INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Context regularity indexed by auditory N1 and P2 event-related potentials

Xiao Wang [1], *Yanhui Zhang*[1], *Gang Peng*[2,3]

[1] The Chinese University of Hong Kong, Hong Kong
[2] The Hong Kong Polytechnic University, Hong Kong
[3] Shenzhen Institutes of Advanced Integration Technology, Chinese Academy of Sciences, China

JoyceWang@link.cuhk.edu.hk, yhzhang@cuhk.edu.hk, gpeng@polyu.edu.hk

Abstract

It is still a question of debate whether the N1-P2 complex is an index of low-level auditory processes or whether it can capture higher-order information encoded in the immediate context. To address this issue, the current study examined the morphology of the N1-P2 complex as a function of context regularities instantiated at the sublexical level. We presented two types of speech targets in isolation and in contexts comprising sequences of Cantonese words sharing either the entire rime units or just the rime segments (thus lacking lexical tone consistency). Results revealed a pervasive yet unequal attenuation of the N1 and P2 components: The degree of N1 attenuation tended to decrease while that of P2 increased due to enhanced detectability of more regular speech patterns, as well as their enhanced predictability in the immediate context. The distinct behaviors of N1 and P2 event-related potentials could be explained by the influence of perceptual experience and the hierarchical encoding of context regularities.

Index Terms: speech perception, context regularity, event-related potentials, N1-P2 complex, amplitude attenuation

1. Introduction

To attune to the rapidly unfolding speech events, humans have evolved the ability to predict incoming acoustic signals based on previous sound input [1, 2]. A well-known example is the elicitation of the mismatch negativity (MMN) by sounds violating the structural regularities in the immediate context [3-8]. However, MMN (peaking around 150–250 ms) is not the only auditory event-related potential (ERP) indexing regularity detection and decomposition. ERPs in earlier or overlapping time windows have also been shown to serve similar functions. Of great interest to this study is the N1-P2 complex and their morphological changes [4].

The N1-P2 complex is best known for its sensitivity to low-level acoustic processing. For instance, studies have shown that the latencies and the amplitudes of the N1 and P2 inflections decrease with increased frequency range [9]. While in response to enhanced stimulus intensity, the amplitude of the complex increases, whether the measurement is taken by the peak-to-peak or the baseline-to-peak method [10, 11]. Clinicians even found it feasible to use the morphology of the N1-P2 complex to predict patients' perceptual thresholds [12].

Nonetheless, N1 and P2 are not simple indexes of auditory processing. As evinced by their habituation [4, 13], changes in the N1-P2 morphology are also sensitive to the sound statistics in the broader context. Besides habituation by repetition, N1 amplitude is also sensitive to stimulus predictability, whose formation requires more elaborate mental computations and knowledge higher on the conceptual hierarchy. Take temporal predictability as an example: Participants in [2] were presented with trains of pure tones. Stimuli in each train started off at a distinct pitch height and remained constant for at least 3 presentations. It was found that the repetition of a pure tone stimulus led to pronounced neural adaptations in the N1 time window, provided that the stimulus onset asynchrony (SOA) was kept constant. By contrast, P2 amplified with repetition in [2], changing in the opposite direction from that reported in [13] and was unconstrained by the temporal predictability of SOA. The authors in [2] attributed these distinctive behaviors of the N1 and P2 components to their differential sensitivity, suggesting that N1 might encode the "when" aspect (i.e., temporal regularity) of the sound events, whereas P2 pertains more closely to the "what" aspect (i.e., the identity) of the sound objects. What is still left open is the question of whether such differential encodings of sound regularities in the N1-P2 window are likewise at play during speech recoding.

Whether differential regularity encoding as reflected by the morphology of the N1-P2 complex can be observed at sublexical phonological levels is currently at issue as well. Most studies in the current literature have focused on the sound regularities realized at the lexical level. Few have manipulated the phoneme makeup of context stimuli with the aim of investigating the scope of the context-dependent regularity encoding or the depth of sublexical processing which the N1 and P2 components can capture. Such a question may be particularly worth exploring among Chinese speakers given their preference for holistic encoding strategies [14, 15]. For under the strong influence of holistic processing, it is possible for sublexical regularities to escape the notice of these participants, which may in turn hinder the efficiency of regularity coding. Nonetheless, the basic prediction goes that if regularity encoding can penetrate sublexical levels, then in the N1 time window, where the directionalities of the repetition and prediction based suppression effects are identical [2, 6], enhanced intelligibility would translate into greater amplitude attenuation. As for the directionality of changes in the P2 time window, it cannot be determined at this point, as well as how sublexical regularities may interact with the temporal predictability of the speech objects.

