Title: Hidden Markov Modeling of Frequency-Following Responses to Mandarin Lexical Tones

Authors: Fernando Llanos, Zilong Xie, Bharath Chandrasekaran

PII: S0165-0270(17)30291-1
DOI: http://dx.doi.org/doi:10.1016/j.jneumeth.2017.08.010
Reference: NSM 7818

To appear in: *Journal of Neuroscience Methods*

Received date: 16-5-2017
Revised date: 3-8-2017
Accepted date: 8-8-2017

Please cite this article as: Llanos Fernando, Xie Zilong, Chandrasekaran Bharath.Hidden Markov Modeling of Frequency-Following Responses to Mandarin Lexical Tones.*Journal of Neuroscience Methods* http://dx.doi.org/10.1016/j.jneumeth.2017.08.010

TITLE

Hidden Markov Modeling of Frequency-Following Responses to Mandarin Lexical Tones

AUTHORS

Fernando Llanos[1], Zilong Xie[1], Bharath Chandrasekaran[1,2,3]

INSTITUTIONS

[1]Department of Communication Sciences & Disorders, Moody College of Communication, The University of Texas at Austin

[2]Department of Psychology, College of Liberal Arts, The University of Texas at Austin

[3]Institute for Neuroscience, College of Liberal Arts, The University of Texas at Austin

TYPE OF ARTICLE

Research paper

CORRESPONDING AUTHOR

Bharath Chandrasekaran (bchandra@utexas.edu; 1-512-471-2035)

2504 Whitis Ave., Austin, TX 78712, USA

RUNNING TITLE

Hidden Markov modeling of FFRs

HIGHLIGHTS

- A new method of neural pitch encoding assessment is described
- This method uses a hidden Markov model to decode neural pitch patterns reflected by frequency following responses
- The accuracy of this model provides a measure that accounts for language experience-dependent plasticity in subcortical encoding of pitch
- This method enables neural pitch encoding assessment with reduced datasets

ABSTRACT

Background. The frequency-following response (FFR) is a scalp-recorded electrophysiological potential reflecting phase-locked activity from neural ensembles in the auditory system. The FFR is often used to assess the robustness of subcortical pitch processing. Due to low signal-to-noise ratio at the single-trial level, FFRs are typically averaged across thousands of stimulus repetitions. Prior work using this approach has shown that subcortical encoding of linguistically-relevant pitch patterns is modulated by long-term language experience.

New method. We examine the extent to which a machine learning approach using hidden Markov modeling (HMM) can be utilized to decode Mandarin tone-categories from scalp-record electrophysiolgical activity. We then assess the extent to which the HMM can capture biologically-relevant effects (language experience-driven plasticity). To this end, we recorded FFRs to four Mandarin tones from 14 adult native speakers of Chinese and 14 of native English. We trained a HMM to decode tone categories from the FFRs with varying size of averages.

Results and comparisons with existing methods. Tone categories were decoded with above-chance accuracies using HMM. The HMM derived metric (decoding accuracy) revealed a robust effect of language

experience, such that FFRs from native Chinese speakers yielded *greater* accuracies than native English speakers. Critically, the language experience-driven plasticity was captured with average sizes significantly smaller than those used in the extant literature.

Conclusions. Our results demonstrate the feasibility of HMM in assessing the robustness of neural pitch. Machine-learning approaches can complement extant analytical methods that capture auditory function and could reduce the number of trials needed to capture biological phenomena.

KEYWORDS

Frequency-following response, hidden Markov model, machine learning, pitch encoding, plasticity

HIDDEN MARKOV MODELING OF FREQUENCY-FOLLOWING RESPONSES TO MANDARIN LEXICAL TONES

1 INTRODUCTION

Pitch is critical to speech and music processing (Ladefoged & Maddieson, 1998; Patel, 2010). For example, speakers of tonal languages (e.g., Chinese) rely on phonologically-relevant pitch patterns (i.e., lexical tones) to convey different word meanings (Gandour, 1983; Ladefoged & Maddieson, 1998). The main cues used for perceptual categorization of such lexical tones are pitch height and pitch direction (Gandour, 1994; Francis & Ciocca, 2003). The neural encoding of pitch is often assessed with the frequency-following response (FFR). FFR is a scalp-recorded electrophysiological potential that reflects phase-locked activity from neural ensembles involved in the processing of low level sound characteristics (Bidelman, 2015; Chandrasekaran & Kraus, 2010; Krishnan, Gandour, & Bidelman, 2010; Coffey, Herholz, Chepesiuk, Baillet, & Zatorre, 2006; Smith, Marsh, & Brown, 1975; Sohmer, Pratt, & Kinarti, 1977). Although it is generally considered that the FFR is entirely generated by auditory subcortical structures (e.g., Krishnan et al., 2005), recent evidence suggests a contribution from auditory cortex (Coffey et al. 2016). An important property of the FFR is that it captures the spectro-temporal correlates of the pitch (e.g., the fundamental

frequency, F0) with high fidelity (Chandrasekaran & Kraus, 2010; Krishnan, Xu, Gandour, & Cariani, 2004) (see Fig-1)

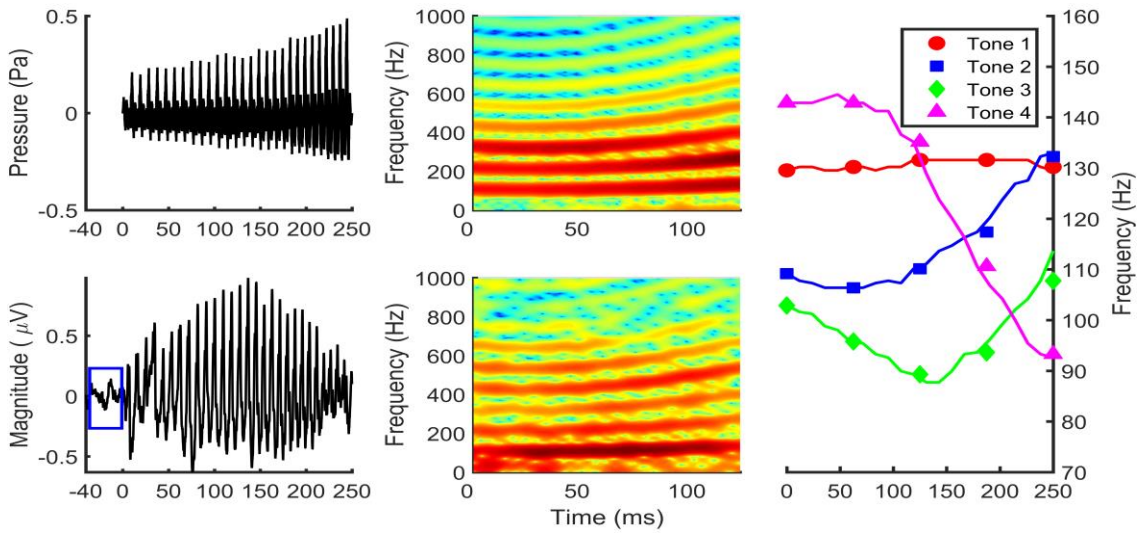

Figure 1. Waveform (A) and spectrogram (B) of a mid-rising Mandarin tone (tone 2), as well as the waveform (C) and spectrogram (D) of the corresponding FFR, elicited from a representative Chinese participant. The FFR was averaged across 1000 trials. The pre-stimulus portion of the FFR in panel C is delimited with a rectangle. Panel E shows the F0 contours of the four stimuli included in the present study.

Prior work has shown that the FFR is malleable to auditory experiences across several timescales (Skoe & Chandrasekaran, 2014). These experiences include long-term language experience (Krishnan, Xu, Gandour, & Cariani, 2005; Krizman, Marian, Shook, Skoe, & Kraus, 2012; Krizman, Skoe, Marian, & Kraus, 2014; Xie, Reetzke, & Chandrasekaran, 2017), musical training (Musacchia, Sams, Skoe, & Kraus, 2007; Wong, Skoe, Russo, Dees, & Kraus, 2007), short-term auditory training (Song, Skoe, Wong, & Kraus, 2008; Xie et al., 2017), online context (Lau, Wong & Chandrasekaran, 2016), and stimulus probability and FFR timing (Gorina-Careta, Zarnowiec, Costa-Faidella & Escera, 2016; Shiga, Althen, Cornella, Zarnowiec, Yabe & Escera, 2015; Slabu, Grimm & Escera, 2012). Following this line of research, recent studies have documented a more faithful subcortical encoding of lexical tones (e.g., Mandarin tones) in native speakers of tonal languages (e.g., Chinese), relative to native speakers of languages that are not tonal (e.g., English) (Krishnan et al., 2010; Krishnan, Gandour, Bidelman, & Swaminathan, 2009; Krishnan et al., 2005; Xie et

al., 2017). These results are theoretically interesting because they demonstrate that subcortical structures are not just passive relay stations, but relate to individual differences in auditory processing (Chandrasekaran, Skoe, & Kraus, 2014).

