allows organisms to extract information from the past (e.g., sounds associated with predators) that is relevant for future survival (Rubin et al., 2016). The demonstration of contextual modulation in speech encoding hence suggests that speech perception is at least partly supported by fundamental properties of general audition as fundamental as at the subcortical auditory system shared among humans and other species.

Interactive effects of linguistic abstraction and stimulus statistics

Importantly, our results reveal that besides stimulus statistics, integrity of FFRs also varies as a function of abstract linguistic relationships between the target tone and stimuli from the surrounding context. Given the same transitional probability of TR occurrence in the listening context (both 34%), integrity of FFR in the *allotone* context condition is higher than in the *contrastive* context.

Due to the identical transitional probability controlled across the *allotone* and *contrastive* context conditions, the previously discussed neural mechanisms, which are known to contribute to context-dependent modulation, are not likely to be at play. Top-down modulation based on stimulus statistics can likely be ruled out due to the identical transitional probability across the two conditions. Meanwhile, the results cannot be attributed to local novelty detection mechanisms such as SSA, since SSA is likely to operate with an *equal* level of intensity across the *allotone* and *contrastive* context since the less frequently occurring target tones therein should be as novel as in each other given the same transitional probability across the

two conditions. Also, unlike in the *repetitive* context which only contained one category, both *allotone* and *contrastive* contexts involved the presentation of two tone categories. The 4500 trials that consisted of the two categories were also presented with an identical pseudorandom stimulus presentation order. As such, the cortical phase-locking inhibition mechanism, which presumably attenuates FFRs when more attention is triggered, is unlikely to be at play since the level of attention triggered should not have differed across the two conditions.

As such, FFR integrity across the *allotone* and *contrastive* contexts likely varied as a function of the independent variable manipulated across the two conditions, i.e., the abstract linguistic relationships among T1, TR, and TD. The psycholinguistics literature has suggested various cognitive mechanisms which may underlie the modulation of speech processing by abstract linguistic abstractions established in prior contexts. For example, the TRACE model (McClelland & Elman, 1986) suggests that there are lexico-semantic and phonological traces activated by prior speech input. Such “traces” can influence or determine the processing of subsequent speech signals. More directly relevant to the current study, the COHORT model and its variants (Marslen-Wilson & Welsh, 1978; Gaskell & Marslen-Wilson, 2002) suggest that during the initial stage of the word recognition process, a “cohort” of words sharing a particular sound sequence with the stimulus will be co-activated. Such co-activation, postulated as the “trace” or the “cohort”, give rise to the *phonological priming* effect (Slowiczek, Nusbaum, & Pisoni, 1987), wherein the processing of a target word which shares certain phonological features with the prime word is facilitated. Specifically on the topic of Mandarin allotony, prior behavioral work utilizing a priming paradigm has suggested that prime words consisting of a TD facilitated lexical decision of the target word which contains a TR (i.e., faster lexical decision), despite the totally different pitch contours of the two tones (Zhou & Marslen-Wilson, 1997; Chien, Sereno, & Zhang, 2016). The priming effect between TD-TR suggests that elicitation of TD in speech processing (i.e., in the “trace” or “cohort” termed in the different models) may also co-activate TR, both of which are presumably stored as context-dependent variants (i.e., allotones) within the same abstract lexical tone category (Chen et al., 2011).

This TD-TR co-activation that has presumably led to the priming effect has also been shown to manifest neurophysiologically. Prior studies have found that presentation of a deviant TR in the context of a TD standard stimulus in an oddball paradigm elicited less robust MMN, relative to when a deviant T1 was presented in the context of a standard TD (Li & Chen, 2015; Politzer-Ahles, Schluter, Wu, & Almeida, 2016). One interpretation to these results was that despite a lack of lexico-morphological context in the

oddball paradigm which would trigger tone sandhi (which are available in priming experiments), a highly repetitive TD may nevertheless co-activate shared properties of *both* TD and TR in the memory trace since both TD and TR are allotones of T3. The co-activation traces of TR elicited by the allotone TD in the listening context might have mitigated the MMN response elicited by TR. The MMN response is known to be inhibited when the probability of occurrence of deviant stimuli is higher, i.e., more predictable occurrence (López-Caballero, Zarnowicz, & Escera, 2016). As such, one interpretation is that the memory trace of TR co-activated by the standard TD has made the deviant TR *less unpredictable* (i.e., by augmenting the probability of TR occurrence beyond the transitional probability) in the oddball paradigm.