2. Method

Sixteen native speakers of Hong Kong Cantonese (male = 7) without any prior history of brain and hearing impairment were recruited for this study. All were right handed and were non-musicians. Participants all gave their written informed consent before the experiment.

http://dx.doi.org/10.21437/Interspeech.2017-658

The current study adapted the paradigm summarized in [2, 5], presenting 128 test trials over the course of 40 minutes. In half of the test trials, target stimuli were presented in isolation; whereas in the rest, targets were presented in contexts consisted of four Cantonese words having either the same rimes (low-variability: /tai1/, /kai1/, /pai1/, /dai1/) or just the same rime segments (high-variability: /tai4/, /kai2/, /pai1/, /dai6/). As can be seen, with the segmental content being identical, the crucial difference between the high- and low-variability contexts was lexical tone consistency at the acoustic and phonemic levels. Moreover, consistent with [7, 16], ERPs elicited by targets in isolation were used as the baseline, included to separate the neuronal responses elicited by stimuli-specific sound properties from those specific to the context. We also included filler trials (11%) presenting noise bursts. In each trial, the position of the noise burst was random: It might substitute either the targets or one of the context stimuli. Naturally, the unpredictability of the noise required sustained and active attention from the participants.

Additionally, to tease apart the effects of repetition and prediction, both of which modulate the amplitudes of the N1 and P2 components, we presented two types of targets in this study, each forming a distinct relation with the preceding context (hereafter, Target-Context Relation, or TCR). At the sublexical level, the targets may conform to the sound patterns of the context in every aspect (i.e., *shared-rime* condition), or they might contain a violation of expectation in the nucleus position (i.e., *nucleus-violation* condition). For instance, /bei1/ in the low-variability context created a case of nucleus violation: It comprised the same semi-vowel and tone categories as did the context stimuli. And just like the context stimuli, the onsets of the targets differed from both the immediately adjacent sounds and their more distant neighbors. The only unpredictable element, in this case, was the nucleus. Across contexts and sessions, the two types of targets occurred with equal probability. All stimuli, noise included, were normalized in duration (500 ms) and intensity (70 dB).

During the experiment, participants needed to make motor responses (i.e., pressing the "N" button on the keyboard) to the noise bursts, not to the target words on which context effects were expected and measured. Trials with high and low stimulus variability were blocked and divided into smaller sessions. For each participant, the presentation order of the filler and target trials were randomized prior to each session; across participants, blocks were presented in counterbalanced order. Practice sessions were provided before each block. Moreover, instead of a constant inter-stimulus interval (ISI), ISI was jittered in the present study (500–800 ms).

64-channel EEG data (1kHz sampling rate) were recorded using the Curry 7 neuroimaging suite. The electrodes were positioned following the International 10-20 system. Prior to statistical analysis, EEG data were band-pass filtered (0.1-30 Hz), rejected (trials with potentials exceeding 100μV at any electrode), corrected for artifacts using the regression-based methods [17], segmented into epochs (-0.2 to 0.79 ms) time locked to the onset of targets, baseline corrected, and off-line re-referenced to the average signals of the mastoids. To isolate context effects while maximally reducing the influence of ERPs idiosyncratic to each target stimulus, all statistical analyses were based on difference waves. ERPs elicited by targets presented in isolation were subtracted from those elicited by corresponding targets presented in contexts. In the following analyses, we mainly focus on changes in ERP amplitudes, rather than latencies, as amplitudes are generally the more sensitive and reliable indexes of context effects.