In the extant literature, the subcortical encoding of pitch is usually assessed with metrics that evaluate the similarity of F0 between the stimulus and the evoked FFR, or the degree of the FFR periodicity (Krishnan et al., 2009; Krishnan et al., 2005; Lau, Wong, & Chandrasekaran, 2016; Wong et al., 2007). The stimulus-to-response similarity is quantified with Euclidean norms that calculate the distance between the stimulus and the evoked FFR's F0 contour (F0 error metric), or normalized crosscorrelation coefficients that maximize at latencies for which the F0 contours being compared become highly correlate (stimulus-to-response correlation metric). To estimate the degree of periodicity, a FFR is compared to itself using the autocorrelation function (peak autocorrelation metric). Due to the low SNR of single-trial responses, FFRs are usually averaged across thousands of stimulus repetitions to faciliate the estimation of these metrics (Krishnan et al., 2010; Krishnan et al., 2005; Xu, Krishnan, & Gandour, 2006). In the present study, we examine the extent to which a machine learning approach using the hidden Markov model (HMM) can be utilized to decode Mandarin tones from FFRs, and capture language experience-dependent plasticity in the neural encoding of pitch. Critically, we examine the extent to which HMM can capture subtle language experience-dependent neural plasticity using subaverages (of the FFR) and datasets that are smaller than those reported in the existing literature.

Our data-driven approach builds upon emerging applications of machine learning methods to model electrophysiological responses (Hausfeld, De Martino, Bonte, & Formisano, 2012; Mesgarani, Cheung, Johnson, & Chang, 2014; Pei, Barbour, Leuthardt, & Schalk, 2011; Yi, Xie, Reetzke, Dimakis, & Chandrasekaran, 2017). The goal of unsupervised machine learning models (Alpaydin, 2014; Mohri, Rostamizadeh, & Talwalkar, 2012) is to learn target patterns (e.g., lexical tones) without being explicitly programmed. Instead, machine learning models are trained to develop broad mathematical

representations that capture the variability of target patterns in the input. These representations are then used to decode target patterns from novel input signals. An important characteristic of machine learning approaches is that they incorporate the modeling of complex distributional input properties (Mohri et al., 2012).

Previous studies have taken advantage of machine learning approaches to decoding cortical electrophysiological responses even on a single-trial basis (Mesgarani et al., 2014). Due to the brainstem sites of origin, FFRs are smaller in magnitude than cortical electrophysiological responses at the scalp, and hence have relatively lower SNR (Chandrasekaran & Kraus, 2010). Although this may limit the power of machine learning models to decode FFRs with small averages, a previous study has demonstrated the feasibility of machine learning methods to decode FFRs to steady-state vowels on a single-trial basis (Yi et al., 2017). Building upon this finding, we aim to investigate the extent to which machine learning methods can also decode FFRs to lexical tones. To the best of our knowledge, this is the first study using the HMM to decode FFR signal properties.

Among extant machine learning models, we chose the HMM because of the characteristic time-varying profile of lexical tones. HMMs are designed to decode series of emissions (e.g., F0 contours) by assuming that the emissions in these series are produced, with certain probabilities, by a series of states that cannot be directly observed (e.g., target lexical tone categories). In the HMM, hidden states are assumed to follow each other with certain transition probabilities established from the distribution of the emissions across sequences (Rabiner, 1989; Rabiner & Juang, 1986). For example, a HMM trained with rising pitch contours will generalize one stochastic representation in which lower F0s are followed by higher F0s with a higher probability than the opposite trend (higher F0s followed by lower F0s). In particular, there are two properties that make the HMM a good candidate to model time-varying pitch patterns. First, HMM can handle a wide variety of time-varying speech signals (Gales & Young, 2008; Huang, Ariki, & Jack, 1990), including Mandarin tones (Yang, Lee, Chang, & Wang, 1988). Second, the HMM

can decode patterns at moderate SNRs (Viikki & Laurila, 1997). Considering the low SNR characteristic of single-trial FFRs, the latter property can be useful to decode lexical tone categories in small FFR datasets.

We recorded 1000 trials of FFRs to each of the four Mandarin lexical tones (T1: high-level; T2: mid-rising; T3: falling-rising; and T4: high-falling) elicited from 14 native speakers of Mandarin Chinese and 14 native speakers of English with no prior tonal language experience. Then, we assessed the neural encoding of Mandarin tones by using a HMM to decode these tones from FFRs. Critically, we manipulated the number of trials used to train and test the HMM, as well as the number of trials used to average the FFRs. The rationale behind these manipulations was to assess the validity of the HMM in decoding FFRs averaged across fewer trials in smaller datasets. We also examined the performance of the HMM at different time points of the FFR. The goal of this additional analysis was to evaluate the effects of language experience-dependent plasticity on the decoding of Mandarin tones over time as well as elucidate the relative contributions of different portions of the FFR to tone classification. Here, the HMM was trained with complete FFRs and tested with FFR portions gradually increasing in 20 steps of 12.5 ms from the onset (12.5 ms corresponded to the 5% of the FFR duration).

To anticipate, the HMM decoded tone categories with above-chance accuracies in both Chinese and English participants. The HMM derived metric (decoding accuracy) exhibited the same pattern of neural plasticity documented in the literature: the Chinese group exhibited a more reliable neural encoding of native lexical pitch patterns than the English group. Critically, these language group differences were elucidated with data sets and averaging sizes that were considerably smaller than the ones documented in the existing literature. Further, FFRs in Chinese listeners yielded faster decoding of Mandarin tone categories relative to English listeners.

2 METHODS

2.1 Participants

Participants were recruited, tested, and compensated in Austin, Texas, by a research protocol approved by the Institutional Board Review of the University of Texas at Austin. We collected data from fourteen native speakers of Chinese (4 male; age: M = 24 years, SD = 3.4 years) and fourteen native speakers of English with no prior tonal language experience (9 male; age: M = 21.9 years, SD = 2.7 years). FFRs are not modulated by gender (Krizman, Skoe, & Kraus, 2012). All participants were right-handed by self-report. They did not report any history of musical training in the past 9 years. English participants were not fluent in a language other than English, and did not report any training in tonal languages. Chinese speakers were late unbalanced bilinguals of English (onset of learning: M = 11 years, SD = 2.8 years; self-rated English proficiency, from 1 to 10 points: M = 6.3 points; SD = 0.9 points). All participants reported no history of hearing problems, neurodevelopmental disorders, or traumatic brain injuries. Audiograms revealed pure-tone thresholds (from 250 to 8000 Hz, octave steps) lower than 25 dB for both air and bone conduction in each ear.

2.2 Stimuli

Stimuli consisted of four 250 ms synthetic Chinese words minimally distinguised by their lexical tones: /yi^{1}/ 'clothing' (T1), high-level tone with F0 equal to 129 Hz; /yi^{2}/ 'aunt' (T2), mid-rising tone with F0 rising from to 109 to 133 Hz (T2); /yi^{3}/ 'chair', low-dipping tone with F0 onset falling from 103 to 89 Hz and F0 offset rising from 89 to 111 Hz (T3); and /yi^{4}/ 'easy' (T4), falling tone with F0 falling from 140 to 92 Hz (T4). F0 contours of these four tones are shown in Fig-1 (panel E). The synthesis was derived from natural male production data (Chandrasekaran, Krishnan, & Gandour, 2007; Xu, 1997). We adjusted all stimulus magnitudes to the same root-mean-square (RMS) intensity value, equivalent to 72 dB sound

pressure level (SPL). SPL levels were measured with a Brüel & Kjaer hand-held analyzer (type 2250-L) connected to a Brüel & Kjaer artificial ear (type 4152).

2.3 Electrophysiological data acquisition

Electrophysiological responses to the stimuli were recorded (sampling rate: 25 kHz) in a sound attenuated booth, using one Ag-AgCl scalp electrode placed on the top-middle part of the head (Cz), with the left and right mastoids as the ground and reference, respectively. The use of a mastoid reference is consistent with our previous work (Xie et al., 2017) and prior work from other labs (Bidelman & Krishnan, 2010; Krishnan et al. 2010). This setup also provided us with a good SNR of FFRs, as shown in Fig-8. All electrode impedances were lower than 5 kΩ. Participants were watching a silent movie of choice with subtitles while listening to the stimuli, binaurally delivered through two insert earphones (ER-3; Etymotic Research, Elk Grove Village, IL). They were instructed to focus their attention on their movies and ignore the auditory stimuli. Stimulus presentation was controlled by a customized E-Prime interface (Schneider et al., 2002), using an inter-stimulus interval (ISI) randomly jittered between 122 to 148 ms. Each of the four Mandarin tones was presented in separate blocks with alternating polarities to minimize the impact of cochlear microphonic artifact on the average FFR (Bidelman, Moreno, & Alain, 2013; Chandrasekaran & Kraus, 2010). The order of the blocks was counterbalanced across participants, with half of the participants listening to blocks of T1 and T2 first, and the other half of participants listening to blocks of T3 and T4 first. The whole recording lasted about 60 minutes.

2.4 Preprocessing of electrophysiological data

Raw electrophysiological responses were preprocessed off-line with BrainVision Analyzer 2.0 (Brain Products, Gilching, Germany). First, responses were bandpass-filtered from 80 to 1,000 Hz (12 dB/octave, zero phase-shift). Then, they were segmented into epochs of 310 ms (40 ms before stimulus onset and 20 ms after stimulus offset). Responses were baseline corrected to the mean voltage of the

noise floor (-40 to 0 ms), and trials with amplitudes exceeding the range of ± 35 μV were rejected. For each tone, we collected 1000 artifact-free trials (500 for each polarity) for further analysis.