We posit that the same underlying mechanism involving TR-TD allotone co-activation (and the lack thereof by the contrastive T1) in the MMN and priming studies may also explain our results. The co-activated TR memory trace by its TD allotone might have interacted with the transitional probability and augmented the probability of the deviant TR's occurrence *beyond* 34% (its transitional probability) in the *allotone* context but not in the *contrastive* context. Such augmented probability of TR occurrence, reflecting interactivities between co-activation and transitional probability, might have **facilitated** neural encoding to TR in the *Allotone* context relative to the *contrastive* context, as reflected in FFRs.

One complication this interpretation may face is that TR, in addition to being an allotone of T3, can also be perceived as a T2. However, this possibility can be ruled out considering our results. T2, like T1, is also contrastive to TD (the citation form of T3). If the deviant TR was perceived *solely* as a T2 (hence also was contrastive as T1), then we would predict TR FFRs in T1 and TD contexts would not differ in integrity, given that the transitional probability of TR was identical in the two conditions. Our results clearly suggest this possibility to be spurious.

Therefore, our findings likely demonstrate that linguistic abstractions such as phonological relationships may constitute part of the contextual information modulating early speech encoding, as indexed by FFRs. Linguistic abstractions may interact with other contextual information such as stimulus statistics to modulate the predictability of upcoming sound top-down to facilitate speech encoding. It must be noted that in the natural speech environment, speech processing rarely occurs with highly repetitive stimuli. The auditory oddball paradigm used in this study was by no means aiming to mimic the natural speech environment. Instead, the highly repetitive auditory oddball paradigm was used for experimental control, and to maintain good signal-to-noise ratio in FFRs through averaging over a thousand of trials for the context-related effects to emerge. However, we

posit that the top-down neural speech encoding mechanism implicated by the results is operative in natural speech environments. Psycholinguistic models like the TRACE model (McClelland & Elman, 1986) suggest that transitional probabilities of speech sounds in natural speech environments can be modulated (i.e., augmented or reduced) by linguistic abstractions at lexical and phonemic levels, as evidence by the modulation of speech perception by lexical (e.g., the Ganong effect Ganong, 1980) and phonemic contexts (e.g., the phonological neighborhood effect Luce & Pisoni, 1998). As such, phonological relationships may also be among the linguistic abstractions that interact with general auditory properties of the neural auditory system (e.g., sensitivity to stimulus statistics) to facilitate speech perception. Such interactivities may facilitate speech perception by fine tuning bottom-up speech signal encoding at the sensory level as early as at the subcortical auditory system, as evidenced by results shown in the FFR which is a dominantly subcortical neurophysiological response. In our previous discussion on top-down modulation in speech encoding, we have brought forward a proposal that context-dependent effects found in FFRs may reflect top-down modulation of subcortical speech encoding through the cortico-fugal feedback loop (Chandrasekaran et al., 2014). Here we extend this proposal by suggesting that together with fundamental properties of the auditory signals such as stimulus statistics, more abstract linguistic representations would also build up the top-down contextual information that the cortico-fugal feedback is sensitive to in the sensory encoding of speech input. As such, this extended proposal would posit that one role the cortex plays to contribute to FFRs (which are thought to be dominated by subcortical sources with cortical contributions) is the computation of mental traces containing linguistic representations from the linguistic context. Such traces may modify the subcortical encoding of upcoming speech information in a feedback-feedforward loop fashion, as reflected in the linguistic-abstraction-dependent FFRs found in this study. To lend further support to this hypothesis, future studies can test for potential interactive effects in FFRs between stimulus statistics and other arrays of linguistic abstraction previously explored in the psycholinguistic literature.