3. Results

For each condition, we exported ERP data from 45 channels, which were further divided into nine areas according to their scalp distributions: left anterior (F3, F5, F7, FC3, FC5), left central (C3, C5, CP3, CP5, TP7), left posterior (P3, P5, P7, O1, PO7), mid anterior (F1, F2, FCZ, FC1, FC2), mid central (C1, CP1, CZ, C2, CP2), mid posterior (P1, P2, PZ, POZ, OZ), right anterior (F4, F6, F8, FC4, FC6), right central (C4, C6, CP4, CP6, TP8), and right posterior (P4, P6, P8, PO8, O2). Figure 1 presents the grand averaged ERP waveforms at the CZ electrode along with the topographic plots of major ERP components collapsed across *TCR* conditions. As can be seen, participants' brain responses show typical N1 (60–135 ms), P2 (135–245 ms) and N4 (245–555 ms) components. ERP windows were determined for each component based on the global field power (see Figure 2).

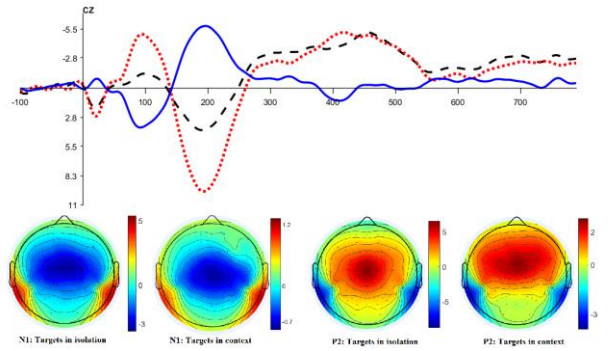

Figure 1. *Overview of N1 and P2 amplitude attenuation. The upper panel shows ERP waveforms recorded at the CZ electrode. Dotted line: ERPs elicited by presenting targets in isolation; dashed line: ERPs elicited by targets presented in contexts; solid line: Difference waves obtained by subtracting the ERPs elicited by targets in contexts from ERPs elicited in isolation. The bottom panel displays the topographic maps in the N1 and P2 windows collapsed across TCR conditions.*

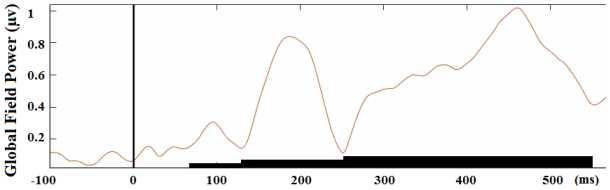

Figure 2. *Global field power. Black bars on the lower x-axis represent the time ranges for the N1, P2, and N4 components.*

3.1. N1 attenuation

To delineate the attenuation patterns for the N1 component, difference waves were calculated for each subject, within each *Context* and *TCR* conditions, and submitted to a four-way repeated measures ANOVA, with *Laterality* (left, midline, right), *Posterity* (anterior, central, posterior), *Context* (low- and, high-variability), and *TCR* (shared-rime, nucleus-violation) as the with-subject variables. If applicable, the Greenhouse-Geisser method was used to correct violations of the Sphericity assumption. Significant main effects were observed for all variables (all *p*s < .05), as well as the significant two-way interactions between all but one variable

pairs (*Laterality* × *Context*: p = .067). The three-way *Laterality* × *Context* × *TCR* interaction also reached statistical significance [F (1.863, 147.162) = 4.199, p = .019], suggesting that while ERP amplitudes showed an extensive reduction in the N1 time window, the degree of attenuation was unequal across conditions. To provide a detailed delineation, three-way (*TCR* × *Context* × *Posterity*) ANOVAs were separately carried out for the left, midline, and right electrode sites.

For the left electrode sites, highly significant main effects were found for *Posterity* [F (2, 158) = 28.861, p < .001]. Post hoc analyses suggested that this was caused by the significantly larger amplitude reductions observed in the anterior and central regions than in the posterior sites. *TCR* was also statistically significant, [F (1, 79) = 21.754, p < .001], as the degree of absolute voltage reduction was much larger in the nucleus-violation condition than in the shared-rime condition. The role of *Context* also reached statistical significance [F (1, 79) = 4.500, p < .05], with the high-variability context eliciting greater attenuation than its low-variability counterpart.

At the same time, highly significant *Context* × *Posterity* [F (1.856, 146.616) = 18.620, p < .001] and *TCR* × *Posterity* [F (1.885, 148.907) = 22.276, p < .001] interactions were found in the three-way ANOVA test. It turned out that N1 amplitude only attenuated significantly more in the nucleus-violation condition in the anterior and central regions (all ps < .001), consistent with the N1 scalp distributions observed in the literature. While for the modulating effects of *Context*, they were mainly captured by electrodes in the anterior, rather than the central (p = 0.081) or posterior (p = 0.128) regions.