3 HIDDEN MARKOV MODELING OF NEURAL PITCH PATTERNS

The HMM classifier consisted of four independent HMMs connected in parallel and implemented in MATLAB R2015b (please see supplementary methods for the implemented MATLAB scripts). These HMMs were trained to decode FFR F0 contours to the four Mandarin tones in individual participants. Tone categories were selected as the HMMs providing the more optimal decoding of target FFRs. Therefore, the decoding of a particular tone category was also influenced by the decoding of the other categories. The F0 contours were estimated from averaged FFRs using the method described in section 3.1 and quantized into sequences of discrete F0 intervals, or codewords, previously established by clustering. The accuracy of the HMM classifier was cross-validated across multiple training and dataset testing selections. Critically, we manipulated the number of FFRs used to train and test the classifier, as well as the number of trials used to average FFRs (averaging size). These steps are schematically illustrated in Fig-2 and described in more detail throughout the next few sections. For more general information about HMM, we refer the readers to the specialized literature (Rabiner, 1989; Rabiner & Juang, 1993)

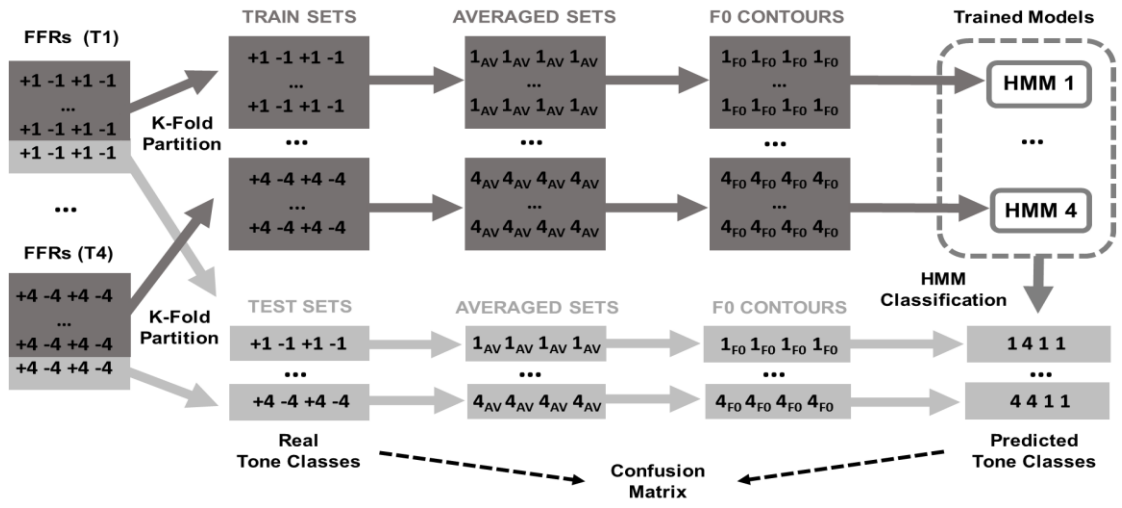

Figure 2. A flowchart illustrating the main steps followed at each cross-validation step. FFRs are labeled with the number of the corresponding tone (e.g., 1 or 4). Stimulus polarities are indexed with a "plus" and "minus" symbol attached to each tone number. FFRs to the same tone (e.g., T1) were separated in training and testing subsets using a K-Fold partition, where K is the possible number of disjoint testing selections available in a given tone dataset (e.g., T1). For example, if the testing size is 100 and the size of the tone data set is 500, the possible number K of disjoint testing selections is 5. The flowchart illustrates the steps followed at each testing selection. Training and testing FFRs were averaged within their corresponding training and testing subsets. F0 contours were extracted from averaged FFRs and quantized into sequences of discrete F0 intervals. Each HMM (one per tone category) was trained with the F0 sequences of the corresponding tone class. Each testing sequence was classified into the tone class of the HMM giving the highest likelihood log-probability of providing the sequence. HMM accuracy was estimated from the confusion matrix that resulted from this classification method across all testing selections.

3.1 F0 extraction procedure

Averaged FFRs were sliced into 22 consecutive time frames of 40 ms each separated by sliding steps of 10 ms. This time frame duration was chosen to improve F0 tracking by including a minimum of 3 cycles per frame. The F0 in each frame was estimated by the short-term autocorrelation method described in Boersma (1993). We searched for the time-lag that maximized the normalized autocorrelation function within a time interval spanning the time-variant periods of 4 to 14.3 ms. We chose this time interval because its reciprocals, of 250 and 70 Hz respectively, encompassed the F0 range of all the stimuli. The reciprocal of the time-lag that maximized the autocorrelation function was considered as the F0 value of the corresponding frame. Time frames were mean de-trended and fully ramped with a standard Hann window, as described in Boersma (1993). Autocorrelation coefficients were divided by their corresponding coefficients in the Hann's autocorrelation function. We applied the same method to extract the F0 contour of each stimulus, previously resampled to the sampling rate of the FFR (i.e., 25 kHz).

3.2 Stochastic topology and vector quantization

Each of the four HMMs in the HMM classifier was defined as one stochastic chain of three self-connected hidden states, feedforward connected to the next two states (i.e., the first hidden state was

connected to itself as well as to the second and third states, whereas the second state was connected to itself and the third state). The F0 contours were quantized into sequences of 22 discrete F0 intervals, or codewords, using a codebook of 50 centroids. To create this codebook, we clustered all the F0 data points available for training in 50 Voronoi cells with a Linde-Buzo-Gray algorithm (Chang & Hu, 1998; Equitz, 1989). The codebook consisted of all the centroids provided by these cells. This choice of parameters (HMM architecture and codebook length) was derived from a previous study that used the HMM to recognize Mandarin tone production (Yang et al., 1988).

3.3 Training, testing, and cross-validation

Each HMM in the classifier was trained with Viterbi (Durbin, 1998) to learn F0 sequences from the same tone class. Trained HMMs were then combined in a parallel architecture to classify novel F0 sequences from any of the four tone classes. Each novel F0 sequence was decoded to a tone category of the HMM providing the highest likelihood log-probability. The accuracy (or correct classification rate) of the classifier was directly computed from the confusion matrix that resulted from this classification method, as the number of true positives and true negatives, divided over the number of true positives, true negatives, false positives, and false negatives (Fielding & Bell, 1997; Gardner & Altman, 1989). HMM accuracy was first estimated for each tone, resulting in a total of four tone-specific accuracy scores (Acc1, Acc2, Acc3, and Acc4) ranging from 0 (minimal accuracy) to 1 (maximal accuracy). We also calculated a general accuracy score across all tones (Acc0), which was computed from the total number of true and false positives and negatives provided by the classifier for all tone classes.

FFRs in the same tone dataset (T1, T2, T3, or T4) were divided into disjoint training and testing subsets. To avoid stimulus artifact bias, training and testing subsets alternated the same number of trials with opposite stimulus polarity (Bidelman et al., 2013; Skoe & Kraus, 2010). To avoid sample selection bias, the selection of testing and training subsets was cross-validated using a K-fold partition, with K equal

to the possible number of disjoint testing selections available in the corresponding tone dataset (Kuhn & Johnson, 2013; Siuly, Li, & Zhang, 2017). For example, if the testing size is 100 trials and the number of FFRs to the tone data set is 500, the possible number of disjoint testing selections is 5. The HMM classifier was trained and tested K times, using different cross-validated selections of testing and training data sets at a time. The accuracy scores (Acc1, Acc2, Acc3, Acc4, and Acc0) were computed from the confusion matrix that resulted from the classification of all F0 sequences at each cross-validation step.

3.4 Averaging method

FFRs in training and testing subsets were averaged with a moving average window (MAW) of length L (we relied on different MAW sizes, which are provided in the next section). This means that each FFR was averaged with respect to the L/2 previous and L/2 following FFRs within the same training or testing subset. This averaging method helped increase the SNR of each FFR while keeping a adequate number of trials for training and testing. However, it also increased the level of redundancy across trials, and thus the risk of under-fitting (Kuhn & Johnson, 2013). To reduce this risk, we performed a exploratory series of HMM runs across different training and MAW sizes. The purpose of this exploratory series was to identify the MAW sizes for which the accuracy of the classifier did not grant much improvement. By excluding these MAW sizes in future runs, we aimed minimize the risk of under-fitting without decreasing the accuracy of the classifier. In this exploratory series, training size was independently fixed to a small size (250 trials), medium size (500 trials) and large size (750 trials). Since the goal here was not to find the minimum number of trials required for the classifier to identify cross-language differences in decoding accuracy, FFR trials that were not used for training were left for testing. The size of the MAW was manipulated, for each training size, from 50 trials to the corresponding training size in increments of 25. Results (see Fig-3) revealed no significant accuracy improvements for MAW sizes larger than approximately the 50% of the training size (mean across training sizes = 55%). The effects of underfitting

were especially salient in the large sample size (see Fig-3, right panel), which covered a wider range of averaging sizes.

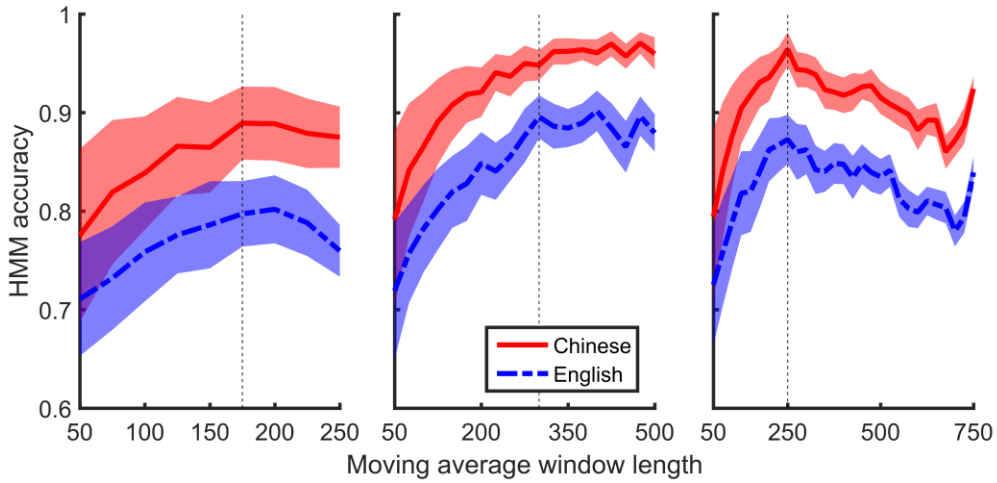

Figure 3. Mean HMM accuracy scores for Chinese and English groups (in solid red and dashed blue, respectively), as a function of the averaging size and for training sizes of 250 trials (left panel), 500 (middle panel), and 750 (right panel). Mean standard errors are color shaded. Dashed vertical lines highlight the onset of the moving average range that did not grant much improvement in HMM accuracy.