There is, however, one potential alternative acoustically based interpretation to our results. Despite the controlled averaged Euclidean distance between T1-TR and TD-TR in this study, the similarity of the *overall shapes* between T1 and TR compared to TD and TR unavoidably covaried with the contrastive vs. allotone status. Perception studies on Mandarin tones have found that TR and TD, both being contour tones, are sometimes confused with each other when presented in isolation (Shen & Lin, 1991; Whalen & Xu, 1992). It must be noted, however, that none of these studies matched the acoustic distance between tone

pairs as in the current study. Nevertheless, one possible explanation would be that under a TD (a contour tone) context, a contour tone (TR) is more predictable; under a T1 context (a level tone), a contour tone (TR) is not as predictable. Despite our level tone T1 stimuli also being rising f_0 contours (albeit slightly), and our use of multiple stimulus tokens which elicit normalization processes that may facilitate lexical tone categorization (Wong & Diehl, 2003), future studies that could replicate our findings would be welcomed. Specifically, future studies could look into contrastive and non-contrastive sounds in other languages that are less dynamic and more easily controlled, e.g., vowels.

Conclusions

In summary, the current study demonstrates that online neural speech encoding of a dynamic linguistic pitch pattern is more robust when a sound is more predictable in a listening context. Predictability of the sound is not only determined by stimulus statistics, but is also likely modulated by linguistic representations elicited by prior speech context at more abstract cognitive levels. Together, we interpret these results as indicative to online interactions between bottom-up and top-down mechanisms which facilitate speech perception. Such interactivities can fine-tune up-coming speech encoding using information from prior listening environments, including an interaction between linguistic and statistic properties of the presentation context.

This current study is also among the first to demonstrate robust influence of linguistic abstraction in FFRs. Due to the interpretability of FFRs at an individual subject's level, the objectiveness of FFR as a diagnostic index, as well as the relatively convenient recording procedure of FFR, a recent trend is to explore the use of individuals' FFRs to predict future learning success, developmental trajectories, and clinical treatment outcomes (Kraus et al., 2017). While the current study was not motivated by such a trend, results of this study may provide the empirical foundation for future research to develop FFR as an index to abstract linguistic sensitivity, other than fundamental processes of the general auditory system. If successful, we speculate that linguistic abstraction-dependent FFRs, compared to conventional FFRs, may be better predictors to future learning, development, and treatment related to phonological awareness due to their involvement of higher-level abstractions which only abstraction-dependent FFRs may be sensitive to.

Acknowledgements This work was supported by the National Institute on Deafness and Other Communication Disorders Grant 1R01-DC-013315 (to B. Chandrasekaran), Research Grants Council of Hong Kong General Research Fund 14117514 (to P.C.M. Wong), Global Parent Child Resource Centre Limited (to P.C.M. Wong), Lui