Qualitatively similar patterns were observed for N1 data recorded along the midline. First, *Posterity* [F (1.596, 126.057) = 70.874, p < .001], *Context* [F (1, 79) = 7.452, p < .01], and *TCR* [F (1, 79) = 21.802, p < .001] all reached statistical significance (ps < .01), together with two-way interactions between all possible pairs of variables (ps < .005). Again, post hoc analysis showed that N1 amplitude decreased more drastically in the nucleus-violation condition as opposed to the shared-rime condition over the anterior and central regions, and was again more so in the high-variability context (p < .001) than in the low-variability context (p = .015); although unlike the left sites, the superior role of the high-variability context in eliciting N1 attenuation propagated over a much wider region along the midline, spreading across the mid-anterior (p < .001) and the mid-central (p = .029) regions.

Data obtained from the right electrode sites were of a highly similar nature as well. The only difference was that for the *Context* by *TCR* interaction [F (1, 79) = 8.907, p < .004], it was driven mainly by the responses to the nucleus-violation condition in the high-variability context (p < .001), not the one with low stimulus variability (p = .627).

3.2. P2 amplitude reduction

It is obvious from Figure 1 that presenting targets in contexts also led to a reduction of P2 amplitude. To explore such a phenomenon, a four-way repeated measures ANOVA was carried out on P2 amplitude data, with *Laterality*, *Posterity*, *Context*, and *TCR* as the within-subject variables. Highly significant main effects of *Laterality* [F (1.825, 144.205) = 102.864, p < .001] and *Posterity* [F (1.415, 111.760) = 102.864, p < .001] were found, along with significant two-way interactions: *Posterity* × *Laterality* [F (2.480, 195.959) = 102.864, p < .001], *Laterality* × *TCR* [F (1.673, 132.136) = 8.626, p < .001], and *Posterity* × *TCR* [F (1.793, 141.659) = 22.369, p < .001]. There was also a significant three-way (*Posterity* × *Laterality* × *TCR*) interaction [F (3.728, 294.505) = 3.445, p < .05], suggesting that similar to N1, P2 amplitude also responded differentially to variable manipulations. To obtain a more detailed delineation, three-way ANOVAs were carried out on data split by *Laterality*.

For the left sites, *Posterity* was a major determinant of P2 morphology [F (2, 158) = 11.146, p < .001]. And consistent with the centro-frontal and parieto-occipital distribution of P2, large amplitude decrements were found over the central and posterior electrode sites. P2 attenuation varied as a function of TCR as well, as evidenced by the *Posterity* by *TCR* interaction [F (1.821, 143.896) = 7.199, p < .005]. Compared to anterior sites, P2 attenuated more in the central (p < .001) and posterior (p = 0.010) areas due to nucleus violations. Whereas in the shared-rime condition, P2 was most pronouncedly attenuated in the central regions (all ps < .005), which even surpassed that induced by violations of nucleus expectancies (p = 0.038).

Moreover, the three-way ANOVA test revealed a *Context* × *TCR* interaction [F (1, 79) = 13.035, p < .001]. In the high-variability context, P2 attenuation in the shared-rime condition was more prominent than that elicited by nucleus violations (p = 0.028). The capacity of the two TCR conditions in eliciting amplitude attenuation was thus reversed across the N1 and P2 windows. Also reversed was the directionality of *Context* effects: Unlike N1, P2 decreased more drastically when targets from the nucleus-violation condition were embedded in the low-variability context (p = 0.007). Analyses for the midline and right electrode sites yielded qualitatively similar results.

In sum, although both N1 and P2 attenuated because of the surrounding contexts, they behaved in qualitatively different ways. First, there was a reversal of the *Context* effects on N1 and P2: While the low-variability context led to larger P2 attenuation in the nucleus-violation condition, in the N1 time window it was the high-variability context that induced greater attenuation. The *TCR* condition driving the amplitude reduction also differed: whereas N1 amplitude relied critically on the ERPs elicited in the nucleus-violation condition; for P2, it was the shared-rime condition that produced the more salient decrements in amplitude.

