Cerebral Cortex, 2023, **33**, 9542–9553
https://doi.org/10.1093/cercor/bhad225
Advance access publication date 21 June 2023
Original Article

Segregation and integration of sensory features by flexible temporal characteristics of independent neural representations

Zhili Han[1,2], Hao Zhu[2,3], Yunyun Shen[1,2,4], Xing Tian[1,2,3,*]

[1]Shanghai Key Laboratory of Brain Functional Genomics (Ministry of Education), School of Psychology and Cognitive Science, East China Normal University, Shanghai 200062, China,
[2]NYU-ECNU Institute of Brain and Cognitive Science at NYU Shanghai, Shanghai 200062, China,
[3]Shanghai Frontiers Science Center of Artificial Intelligence and Deep Learning; Division of Arts and Sciences, NYU Shanghai Shanghai 200126, China,
[4]Cognitive Neuroimaging Unit, INSERN, CEA, CNRS, Universite Paris-Saclay, Neuronspin Center, Gif Yvette 91191, France

*Corresponding author: New York University Shanghai, Shanghai 200122, China. Email: xing.tian@nyu.edu

Segregation and integration are two fundamental yet competing computations in cognition. For example, in serial speech processing, stable perception necessitates the sequential establishment of perceptual representations to remove irrelevant features for achieving invariance. Whereas multiple features need to combine to create a coherent percept. How to simultaneously achieve seemingly contradicted computations of segregation and integration in a serial process is unclear. To investigate their neural mechanisms, we used loudness and lexical tones as a research model and employed a novel multilevel oddball paradigm with Electroencephalogram (EEG) recordings to explore the dynamics of mismatch negativity (MMN) responses to their deviants. When two types of deviants were presented separately, distinct topographies of MMNs to loudness and tones were observed at different latencies (loudness earlier), supporting the sequential dynamics of independent representations for two features. When they changed simultaneously, the latency of responses to tones became shorter and aligned with that to loudness, while the topographies remained independent, yielding the combined MMN as a linear additive of single MMNs of loudness and tones. These results suggest that neural dynamics can be temporally synchronized to distinct sensory features and balance the computational demands of segregation and integration, grounding for invariance and feature binding in serial processing.

Key words: mismatch negativity (MMN); loudness; lexical tones; feature binding; synchronization.

Introduction

"The sound must seem an echo to the sense:
Soft is the strain when Zephyr gently blows,
And the smooth stream in smoother numbers flows;
But when loud surges lash the sounding shore,
The hoarse, rough verse should like the torrent roar."—Alexander Pope.

The induction of any percepts depends on external signals whose intensity is above an agent's sensory threshold. The perception of intensity, for example loudness in the auditory domain, can be established independently from other features of the stimuli or various contexts (e.g. hearing "loud" regardless of whether being "surges" or "roar") (Zahorik and Wightman 2001). However, effective interaction with the environment requires stable perception despite variations in basic attributes such as signal intensity (e.g. understanding the word "poem" no matter whether one says it softly or loudly). That is, abstract representation must be established consistently and independently from the variation in the physical realization—achieving the invariance of perception, such as invariance to intensity, or invariance to different contexts (Liberman et al. 1967; Barbour 2011). On the other hand, basic attributes such as loudness need to combine with other features (such as pitch) to create a coherent percept, such as "gently blows" versus "torrent roar." That is, in addition to the segregation of features to achieve invariance, efficient perception necessitates *integration* across different levels of features (Treisman and Gelade 1980). How the seemingly contradicted computations—*segregation* for achieving invariance and *integration* for binding features and creating coherent percepts—are implemented simultaneously is unclear (Deco et al. 2015).