Che Woo Institute of Innovative Medicine (to P. C. M. Wong), and Dr. Stanley Ho Medical Development Foundation (to P.C.M. Wong). We also wish to thank Christopher Chan, Kirin Cheung, Tianfan Liu, Grace Pan, Binghui Shen, Xiaohui Sun, and Yi Wu for their assistance with data collection.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- Anderson, L., & Malmierca, M. (2013). The effect of auditory cortex deactivation on stimulus-specific adaptation in the inferior colliculus of the rat. *European Journal of Neuroscience*, 37(1), 52–62.
- Bidelman, G. M. (2015). Multichannel recordings of the human brainstem frequency-following response: scalp topography, source generators, and distinctions from the transient ABR. *Hearing Research*, 323, 68–80.
- Bidelman, G. M. (2018). Subcortical sources dominate the neuroelectric auditory frequency-following response to speech. *NeuroImage*, 175, 56–69.
- Bidelman, G. M., Gandour, J. T., & Krishnan, A. (2011). Musicians and tone-language speakers share enhanced brainstem encoding but not perceptual benefits for musical pitch. *Brain and Cognition*, 77(1), 1–10.
- Boersma, P., & Weenink, D. (2014). *Praat: doing phonetics by computer [computer program]*. (Version 5. 3. 73, retrieved 21 April 2014 from <http://www.praat.org/>).
- 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(3), 311–319.
- Chandrasekaran, B., & Kraus, N. (2010a). The scalp-recorded brainstem response to speech: neural origins and plasticity. *Psychophysiology*, 47(2), 236–246.
- Chandrasekaran, B., Sampath, P. D., & Wong, P. C. (2010b). Individual variability in cue-weighting and lexical tone learning. *The Journal of the Acoustical Society of America*, 128(1), 456–465.
- Chandrasekaran, B., Skoe, E., & Kraus, N. (2014). An integrative model of subcortical auditory plasticity. *Brain Topography*, 27(4), 539–552.
- Chen, Y., Shen, R., & Schiller, N. O. (2011). Representation of allophonic tone sandhi variants. In *Proceedings of psycholinguistics representation of tone. Satellite Workshop to ICPHS, Hong Kong* (pp. 38–41).
- Chien, Y. F., Sereno, J. A., & Zhang, J. (2016). Priming the representation of Mandarin tone 3 sandhi words. *Language, Cognition and Neuroscience*, 31(2), 179–189.
- Coffey, E. B., Herholz, S. C., Chepesiuk, A. M., Baillet, S., & Zatorre, R. J. (2016). Cortical contributions to the auditory frequency-following response revealed by MEG. *Nature Communications*, 7.
- Coffey, E. B., Musacchia, G., & Zatorre, R. J. (2017). Cortical correlates of the auditory frequency-following and onset responses: EEG and fMRI evidence. *Journal of Neuroscience*, 37(4), 830–838.
- Connolly, J. F., & Phillips, N. A. (1994). Event-related potential components reflect phonological and semantic processing of the terminal word of spoken sentences. *Journal of Cognitive Neuroscience*, 6(3), 256–266.
- Denham, S. L., & Winkler, I. (2017). Predictive coding in auditory perception: challenges and unresolved questions. *European Journal of Neuroscience*.
- Diaz, M. T., & Swaab, T. Y. (2007). Electrophysiological differentiation of phonological and semantic integration in word and sentence contexts. *Brain Research*, 1146, 85–100.
- Diehl, R. L. (1987). Auditory constraints on speech perception. In Schouten, M. (Ed.) *The psychophysics of speech perception*, (pp. 39:210–219). Martinus-Nijhoff: Dordrecht.
- Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Review of Psychology*, 55, 149–179.
- Escera, C., Yago, E., Corral, M. J., Corbera, S., & Nuñez, M. I. (2003). Attention capture by auditory significant stimuli: semantic analysis follows attention switching. *European Journal of Neuroscience*, 18(8), 2408–2412.
- Eulitz, C., & Lahiri, A. (2004). Neurobiological evidence for abstract phonological representations in the mental lexicon during speech recognition. *Journal of Cognitive Neuroscience*, 16(4), 577–583.
- Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences*, 6(2), 78–84.
- Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London, Series B*, 360(1456), 815–836.
- Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, 6(1), 110.
- Gaskell, M. G., & Marslen-Wilson, W. D. (2002). Representation and competition in the perception of spoken words. *Cognitive Psychology*, 45(2), 220–266.
- Giard, M. H., Perrin, F., Pernier, J., & Bouchet, P. (1990). Brain generators implicated in the processing of auditory stimulus deviance: a topographic event-related potential study. *Psychophysiology*, 27(6), 627–640.
- Hagoort, P., & Brown, C. M. (2000). ERP Effects of listening to speech: semantic ERP effects. *Neuropsychologia*, 38(11), 1518–1530.
- Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8, 393–402.
- Hickok, G. (2012). The cortical organization of speech processing: feedback control and predictive coding the context of a dual-stream model. *Journal of Communication Disorders*, 45(6), 393–402.
- Holt, L. L., Lotto, A. J., & Kluender, K. R. (1998). Incorporating principles of general learning in theories of language acquisition. *Chicago Linguistic Society*, 34, 253–268.
- Jääskeläinen, I. P., Ahveninen, J., Bonmassar, G., Dale, A. M., Ilmoniemi, R. J., Levänen, S., et al. (2004). Human posterior auditory cortex gates novel sounds to consciousness. *Proceedings of the National Academy of Sciences of the United States of America*, 101(17), 6809–6814.
- Kraus, N., Anderson, S., & White-Schwoch, T. (2017). The frequency-following response: a window into human communication. In *The frequency-following response*, (pp. 1–15): Springer.
- 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.
- Krishnan, A., Xu, Y., Gandour, J. T., & Cariani, P. A. (2004). Human frequency-following response: representation of pitch contours in Chinese tones. *Hearing Research*, 189(1), 1–12.
- Krishnan, A., & Gandour, J. T. (2009). The role of the auditory brainstem in processing linguistically-relevant pitch patterns. *Brain and Language*, 110(3), 135–148.
- Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (de) constructing the n400. *Nature Reviews Neuroscience*, 9(12), 920–933.