3.5 Manipulation of training, testing, and averaging size

We manipulated the training, testing and averaging size across different runs in individual participants. The sample size was denoted as the training size plus the testing size. The training size ranged from 100 to 900 trials per tone in steps of 50 across runs. The averaging size was manipulated as a function of the training size to minimize the risk of underfitting (see section 3.4). Specifically, for each training size, the averaging size varied from 50 trials to the half the corresponding training size in increments of 25. The testing size, in each run, was set to be the double of the corresponding averaging size to minimize the risk of under-fitting while keeping a reduced but a decent number of trials per run. Combinations of training and testing sizes that led to sampling sizes larger than 1000 were excluded. The resulting combinations of training, testing and averaging size are depicted in Fig-5 and Fig-6 (section 4.2). For each combination of

sizes, we obtained a total of five accuracy scores per participant: four tone-specific accuracy scores (Acc1, Ac2, Acc3, and Acc4), and one general accuracy score for all tones altogether (Acc0).

3.6 Tone decoding accuracy over time

In addition to the modeling described in the previous sections (3.2 – 3.5), we assessed the performance of the classifier in decoding Mandarin tones over time. Our goal here was to investigate the extent to which FFRs from Chinese listeners yielded earlier decoding of tone categories than English listeners. In such a case, we would expect the HMM to require less time (shorter FFR portions) to decode Mandarin tones from FFRs in Chinese listeners. To test this, we trained the HMM with complete F0 sequences. Then, we estimated the tone class of each novel F0 sequence along 20 consecutive time points gradually increasing in steps of 12.5 ms (12.5 ms corresponded to the 5% of the FFR duration). The tone class at each time point was estimated with the classification criteria described in section 3.2 (maximum likelihood log-probability). Portions of F0 sequences from the onset of the FFR to the corresponding time point were decoded to the tone category that provided the highest likelihood log-probability. This analysis differs from existing analyses of pitch-tracking accuracy across different FFR chunks (Krishnan et al., 2009) in that our goal here was to assess the amount of information required over time for the HMM to decode tone categories correctly, instead of identifying the portions of the FFR that were more reliably decoded. In this context, the use of complete FFRs during the training phase was meant to replicate the experimental circumstances in which FFRs were elicited (participants were listening to complete stimulus waveforms).

Our analysis of time-dependent accuracy focused on the combinations of training, testing and averaging size that maximized and minimized cross-language differences in general accuracy (Acc0). We focused on these combinations of sizes because they represented the best and worst circumstances in

which language group differences were expected to occur. The exact size values are provided in section 4.3.

4 STATISTICAL ANALYSES AND RESULTS

4.1 FFR dataset validation

Prior to the HMM modeling, we used our FFR data set to validate the pattern of language experience-dependent plasticity documented in the literature with the existing metrics: F0 error metric, the stimulus-to-response correlation, and peak autocorrelation (see Xie et al., 2007, for a more detailed description of these metrics). These metrics have been used to evaluate the effects of long-term auditory experience (Bidelman, Gandour, & Krishnan, 2011; Krishnan et al., 2005; Wong et al., 2007; Xu et al., 2006) and short-term auditory training (Chandrasekaran, Kraus, & Wong, 2012; Song et al., 2008; Xie et al., 2017) in the subcortical encoding of Mandarin tones in native speakers of tonal and non-tonal languages. As expected, the Chinese group revealed a more optimal neural encoding of pitch than the English group across the three metrics, as evidenced by significant effects of the language group [F0 error: $F(1,26) = 11.20$, $p = 0.002$, $\eta_p^2 = 0.18$; stimulus-to-response correlation: $F(1,26) = 6.61$, $p = 0.01$, $\eta_p^2 = 0.09$; peak autocorrelation: $F(1,26) = 7.44$, $p = 0.01$, $\eta_p^2 = 0.12$]. Specifically, FFRs from Chinese listeners, relative to those from English listeners, exhibit lower F0 errors and higher stimulus-to-response and peak autocorrelation coefficients [F0 error: Chinese, M = 4.92 (SD = 4.23) vs. English, M = 9.97 (SD = 7); stimulus-to-response correlation: Chinese, M = 0.56 (SD = 0.23) vs. English, M = 0.48 (SD = 0.21); peak autocorrelation: Chinese, M = 0.60 (SD = 0.13) vs. English, M = 0.43 (SD = 0.13)].

Fig-4 depicts the F0 contours of the four stimulus waveforms with the F0 contours of their elicited FFRs, averaged by language group. In this figure, stimulus pitch contours are more faithfully reproduced in Chinese than in English. We note that the results for the HMM metric, introduced in the next sections,

are not entirely consistent with the fidelity of the FFR. The HMM classifier is structured as four tone-specific HMMs connected in parallel. Therefore, its performance is expected to be modulated by F0 contour similarities across tone categories. This type of modulation is not expected to occur in previous FFR metrics, which focuses on stimulus and FFR signal properties within the same tone category but not across tone categories.

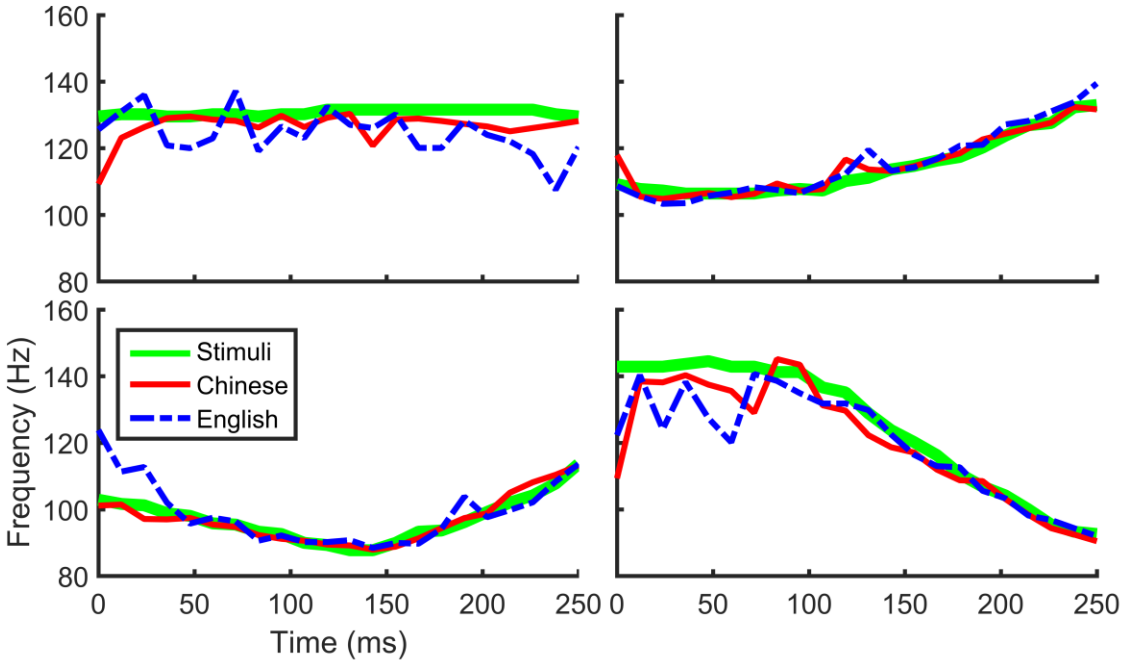

Figure 4. F0 contours of stimulus signals (thick green line) and mean F0 contours of averaged FFRs across Chinese speakers (thin red line) and English speakers (thin blue-dashed line) listening to T1 (top left), T2 (top right), T3 (bottom left), and T4 (bottom right).

4.2 Tone decoding accuracy

Mean accuracy scores for each combination of language group (Chinese and English) and tone (T1, T2, T3, and T4) were higher than the level of chance (0.25), and ranged from 0.74 to 0.88 across these combinations. Fig-5 depicts the accuracy scores (Acc0, Acc1, Acc2, Acc3, and Acc4), averaged across all

participants, for each combination of training, testing and averaging size. T4 provided the poorest mean accuracy scores. These means were lower than 0.75 when the averaging size was smaller than 100 trials. In contrast, the mean accuracies of the other tones were higher than 0.75 for averaging sizes as small as 50 trials, with a few exceptional cases, and were higher than 0.85 for averaging sizes approximately larger than 125 trials (see Fig-5).