4. Discussion

Results of this study corroborated existing literature by showing that (1) the human auditory system has the ability to detect and encode context regularities [1, 2, 4, 7], and that (2) such a process gradually unfolds in the time windows of the N1 and P2 components [2, 6, 13]. However, unlike previous research, the present study increased the depth of processing that was needed to extract context regularities. We varied the amount of phonemic overlap among context stimuli, extending context regularities to the sublexical, rather than the lexical, level. Our data also showed that sublexical regularity encoding could be extended to tonal language users having a strong preference for holistic speech encoding [14]. It thus seems that the need to register regularities in the context has overridden the influence of long-term, tonal-language experience, thereby modulating the lexical retrieval units and strategies listeners might use. This is the first observation in this study that merits further investigations. And given its robustness, this dynamic interaction between the immediate sound input and the speech habit fostered by previous perceptual experience needs to be accounted for by human speech perception models.

Moreover, this study constructed two types of contexts to explore the source of the N1 and P2 amplitude changes observed in the literature: whether they are driven by stimuli repetition or the ease of prediction. One context had lower variability at the acoustic and phonemic levels, because it comprised stimuli sharing both the rime segments and the lexical tones. The other context had higher variability, since it contained stimuli having identical rime segments but distinct lexical tones. Intuitively, decreased stimulus variability would increase the intelligibility (i.e., detectability and salience) of sublexical sound patterns, which would, in turn, lead to greater repetition suppression effects observed in [2, 4]. Interestingly, though, such a prediction was not borne out by our data. We found that while suppression was a prevalent phenomenon in the 60–135 ms time window, N1 attenuated significantly more in the high-variability context. Such indifference of N1 toward stimulus variability is beyond the scope of the repetition suppression account.

Results also revealed a higher degree of N1 attenuation in the nucleus-violation condition as compared with the shared-rime condition. This is another finding that runs counter to the repetition suppression theory. For instead of replication, major sublexical components unfolding in the nucleus-violation condition in the N1 window, i.e. onset and nucleus, differed between targets and context stimuli. If repetition suppression were to apply to this condition, smaller attenuation would be expected instead. Studies also observed that N1 attenuated as attention decreased [18], but this does not explain our data either, as attention is typically re-oriented toward deviances, rather than being directed away from them [7, 8].

What then caused larger N1 attenuation to surface in the high-variability context? To answer this question, it may be useful to consider the influence of short-term perceptual experience. It is possible that presenting the high- and low-variability trials in separate blocks led to differential adaptations to stimulus variability: In the high-variability block, acoustic and phonemic variabilities themselves might have been accepted as the norm due to continuous exposure to stimulus variations. Similarly, the consistency of rimes in the context stimuli of the low-variability block reduced listeners' perceptual tolerance for sound variations. This makes targets containing nucleus violations a much poorer fit to the low-variability context relative to its high-variability counterpart.

Viewed in a broader context, the modulating effects of contextual sound regularities discussed above aligned well with the literature on cross-linguistic speech perception. In [19], for example, the authors used hemodynamic responses to examine the neural correlates of Japanese vowel length contrasts. They found that Korean learners of Japanese did not activate the left temporal lobe as much as native Japanese speakers did, showing a relative indifference to durational changes. Differential brain responses observed in native and learner brains was especially apparent when pairs of stimuli straddling the categorical boundaries of Japanese long and short vowels were presented. One explanation for this is the non-contrastive nature and the communicative functions of vowel duration in Korean [20], as they allow Koreans to be more liberal in speech productions. To adapt to this variability, it follows that Korean users need to be more lenient with duration changes. Behaviorally, this leniency translates into enlarged perceptual tolerability and duration thresholds; neurologically, it results in lower activation levels in learners' primary auditory cortex [19]. [21] obtained similar results, reporting weaker cortical activities in learners with low L2 proficiency. That perceptual experience shapes our tolerability for sound variations has also been attested across domain boundaries. For instance, due to the functional significance of semitones and fine-grained pitch movements, pitch perception tends to be more stringent in music than in speech [20].