- Lau, J. C., Wong, P. C., & Chandrasekaran, B. (2017). Context-dependent plasticity in the subcortical encoding of linguistic pitch patterns. *Journal of Neurophysiology*, 117(2), 594–603.
- Li, X., & Chen, Y. (2015). Representation and processing of lexical tone and tonal variants: evidence from the mismatch negativity. *PLOS One*, 10(12), e0143097.
- Liu, F., Maggu, A. R., Lau, J. C., & Wong, P. C. (2014). Brainstem encoding of speech and musical stimuli in congenital amusia: evidence from Cantonese speakers. *Frontiers in Human Neuroscience*, 8, 1029.
- López-Caballero, F., Zarnowiec, K., & Escera, C. (2016). Differential deviant probability effects on two hierarchical levels of the auditory novelty system. *Biological Psychology*, 120, 1–9.
- Lotto, A. J. (2000). Language acquisition as complex category formation. *Phonetica*, 57(2–4), 189–196.
- Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: the neighborhood activation model. *Ear and Hearing*, 19(1), 1.
- Lupyan, G., & Clark, A. (2015). Words and the world predictive coding and the language–perception–cognition interface. *Current Directions in Psychological Science*, 24(4), 279–284.
- Malmierca, M. S., Cristaudo, S., Pérez-González, D., & Covey, E. (2009). Stimulus-specific adaptation in the inferior colliculus of the anesthetized rat. *Journal of Neuroscience*, 29(17), 5483–5493.
- Malmierca, M. S., Anderson, L. A., & Antunes, F. M. (2015). The cortical modulation of stimulus-specific adaptation in the auditory midbrain and thalamus: a potential neuronal correlate for predictive coding. *Frontiers in Systems Neuroscience*, 9, 19.
- Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive psychology*, 10(1), 29–63.
- McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18(1), 1–86.
- Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5–6), 453–467.
- Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: a review. *Clinical Neurophysiology*, 118(12), 2544–2590.
- Natan, R. G., Briguglio, J. J., Mwilambwe-Tshilobo, L., Jones, S. I., Aizenberg, M., Goldberg, E. M., & Geffen, M. N. (2015). Complementary control of sensory adaptation by two types of cortical interneurons. *Elife*, 4, e09868.
- Nourski, K. V., Banks, M. I., Steinschneider, M., Rhone, A. E., Kawasaki, H., Mueller, R. N., & Howard III, M. A. (2017). Electroencephalographic delineation of human auditory cortical fields based on effects of propofol anesthesia. *Neuroimage*, 152, 78–93.
- Nusbaum, H., & Magnuson, J. (1997). Talker normalization: phonetic constancy as a cognitive process. In Mullenn, K. J. J. (Ed.) *Talker variability in speech processing*, (pp. 109–132). New York: Academic Press.
- Parbery-Clark, A., Strait, D., & Kraus, N. (2011). Context-dependent encoding in the auditory brainstem subserves enhanced speech-in-noise perception in musicians. *Neuropsychologia*, 49(12), 3338–3345.
- Pérez-González, D., Malmierca, M. S., & Covey, E. (2005). Novelty detector neurons in the mammalian auditory midbrain. *European Journal of Neuroscience*, 22(11), 2879–2885.
- Politzer-Ahles, S., & Zhang, J. (2012). The role of phonological alternation in speech production: evidence from Mandarin tone sandhi. In *Proceedings of meetings on acoustics 164ASA* (vol. 18, No. 1, p. 060001). ASA.
- Politzer-Ahles, S., Schluter, K., Wu, K., & Almeida, D. (2016). Asymmetries in the perception of Mandarin tones: evidence from mismatch negativity. *Journal of Experimental Psychology: Human Perception and Performance*, 42(10), 1547.
- Rauschecker, J. P., & Scott, S. K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature Neuroscience*, 12(6), 718–724.