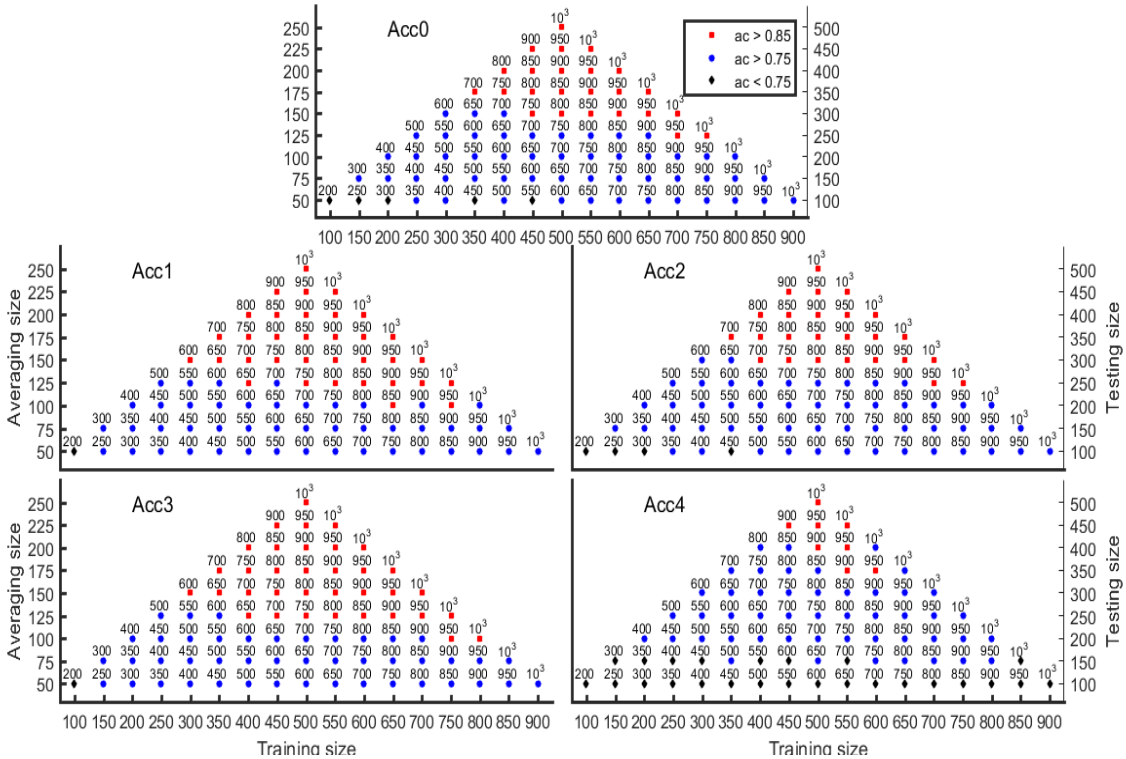

Figure 5. HMM accuracy scores (from 0 to 1) averaged across all participants, as a function of the training, testing, and averaging size. General accuracy scores (Acc0) for all tones are shown in the top panel. The panels below are labeled with the names of their corresponding tone-specific accuracy scores (Acc1, Acc2, Acc3, and Acc4). Accuracy scores are encoded with a color-shape scale that groups them in different intervals from lower to higher accuracy. Numbers on top of each colored squared correspond to the sampling size (training plus testing size).

Mean accuracy scores were consistently higher for the Chinese group, relative to the English group [Acc1: Chinese, M = 0.88 (SD = 0.1) vs. English, M = 0.79 (SD = 0.1); Acc2: Chinese, M = 0.86 (SD = 0.1) vs. English, M = 0.78 (SD = 0.1); Acc3: Chinese, M = 0.88 (SD = 0.09) vs. English, M = 0.78 (SD = 0.1); Acc4: Chinese, M = 0.81 (SD = 0.12) vs. English, M = 0.74 (SD = 0.11)]. When all the accuracy scores were

collapsed by tone, Acc4 (M = 0.77, SD = 0.12) was smaller than Acc1 (M = 0.84, SD = 0.11), Acc2 (M = 0.82, SD = 0.11), and Acc3 scores (M = 0.83, SD = 0.11). Pairwise comparisons via a two-sample t-test (Mann-Whitney test for non-normal and/or heteroscedastic score distributions) with T4 as the baseline tone confirmed that the Acc4 was significantly lower than the other three accuracy scores (Acc1, Acc2, and Acc3; all ps < 0.001). This indicates that T4 was more poorly decoded than the other three tones.

4.3 Language group differences

Language group differences for each accuracy score (Acc1, Acc2, Acc3, Acc4, and Acc0) at each combination of training, testing and averaging size were evaluated using two-sample t-tests (Mann-Whitney tests for non-normal and/or heteroscedastic score distributions). In each test, accuracy score and language group were the dependent and independent variables, respectively.

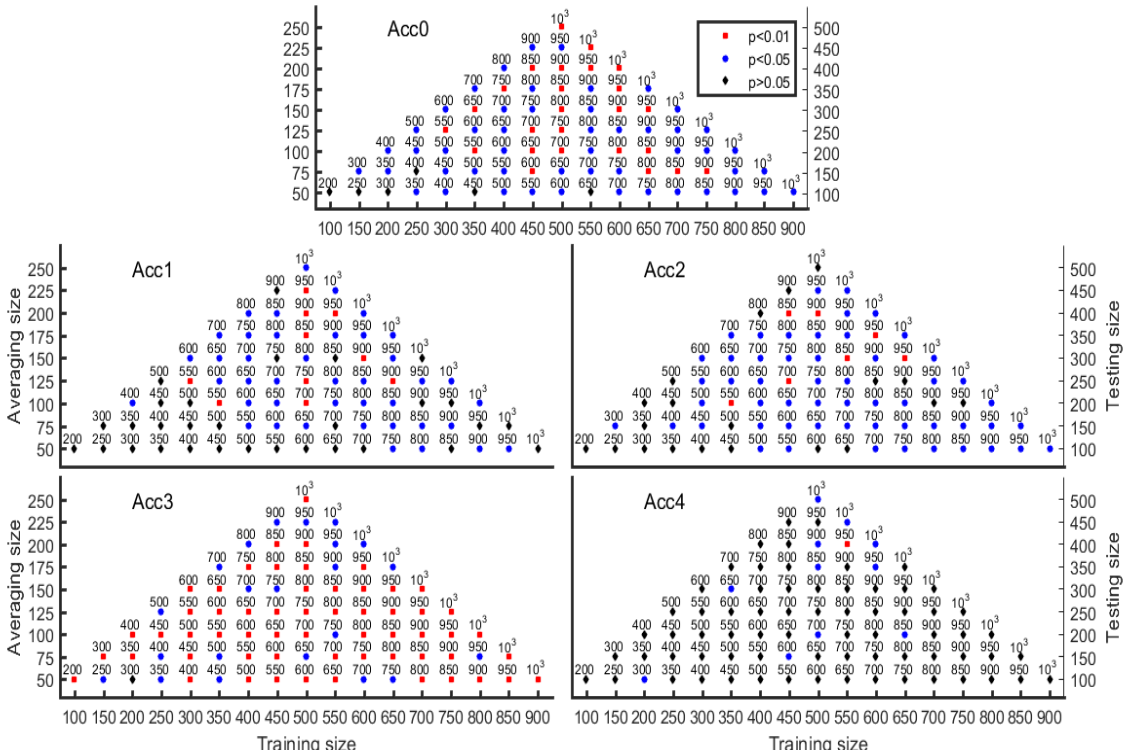

Figure 6. T-test results for language group differences (Chinese vs. English) in decoding accuracy as a function of the training, testing and averaging size. General accuracy scores for all tones (Acc0) are shown in the top panel. The panels below are labeled with the name of the corresponding tone-specific accuracy

score (Acc1, Acc2, Acc3, and Acc4). Statistical significance is encoded with a color-shape scale. Numbers on top of each colored squared correspond to the sampling size (training plus testing size).

Test results are displayed in Fig-6. This figure uses a color scale to display the statistical significance of each test. In all tests yielding $p < 0.05$, the mean accuracy score of the Chinese group was higher than the one of the English group. Test results for general accuracy (Acc0; Fig-6, top panel) provided a big area of statistical significance across almost all combinations of training, testing and averaging size, with only six combinations that were not statistically significant. Tests results for Acc3 (Fig-6, bottom-left panel) revealed statistically significant language group differences across all size combinations except one (training size = 200, averaging size = 50, testing size = 400). In contrast, the results for Acc4 (Fig-6, bottom-right panel) provided the smallest number of tests to reach statistical significance (12 out of 81). In this case, language group differences became statistically significant with sampling sizes larger than approximately 700 trials and averaging sizes larger than approximately 350 trials. In regards to Acc1 and Acc2 (Fig-6, mid-left and -right panels), language group differences reached statistical significance at sampling and averaging sizes larger than approximately 550 and 150 trials, respectively, with a total number of 51 (Acc1) and 61 (Acc2) tests showing statistically significant language group differences. A series of four binary Chi-square tests confirmed that Acc4 provided the smallest proportion of statistically significant language group differences ($qs > 17.92$, $ps < 0.001$).

4.4 Tone decoding accuracy over time

Language group differences in decoding accuracy over time (FFR portions) were statistically evaluated via two-sample t-tests (Mann-Whitney tests for non-normal and heteroscedastic distributions). Language group differences were evaluated for each decoding accuracy score (Acc0, Acc1, Acc2, Acc3 and Acc4) at each time point (20 time points total) in each condition (optimal and suboptimal). In the optimal condition, we focused on the combination of training, averaging and testing size that maximized language

group differences in general accuracy (training size = 500 trials, averaging size = 200 trials, sampling size = 900 trials). In the suboptimal condition, we focused on the combination of sizes that did not yield differences between language groups (training size = 150 trials, averaging size = 50 trials, sampling size = 250). These two conditions represented the best and worst circumstances in classifier performance (with the ability to capture language differences). Test results are shown in Fig-7 (panels B - E).