Unequal decrements following the encoding of context regularities have also been observed in this study in the P2 time window. However, different from N1, P2 elicited by the nucleus-violation condition attenuated to a greater degree in the low-variability context. Explanations for this may lie in the hierarchical models of sensory processing and predictive coding [2, 6]. Both of these models contend that in addition to monotonous repetition, neural activities diminish when signals align with higher-order expectations derived from complex context regularities, although attenuation by expectation tends to have a longer latency than that by repetition. Given that targets with the same rime segments as the context stimuli are more regular and predictable (in terms of their sublexical structures) than targets with nucleus violations, attenuation in the P2 time window was more likely the result of prediction-based, higher-order cognitive computations taking control over lower-level processes underlying repetition suppression. As such, our data also lend support to the claim that N1 and P2 can encode distinct neural activities and may thus be viewed as independent components rather than integral parts of the vertex potential complex [22].

In addition, the distinct behaviors of N1 and P2 may shed light on the hybrid view of speech representations, which maintains that phoneme categories are stored in the brain not just by abstract phonological features, but also by sensory memory traces encoding speaker- and stimuli-specific information [23, 24]. In line with this claim, our data showed that N1 and P2 morphology are modulated by sound patterns at both the surface (e.g., global impression of stimulus variability) and the sublexical phonological levels (e.g., rime consistency). As to the level of information that plays the more predominant role, it may be determined by the encoding efficiency of the neural networks underlying N1 and P2. Conceivably, a lot of individual differences can be expected in the way speech regularities are registered [25].

Finally, our data showed that despite the lack of temporal predictability of the context and target stimuli (as a result of ISI jittering), N1 and P2 attenuation reliably occurred. This is at odds with the findings in both [26] and [2] and the "what"-and-"when" explanation proposed therein. When presented with speech samples from their native languages, rather than pure tones as in [2], it seems that neither fixed SOAs nor temporal cueing is a prerequisite for deriving sound expectations. The ability of the human auditory system to filter out surface variations thus seems to operate in the frequency [23] as well as the temporal dimensions. It is also possible that similar to frequency normalization [23, 24], the tolerability for temporal variations is another evolutionarily motivated and non-modality-specific human trait. Future studies are needed to verify such a claim and to delineate the scope and workings of the mechanisms supporting such temporal tolerance.

5. Acknowledgements

This study was supported in part by grants from National Natural Science Foundation of China (NSFC: 11474300), National Social Science Fund of China (13&ZD189), and Research Grant Council of Hong Kong (GRF: 14411314).

6. References

[1] I. Winkler, L. D. Susan, and N. Israel, "Modeling the auditory scene: predictive regularity representations and perceptual objects," *Trends in Cognitive Sciences*, vol. 13, no. 12, pp. 532–540, 2009.

[2] J. Costa-Faidella, T. Baldeweg, S. Grimm, and C. Escera, "Interactions between "what" and "when" in the auditory system: temporal predictability enhances repetition suppression," *Journal of Neuroscience*, vol. 31, no. 50, pp.18590–18597, 2011.

[3] H. Tiitinen, P. May, K Reinikainen, and R. NÃ, "Attentive novelty detection in humans is governed by pre-attentive sensory memory," *Nature*, vol. 372, no. 6501, pp. 90–92, 1994.

[4] R. Näätänen and T. Rinne. "Electric brain response to sound repetition in humans: an index of long-term-memory–trace formation?" *Neuroscience Letters*, vol. 318, no. 1, pp. 49–51, 2002.

[5] E. Sussman, I. Winkler, M. Huotilainen, W. Ritter, and R. Näätänen, "Top-down effects can modify the initially stimulus-driven auditory organization," *Cognitive Brain Research*, vol. 13, no. 3, pp. 393–405, 2002.

[6] A. Todorovic and F. P. de Lange, "Repetition suppression and expectation suppression are dissociable in time in early auditory evoked fields," *Journal of Neuroscience*, vol. 32, no. 39, pp. 13389–13395, 2012.

[7] E. S. Sussman, S. Chen, J. Sussman-Fort, and E. Dinces, "The five myths of MMN: redefining how to use MMN in basic and clinical research," *Brain Topography*, vol. 27, no. 4, pp. 553–564, 2014.

[8] A. Calcus, P. Deltenre, I. Hoonhorst, G. Collet, E. Markessis, and C. Colin, "MMN and P300 are both modulated by the featured/featureless nature of deviant stimuli," *Clinical Neurophysiology*, vol. 126, no. 9, pp. 1727–1734, 2015.