- Rubin, J., Ulanovsky, N., Nelken, I., & Tishby, N. (2016). The representation of prediction error in auditory cortex. *PLoS Computational Biology*, 12(8), e1005058.
- Russo, N., Nicol, T., Musacchia, G., & Kraus, N. (2004). Brainstem responses to speech syllables. *Clinical Neurophysiology*, 115(9), 2021–2030.
- Shen, X. S., & Lin, M. (1991). A perceptual study of Mandarin tones 2 and 3. *Language and Speech*, 34(2), 145–156.
- Skoe, E., & Kraus, N. (2010). Auditory brainstem response to complex sounds: a tutorial. *Ear and Hearing*, 31(3), 302.
- Skoe, E., Chandrasekaran, B., Spitzer, E. R., Wong, P. C., & Kraus, N. (2014). Human brainstem plasticity: the interaction of stimulus probability and auditory learning. *Neurobiology of Learning and Memory*, 109, 82–93.
- Slabu, L., Grimm, S., & Escera, C. (2012). Novelty detection in the human auditory brainstem. *Journal of Neuroscience*, 32(4), 1447–1452.
- Slowiaczek, L. M., Nusbaum, H. C., & Pisoni, D. B. (1987). Phonological priming in auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(1), 64.
- Song, J. H., Skoe, E., Wong, P. C., & Kraus, N. (2008). Plasticity in the adult human auditory brainstem following short-term linguistic training. *Journal of Cognitive Neuroscience*, 20(10), 1892–1902.
- Strait, D. L., Hornickel, J., & Kraus, N. (2011). Subcortical processing of speech regularities underlies reading and music aptitude in children. *Behavioral and Brain Functions*, 7(1), 1.
- Strait, D. L., Parbery-Clark, A., Hittner, E., & Kraus, N. (2012). Musical training during early childhood enhances the neural encoding of speech in noise. *Brain and Language*, 123(3), 191–201.
- Suga, N. (2008). Role of corticofugal feedback in hearing. *Journal of Comparative Physiology A*, 194(2), 169–183.
- Whalen, D. H., & Xu, Y. (1992). Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica*, 49(1), 25–47.
- Winer, J. A., Larue, D. T., Diehl, J. J., & Hefti, B. J. (1998). Auditory cortical projections to the cat inferior colliculus. *Journal of Comparative Neurology*, 400(2), 147–174.
- Wong, P. C., & Diehl, R. L. (2003). Perceptual normalization for inter- and intratalker variation in Cantonese level tones. *Journal of Speech, Language, and Hearing Research*, 46(2), 413–421.
- Wong, P. C., 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.
- Xie, Z., Reetzke, R., & Chandrasekaran, B. (2017). Stability and plasticity in neural encoding of linguistically relevant pitch patterns. *Journal of Neurophysiology*, 117(3), 1409–1424.
- Xie, Z., Reetzke, R., & Chandrasekaran, B. (2018). Taking attention away from the auditory modality: context-dependent effects on early sensory representation of speech. *Neuroscience*, 384, 64–75.
- Yip, M. (2002). *Tone*. Cambridge: Cambridge University Press.
- Zhang, C., Xia, Q., & Peng, G. (2015). Mandarin third tone sandhi requires more effortful phonological encoding in speech production: evidence from an ERP study. *Journal of Neurolinguistics*, 33, 149–162.
- Zhou, X., & Marslen-Wilson, W. (1997). The abstractness of phonological representation in the Chinese mental lexicon. In *Cognitive Processing of Chinese and other Asian languages* (pp. 3–26).

Affiliations

Joseph C. Y. Lau^{1,2} · Patrick C. M. Wong^{1,2} · Bharath Chandrasekaran³

¹ Department of Linguistics and Modern Languages, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China

² Brain and Mind Institute, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China

³ Department of Communication Science and Disorders, School of Health and Rehabilitation Sciences, University of Pittsburgh, Pittsburgh, PA, USA