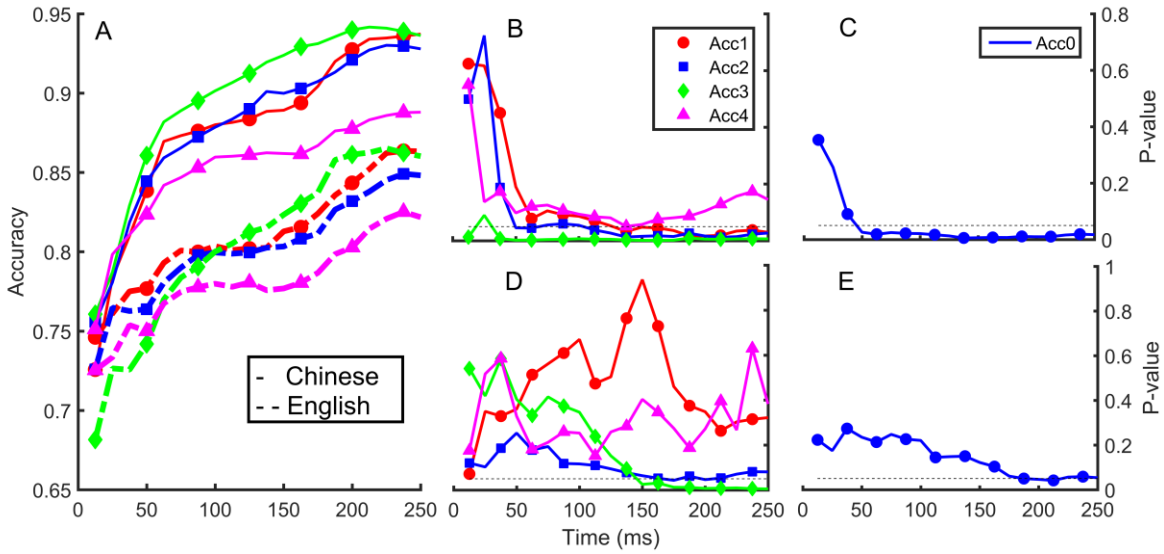

Figure 7. Panel A shows the evolution of tone-specific accuracy scores (Acc1, Acc2, Acc3, and Acc4), over time, in Chinese (solid lines) and English (dashed lines) in the optimal condition (training size = 500, averaging size = 10, sampling size = 900). The other panels (B-E) illustrate the statistical significance (p-values) of language group differences in accuracy scores over time. P-values in B and D refer to language group differences in tone-specific accuracies over time and p-values in C and E refer to the same differences in general accuracy over time. P-values in B and C refer to the optimal condition and p-values in D and E refer to the suboptimal condition (training size = 150, averaging size = 50, sampling size = 250). The dotted lines in B-E mark the level of statistical significance (0.05).

In the optimal condition, the Chinese tone-specific accuracy scores were higher than the English ones across all time points (Fig-7, panel A). Within the Chinese group, T3 was decoded earlier than T1 and T2. T4 was poorly decoded across time, relative to the other tones. In English, however, T1 provided the best decoding accuracies during the first 100 ms. After 100 ms, the evolution of tone-specific accuracy scores over time was similar across language groups: T3 provided the best over-time accuracy scores,

followed by T2 and T1 at similar over-time accuracy scores, which, in turn, were better than the ones of T4. Further, language group differences in accuracy reached statistical significance (i.e. $p < 0.05$) during the first 200 ms (Fig-7, panel B). The first tone to reach $p<0.05$ was T3, at 37.5 ms, followed by T2 (50 ms), T1 (137.5 ms) and T4 (137.5 ms), respectively. From these results, we infer that, in the optimal condition, Chinese FFRs yielded faster decoding of tone categories than English FFRs. Also, T3 was decoded faster than the other three tones, except in English FFRs, which provided a faster decoding of T1, at least during their first 100 ms. Finally, T4 was decoded slower than the other three tones in both language groups.

In the suboptimal condition (Fig-7, panel D), only T3 (150 ms) and T2 (187.5 ms) reached statistical significance in language group comparisons of accuracy over time ($p<0.05$). The temporal evolution of language group differences in general accuracy (Acc0) is shown in panels C (optimal condition) and E (suboptimal condition) of Fig-7. In this case, language group differences reached statistical significance as early as 50 ms (optimal condition) and 137.5 ms (suboptimal condition). Altogether, these results indicate that the decoding of Mandarin tones from FFRs was modulated, over time, by both language-dependent factors and tone categories.

5 GENERAL DISCUSSION

5.1 Results and major contributions

We examined the extent to which hidden Markov modeling (HMM) can be utilized to decode Mandarin tones from FFRs. A second goal was to assess the extent to which this metric (decoding accuracy, tone decoding accuracy over time-decoding accuracy) was able to capture subtle language experience-dependent plasticity in the neural encoding of Mandarin tones as documented in the literature (Krishnan et al., 2005; Krizman, Marian, et al., 2012; Krizman et al., 2014; Xie et al., 2017).

Specifically, we investigated the extent to which the Chinese FFRs provided a better and faster decoding of Mandarin tones than English ones.

The HMM decoded tone category information from FFR data with high, above-chance accuracy scores. Importantly, these scores were achieved with averaging sizes as small as 50 or 100 trials, depending on the tone (see Fig-5). Mandarin tones were more accurately decoded from Chinese FFRs, suggesting a more reliable neural encoding of pitch in the Chinese group, relative to the English group (see Fig-6). This finding is consisent with the language experience-dependent plasticity documented in the literature (Krishnan et al., 2005; Xie et al., 2017; Xu et al., 2006).

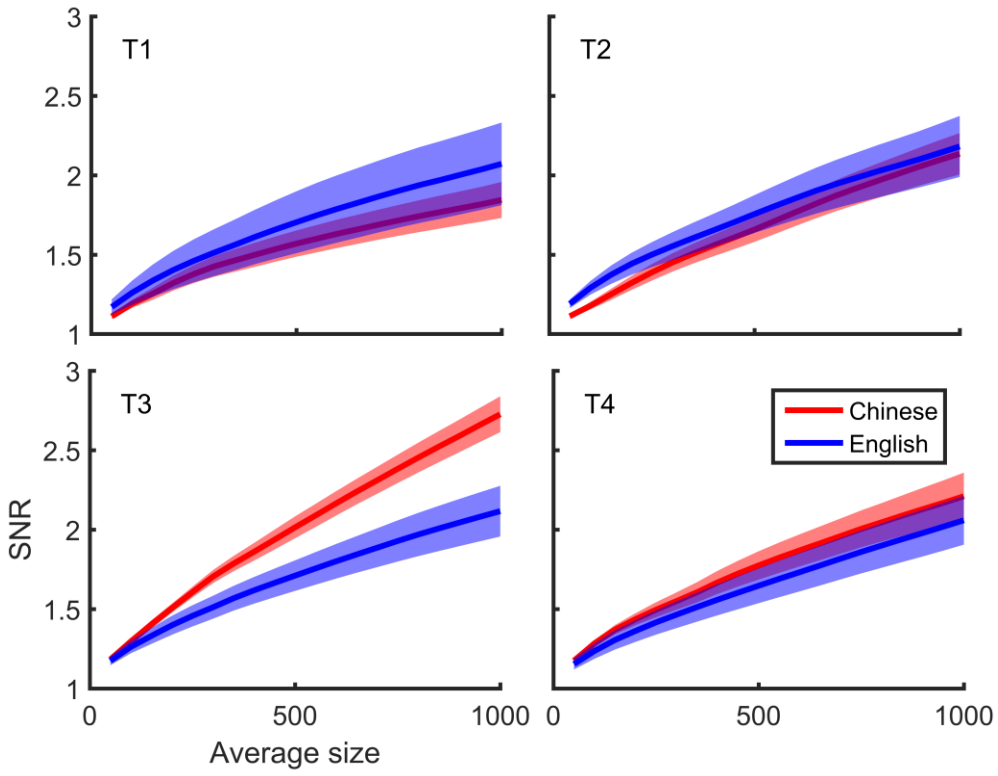

Figure 8. Mean signal-to-noise ratios (SNRs) as a function of the averaging size, language group (Chinese, and English) and tone (T1, T2, T3, T4). SNRs (y-axes) were estimated with the procedure described in section 3.4. For every average size L (x-axes), FFR and pre-stimulus signals were averaged with respect to the L/2 preceding and L/2 following signals. The mean amplitude of the averaged FFR signals were then divided over the mean amplitude of their corresponding averaged pre-stimulus signals. The ratio was take

as a value of SNR. The resulting SNRs were then averaged into a single SNR score per participant. Language group variances are shaded in the corresponding color.

Critically, in the present study, this pattern of results was elucidated with FFR data sets as small as 200, 550 or 700 trials, and was found to be tone-dependent (T3 and T4 required the smallest and largest sample sizes, respectively). Also, Chinese FFRs provided a faster decoding of tone categories than English FFRs, as early as 50 ms in the optimal condition, relative to 137.5 ms in the suboptimal one (see Fig-7, panels C and E). Taken together, these results demonstrate that the HMM is robust in capturing biologically-relevant information in the FFR, even at low SNR evidenced in smaller sample sizes and averaging sizes (these SNRs are shown in Fig-8).