[9] J. L. Wunderlich and B. K. Cone-Wesson, "Effects of stimulus frequency and complexity on the mismatch negativity and other components of the cortical auditory-evoked potential," *The Journal of the Acoustical Society of America*, vol. 109, no. 4, pp. 1526–1537, 2001.

[10] T. W. Picton, W. S. Goodman, and D. P. Bryce, "Amplitude of evoked responses to tones of high intensity," *Acta Oto-Laryngologica*, vol. 70, no. 2, pp. 77–82, 1970.

[11] G. Adler and J. Adler, "Influence of stimulus intensity on AEP components in the 80-to 200-millisecond latency range," *Audiology*, vol. 28, no. 6, pp. 316–324, 1989.

[12] G. Lightfoot, "The N1-P2 cortical auditory evoked potential in threshold estimation," *Insights in Practice for Clinical Audiology*, 2010.

[13] J. Rust, "Habituation and the orienting response in the auditory cortical evoked potential," *Psychophysiology*, vol. 14, no. 2, pp. 123–126, 1977.

[14] T. Liu and J. H. W. Hsiao, "Holistic Processing in Speech Perception: Experts' and Novices' Processing of Isolated Cantonese Syllables," In *Cogsci 2014 – 36th Annual Conference of the Cognitive Science Society, July 23-26, Québec City, Canada, Proceedings*, 2014, pp. 869–874.

[15] X. Wang and G. Peng, "Cantonese Spoken Word Retention by Speakers with and without Congenital Amusia: Implications from Phonological Similarity and Cognitive Load Effects," *IEEE SigPort*, 2016.

[16] N. Kraus, T. McGee, T. D. Carrell, and A. Sharma, "Neurophysiologic bases of speech discrimination," *Ear and Hearing*, vol. 16, no. 1, pp. 19–37, 1995.

[17] G. Gratton, M. G. Coles, and E. Donchin. "A new method for off-line removal of ocular artifact," *Electroencephalography and Clinical Neurophysiology*, vol. 55, no. 4, pp. 468–484, 1983.

[18] S. A. Hillyard, R. F. Hink, V. L. Schwent, and T. W. Picton, "Electrical signs of selective attention in the human brain," *Science*, vol. 182, no. 4108, pp. 177–180, 1973.

[19] Y. Minagawa-Kawai, K. Mori, and Y. Sato, "Different brain strategies underlie the categorical perception of foreign and native phonemes," *Journal of Cognitive Neuroscience*, vol. 17, no. 9, pp. 1376–1385, 2005.

[20] Y. Xu, "Speech melody as articulatorily implemented communicative functions," *Speech Communication*, vol. 46, no. 3, pp. 220–251, 2005.

[21] S. Dehaene, E. Dupoux, J. Mehler, L. Cohen, E. Paulesu, D. Perani, P. F. Van de Moortele, S. Lehéricy, and D. Le Bihan, "Anatomical variability in the cortical representation of first and second language," *Neuroreport*, vol. 8, no. 17, pp. 3809–3815, 1997.

[22] K. E. Crowley and I. M. Colrain. "A review of the evidence for P2 being an independent component process: age, sleep and modality," *Clinical neurophysiology*, vol. 115, no. 4, pp. 732–744, 2004.

[23] P. C. Wong, H. C. Nusbaum, and S. L. Small, "Neural bases of talker normalization," *Journal of Cognitive Neuroscience*, vol. 16, no. 7, 1173–1184, 2004.

[24] C. C. Zhang, "Perceptual Normalization of Inter-and Intra-talker Variations in Tone Categorization," Ph.D. dissertation, Dept. Ling., CUHK, Hong Kong, 2014.

[25] T. K. Perrachione, J. Lee, L. Y. Y. Ha, and P. C. M. Wong, "Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design," *The Journal of the Acoustical Society of America*, vol. 130, no. 1, 461–472, 2011.

[26] P. F. Sowman, A. Kuusik, and B. W. Johnson, "Self-initiation and temporal cueing of monaural tones reduce the auditory N1 and P2," *Experimental Brain Research*, vol. 222, no. 1-2, pp. 149–157, 2012.