5.2 Language-dependent neural plasticity

The language experience-dependent plasticity in the neural encoding of pitch patterns (e.g., Mandarin tones) has been attributed to differential amounts of exposure to dynamic pitch patterns and relative differences in the linguistic-relevance of dynamic pitch patterns (Jeng, Dickman, Montgomery-Reagan, Tong, Wu & Lin, 2011; Krishnan & Gandour, 2009; Krishnan et al. 2012). Mandarin Chinese listeners, relative to English listeners, receive extensive exposure to dynamic, linguistically-relevant pitch patterns during their language development. As a result, the subcortical auditory circuitry in Chinese listeners is likely to be reorganized to facilitate the encoding of dynamic pitch patterns that are frequently occurring in their auditory environment. This account is supported by animal models suggesting that neural circuitry calibrates locally to selectively fine-tune representation of frequently occurring signals in one's auditory environment during development (Keuroghlian & Knudsen, 2007; Knudsen, 2002) Such local reorganization can persist through adulthood, and continue to modulate auditory processing (Keuroghlian & Knudsen, 2007; Linkenhoker, von der Ohe, & Knudsen, 2005)

Our results indicate that the neural encoding of pitch is not only affected by language experience but also by tone category. In particular, the decoding of T4 was considerably less accurate than any of the other three tones. This result is in line with previous findings showing a suboptimal subcortical encoding of falling tonal patterns (Krishnan et al., 2010; Krishnan et al., 2009; Krishnan et al., 2004; Liu, Maggu, Lau, & Wong, 2015). For example, Krishnan et al. (2010) used normalized autocorrelation to estimate the strength of periodicity in the FFR (peak autocorrelation metric). They found a suboptimal neural encoding of periodicity in FFRs to falling linguistically-relevant tonal contours, relative to rising contours. This pitch direction asymmetry has also been documented in FFRs to non-speech signals (Krishnan & Parkinson, 2000; Krishnan et al., 2004) and in psychoacoustic experiments (Collins & Cullen Jr, 1978) in which rising tones differences were better discriminated than the ones between falling tones.

In the present study, the poorer decoding of the falling tone (T4) contrasts with the optimal accuracy scores of the Mandarin tones with rising F0 trajectories (T3 and T2). This result, is also reflected in the FFR literature (Krishnan, Gandour, & Bidelman, 2012; Krishnan et al., 2004). Speakers of tonal languages tend to exhibit a better neural encoding of pitch directional cues (critical to dynamic tones like T3 and T2) relative to speakers of languages that are not tonal. In our study, SNRs are also tone-specific (see Fig-8). Chinese SNRs were higher than the English ones for all tones except the level T1. When averaging size was maximum (1000 trials, optimal SNR), T2 and T3 maximized the differences between mean Chinese SNRs (T1: 2.00; T2: 2.47; T3: 2.94; T4: 2.28) and mean English SNRs (T1: 2.00; T2: 2.18; T3: 2.11; T4: 2.05). Language group differences in SNR reached statistical significance ($p<0.05$) with T3, the tone involving more complex pitch directional changes. These group differences systematically reflect the relative differences in the tuning to dynamic pitch.

5.3 Tone decoding accuracy over time

Our results suggest that tone decoding accuracy over time is also dependent on the tone category. Decoding of T4 was considerably more delayed than any of the other three tones in both Chinese and English speakers. This result could be a consequence of the poor decoding accuracy of T4 combined with the acoustic similarity of the onsets of T4 and T1 during approximately the first 70 ms. Language experience also mediated tone-specific differences in tone decoding accuracies over time. In Chinese, T3 and T2 were decoded faster than T1 and T4. Accuracies over time were different in English, especially during the first 100 ms, where T1 was decoded faster the other tones. Interestingly, also in English, the decoding accuracy of T3 increased considerably in magnitude, after approximately 150 ms, consistent with T3 turning point (Krishnan et al., 2010). In general, Chinese scores for general accuracy over time (Acc0, all tones combined) were significantly better than the English ones even with small sampling and averaging sizes (see Fig-7, panels C and E). This result demonstrates that the time-based decoding of Mandarin tones, as reflected by the HMM, is highly influenced by long-term auditory experience. Despite this, our time-based analysis has two important limitations. First, we used duration-normalized stimuli, so this may not reflect real-world differences. Second, we did not have behavioral data (tone gating paradigm) that could provide information on behavioral-relevance. These limitations can be addressed in future studies.

In previous tone gating studies, Chinese listeners required more time to identify T3, relative to the other three Mandarin tones (Whalen & Xu, 1992; Wu & Shu, 203), while T1 is typically identified the earliest (e.g., Whalen & Xu, 1992; however, T2 in Wu & Shu, 2003). Consistent with previous studies, our HMM analysis showed that T1 is decoded earlier than T4. In contrast to previous studies (Whalen & Xu, 1992), we found that T3 was decoded faster relative to the other three tones. This divergence could be due to differences in tone specific duration

across studies. Previous tone gating studies preserved the tone specific duration differences observed in Mandarin Chinese production, such that T3 is longer than T2, which in turn is usually longer than T4 and T1. The longer duration of T3 might have contributed to its longer identification latency, relative to the other Mandarin tones. In our study, however, all the tones had the same duration. This, combined with the lower F0 onset of T3, relative to the onset of the other tones, might have facilitated an earlier temporal decoding of T3. Another important difference worth noting here is that listeners' performance in previous tone gating tasks was modulated by attentional and decisional factors. Our data acquisition paradigm, based on passive listening, may be less affected by attentional and decisional factors. The difference in the attentional and decisional factors between our study and prior tone gating work may explain the disparity of findings. The influence of attentional and decisional factors on pitch representation could be further investigated with FFR elicitation paradigms that engage varying levels of attention.

5.4 Conclusion and future directions

Our results demonstrate the feasibility of machine learning approaches to characterize the FFR, which is considered a promising biomarker of auditory function (Chandrasekaran & Kraus, 2010; Chandrasekaran et al., 2012; Chandrasekaran., Skoe, & Kraus, 2014). Importantly, our study shows that the machine learning derived metrics can reflect biologically-relevant influences on auditory processing. One important finding worth highlighting is that this machine learning approach can be used to quantify FFR with trial numbers well below those reported in the existing literature. Studies using Mandarin tones often recorded FFRs to at least 1500 stimulus repetitions (Bidelman et al., 2013; Krishnan et al., 2010; Krishnan et al., 2005; Wong et al., 2007; Xu et al., 2006). When all tones are considered, the language

experience effect becomes statistically reliable ($p<0.01$) at or above 550 trials (training size = 350, testing size = 200, averaging size = 100). When only T3 is considered, this threshold moves to 300 trials (training size = 150, averaging size = 57, testing size = 150). Hence, with the machine learning approaches, we could potentially move away from typical experimental designs where we record FFRs to thousands of stimuli repetitions, and adopt designs that record FFRs to less stimuli repetitions. An immediate benefit may be to reduce the time, effort, and cost associated with FFR data collection. Shorter experimental durations could also be valuable in FFR studies involving difficult-to-test populations (e.g., infants, older adults and hearing-impaired listeners). These approaches also have the potential to provide information, akin to gating studies, on the minimum amount of information required for decoding.

Given than the FFR is very stable within participants across different experimental sessions, as shown by Xie et al (2017), the HMM could be trained with input across different experimental sessions. This method would improve the reliability of the model and save even more time in future experiment using the same stimulus signals.

Despite this promising methodological perspective, we note that the minimal number or trials required for the HMM to decode speech categories or group differences may be modulated by task related factors such as attention in active listening, stimulus properties such as the type of background noise, or the signal property under investigation (e.g., low vowel formants). For example, background noise may decrease the SNR of the FFR and thus increase the minimal number of trials required for the HMM to operate efficiently. Similarly, vowel formants located at higher frequency bands may also require of a larger number of trials in order to be decoded by the HMM, due to the poorer spectro-temporal resolution of the FFR at higher frequencies (Chandrasekaran & Krauss, 2010; Skoe & Krass, 2010). Further research is needed to clarify these aspects.

ACKNOWLEDGMENTS

This work was supported by the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health, Bethesda, Maryland [grant number R01DC013315 B.C.]. The content is solely the responsibility of the authors and does not represent the official view of the National Institutes of Health. The authors thank Rachel Reetzke, Jacie Richardson, and the research assistants for significant contributions in data collection and processing. We thank Han-Gyol Yi for his feedback on the analyses. We thank the two peer reviewers of the manuscript for their constructive feedback and their ideas on how to reduce the number of FFR trials required for the HMM to decode Mandarin tones within subjects across different experimental sections.

ETHICAL STANDARDS

I have read and have abided by the statement of ethical standards for manuscripts submitted to the Journal of Neuroscience Methods.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial, or otherwise.

REFERENCES

Alpaydin, E. (2014). *Introduction to machine learning*: MIT press.

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., & Krishnan, A. (2010). Effects of reverberation on brainstem representation of speech in musicians and non-musicians. *Brain research*, *1355*, 112-125.

Bidelman, G. M., Gandour, J. T., & Krishnan, A. (2011). Cross-domain effects of music and language experience on the representation of pitch in the human auditory brainstem. *Journal of cognitive neuroscience, 23*(2), 425-434.

Bidelman, G. M., Moreno, S., & Alain, C. (2013). Tracing the emergence of categorical speech perception in the human auditory system. *Neuroimage, 79*, 201-212.

Boersma, P. (1993). *Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound.* Paper presented at the Proceedings of the institute of phonetic sciences.

Chandrasekaran, B., & Kraus, N. (2010). The scalp-recorded brainstem response to speech: Neural origins and plasticity. *Psychophysiology, 47*(2), 236-246.

Chandrasekaran, B., Kraus, N., & Wong, P. C. (2012). Human inferior colliculus activity relates to individual differences in spoken language learning. *Journal of Neurophysiology, 107*(5), 1325-1336.

Chandrasekaran, B., Krishnan, A., & Gandour, J. T. (2007). Mismatch negativity to pitch contours is influenced by language experience. *Brain research, 1128*, 148-156.

Chandrasekaran, B., Skoe, E., & Kraus, N. (2014). An integrative model of subcortical auditory plasticity. *Brain topography, 27*(4), 539-552.

Chandrasekaran., Skoe, E., & Kraus, N. (2014). An integrative model of subcortical auditory plasticity. *Brain Topogr, 27*(4), 539-552. doi:10.1007/s10548-013-0323-9

Chang, C. C., & Hu, Y. C. (1998). A fast LBG codebook training algorithm for vector quantization. *IEEE Transactions on Consumer Electronics, 44*(4), 1201-1208.

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., & Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nature neuroscience, 13*(11), 1428-1432.

Collins, M. J., & Cullen Jr, J. K. (1978). Temporal integration of tone glides. *The Journal of the Acoustical Society of America, 63*(2), 469-473.

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*.

Equitz, W. H. (1989). A new vector quantization clustering algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 37*(10), 1568-1575.

Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental conservation, 24*(01), 38-49.

Francis, A. L., Ciocca, V., & Ng, B. K. C. (2003). On the (non) categorical perception of lexical tones. *Perception & psychophysics*, *65*(7), 1029-1044.

Gales, M., & Young, S. (2008). The application of hidden Markov models in speech recognition. *Foundations and trends in signal processing, 1*(3), 195-304.

Gandour, J. T. (1983). Tone perception in far eastern-languages. *Journal of Phonetics, 11*(2), 149-175.

Gandour, J. T. (1994). Phonetics of tone. *The encyclopedia of language & linguistics, 6*, 3116-3123.

Gardner, M. J., & Altman, D. G. (1989). Statistics with confidence: confidence intervals and statistical guidelines.

Gorina-Careta, N., Zarnowiec, K., Costa-Faidella, J., & Escera, C. (2016). Timing predictability enhances regularity encoding in the human subcortical auditory pathway. *Scientific reports*, *6*, 37405.

Hausfeld, L., De Martino, F., Bonte, M., & Formisano, E. (2012). Pattern analysis of EEG responses to speech and voice: Influence of feature grouping. *Neuroimage, 59*(4), 3641-3651.

Huang, X. D., Ariki, Y., & Jack, M. A. (1990). *Hidden Markov models for speech recognition* (Vol. 2004): Edinburgh university press Edinburgh.

Jeng, F. C., Hu, J., Dickman, B., Montgomery-Reagan, K., Tong, M., Wu, G., & Lin, C. D. (2011). Cross-linguistic comparison of frequency-following responses to voice pitch in American and Chinese neonates and adults. *Ear and hearing*, *32* (6), 699-707.

Kekre, H., & Sarode, T. K. (2008). Speech data compression using vector quantization. *WASET International Journal of Computer and Information Science and Engineering (IJCISE), 2*(4), 251-254.

Keuroghlian, A. S., & Knudsen, E. I. (2007). Adaptive auditory plasticity in developing and adult animals. *Progress in neurobiology, 82*(3), 109-121.

Knudsen, E. I. (2002). Instructed learning in the auditory localization pathway of the barn owl. *Nature, 417*(6886), 322-328.

Krishnan, A., Bidelman, G. M., & Gandour, J. T. (2010). Neural representation of pitch salience in the human brainstem revealed by psychophysical and electrophysiological indices. *Hearing research*, *268*(1), 60-66.

Krishnan, A., Gandour, J. T., & Bidelman, G. M. (2010). The effects of tone language experience on pitch processing in the brainstem. *Journal of Neurolinguistics, 23*(1), 81-95.

Krishnan, A., Gandour, J. T., & Bidelman, G. M. (2012). Experience-dependent plasticity in pitch encoding: from brainstem to auditory cortex. *Neuroreport, 23*(8), 498.

Krishnan, A., Gandour, J. T., Bidelman, G. M., & Swaminathan, J. (2009). Experience dependent neural representation of dynamic pitch in the brainstem. *Neuroreport, 20*(4), 408.

Krishnan, A., & Parkinson, J. (2000). Human frequency-following response: representation of tonal sweeps. *Audiology and Neurotology, 5*(6), 312-321.

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

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.

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

Krizman, J., Skoe, E., & Kraus, N. (2012). Sex differences in auditory subcortical function. *Clinical Neurophysiology, 123*(3), 590-597.

Krizman, J., Skoe, E., Marian, V., & Kraus, N. (2014). Bilingualism increases neural response consistency and attentional control: Evidence for sensory and cognitive coupling. *Brain and language, 128*(1), 34-40.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26): Springer.

Ladefoged, P., & Maddieson, I. (1998). The sounds of the world's languages. *Language, 74*(2), 374-376.

Lau, J. C., Wong, P. C., & Chandrasekaran, B. (2016). Context-dependent plasticity in the subcortical encoding of linguistic pitch patterns. *Journal of Neurophysiology*, jn. 00656.02016.

Linkenhoker, B. A., von der Ohe, C. G., & Knudsen, E. I. (2005). Anatomical traces of juvenile learning in the auditory system of adult barn owls. *Nature neuroscience, 8*(1), 93-98.

Lisker, L. (1986). “Voicing” in English: a catalogue of acoustic features signaling/b/versus/p/in trochees. *Language and speech, 29*(1), 3-11.

Liu, F., Maggu, A. R., Lau, J. C., & Wong, P. (2015). Brainstem encoding of speech and musical stimuli in congenital amusia: evidence from Cantonese speakers. *Frontiers in human neuroscience, 8*, 1029.

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological review, 118*(2), 219.

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science, 343*(6174), 1006-1010.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of machine learning*: MIT press.

Musacchia, G., Sams, M., Skoe, E., & Kraus, N. (2007). Musicians have enhanced subcortical auditory and audiovisual processing of speech and music. *Proceedings of the National Academy of Sciences, 104*(40), 15894-15898.

Patel, A. D. (2010). *Music, language, and the brain*: Oxford university press.

Pei, X., Barbour, D. L., Leuthardt, E. C., & Schalk, G. (2011). Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans. *Journal of neural engineering, 8*(4), 046028.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257-286.

Rabiner, L. R., & Juang, B.-H. (1986). An introduction to hidden Markov models. *ieee assp magazine, 3*(1), 4-16.

Rabiner, L. R., & Juang, B.-H. (1993). Fundamentals of speech recognition.

Remez, R. E. (2005). Perceptual organization of speech. *The handbook of speech perception*, 28-50.

Shiga, T., Althen, H., Cornella, M., Zarnowiec, K., Yabe, H., & Escera, C. (2015). Deviance-related responses along the auditory hierarchy: Combined FFR, MLR and MMN evidence. *PloS one*, *10*(9), e0136794.

Siuly, S., Li, Y., & Zhang, Y. (2017). *EEG Signal Analysis and Classification: Techniques and Applications*: Springer.

Skoe, E., & Chandrasekaran, B. (2014). The layering of auditory experiences in driving experience-dependent subcortical plasticity. *Hearing research, 311*, 36-48.

Skoe, E., & Kraus, N. (2010). Auditory brainstem response to complex sounds: a tutorial. *Ear and hearing, 31*(3), 302.

Slabu, L., Grimm, S., & Escera, C. (2012). Novelty detection in the human auditory brainstem. *Journal of Neuroscience*, *32*(4), 1447-1452.

Smith, J. C., Marsh, J. T., & Brown, W. S. (1975). Far-field recorded frequency-following responses: evidence for the locus of brainstem sources. *Electroencephalography and clinical neurophysiology, 39*(5), 465-472.

Sohmer, H., Pratt, H., & Kinarti, R. (1977). Sources of frequency following responses (FFR) in man. *Electroencephalography and clinical neurophysiology, 42*(5), 656-664.

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.

Viikki, O., & Laurila, K. (1997). *Noise robust HMM-based speech recognition using segmental cepstral feature vector normalization.* Paper presented at the Robust Speech Recognition for Unknown Communication Channels.

Whalen, D. H., & Xu, Y. (1992). Information for Mandarin tones in the amplitude contour and in brief segments. Phonetica, 49(1), 25-47.

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.

Wu, N. and Shu, H. (2003). The gating paradigm and spoken word recognition of Chinese. *Acta Psychologica,* 35 (5): 582-590.

Xie, Z., Reetzke, R. D., & Chandrasekaran, B. (2017). Stability and plasticity in neural encoding of linguistically-relevant pitch patterns. *Journal of Neurophysiology*, jn. 00445.02016.

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of phonetics*, *25*(1), 61-83.

Xu, Y., Krishnan, A., & Gandour, J. T. (2006). Specificity of experience-dependent pitch representation in the brainstem. *Neuroreport, 17*(15), 1601-1605.

Yang, W.-J., Lee, J.-C., Chang, Y.-C., & Wang, H.-C. (1988). Hidden Markov model for Mandarin lexical tone recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 36*(7), 988-992.

Yi, H.-G., Xie, Z., Reetzke, R., Dimakis, A. G., & Chandrasekaran, B. (2017). Vowel decoding from single-trial speech-evoked electrophysiological responses: A feature-based machine learning approach. *Brain and Behavior*.