The sensation of intensity and frequency highlights the independent yet intertwined perceptual processes, making the loudness and pitch an optimal research model to investigate the mechanisms that balance segregation and integration. The percept of loudness is closely related to the intensity of a sound. The firing rate and the distribution of neurons are hypothesized to mediate the encoding of signal intensity, which is independent in distinct frequency channels in cochlear and combines nonlinearly at some later stages at the cortical level (Moore 1995, Schreiner and Malone 2015). Whereas pitch is a subjective percept related to the acoustic property of fundamental frequency (*f*0). The percept of pitch is arguably encoded nonlinearly in a seemingly tonotopic map in auditory cortices (Pantev et al. 1989; Bendor and Wang 2005). In the hierarchical processing of speech, invariance to background noise is more in nonprimary auditory areas after processes in the primary auditory cortex (Kell and McDermott 2019), suggesting that the abstract representation of speech is established in a specific spatial and temporal manner.

Received: April 24, 2023. **Revised:** June 3, 2023. **Accepted:** June 5, 2023
© The Author(s) 2023. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Downloaded from https://academic.oup.com/cercor/article/33/16/9542/7204411 by New York University Libraries user on 04 September 2023

The distinct ways of encoding and representations indicate the independence of loudness and pitch perception.

Despite the distinct ways of encoding intensity and frequency, the perceptions of loudness and pitch interact. For example, tracking loudness changes was influenced when sound frequency dynamically fluctuated, suggesting that the interaction of pitch and loudness occurs in the central auditory system (Neuhoff et al. 1999). Moreover, in Mandarin Chinese, loudness contour can have similar functions as pitch contour (c.f., lexical tone) that serves as a primitive prosodic cue to derive phonological and lexical-semantic information (Duanmu 2007). The loudness with pitch contour together can boost the success rate of Mandarin tone recognition in both simulated and actual cochlear implant hearing (Meng et al. 2017). Recent studies further demonstrate that the pitch process interacts with many acoustic attributes in the context of complex stimuli such as speech and music (McPherson and McDermott 2023).

The oddball paradigm can be an optimal protocol to investigate the mechanisms that balance the segregation and integration between multiple levels of features. In the oddball paradigm, a novel token infrequently occurs within a sequence of repetitive tokens. A prominent negative-going peak, termed mismatch negativity (MMN), is evoked by the low probability stimuli (deviant) compared with more frequently presented stimuli (standard) (Jääskeläinen et al. 2004; Näätänen et al. 2005). The fundamental presupposition of MMN is the establishment of deviants' neural representation that is different from the one of standards. Therefore, MMN is an effective neural measure for exploring neural representation (Näätänen 1995) and proven in the auditory domain like frequency (Jacobsen and Schröger 2001) and amplitude (Jacobsen et al. 2003). Moreover, the latency of MMN is the upper bound of establishing the representation of deviants. If the representation of loudness is first established before tones, the deviant in the loudness would induce MMN early than the deviant of tones. However, if the deviants in both features are available and integrated, the MMN would reflect the dynamics of integration. That is, by manipulating the inclusion of deviant feature(s), MMN can "temporally isolate" the timing of the establishment of representation, as well as reflect the timing of feature integration. Previous MMN studies mostly investigate multiple auditory features separately, for example, duration and intensity (Näätänen et al. 2004; Fisher et al. 2011), frequency, and phonological features (Pakarinen et al. 2007; Lovio et al. 2009; Honbolygó et al. 2017). Here, we expanded the oddball paradigm by including independent and simultaneous manipulations of auditory attributes of intensity and frequency and explored the mechanisms of separation and interaction between the establishment processes of multilevel neural representations.

The independent yet interactive nature of loudness and pitch perception yields a hypothesis that a mechanism that balances the segregation and integration between two neural representations should be available. In this study, EEG recordings were employed to investigate the neural dynamics of segregation and integration in responses to the changes of loudness, tones, as well as their combination in a multilevel oddball paradigm, in which we manipulated the sound type of the standards and deviants and yielded *Loudness Deviant (LD)*, *Tones Deviant (TD)*, and *Combined Deviant (CD)* conditions (Fig. 1A). Because of segregation for establishing independent abstract loudness and pitch representations, we predict that MMNs to *LD* and *TD* have independent neural sources, but the latency of MMN to *LD* is earlier than that to *TD* for manifesting the separation between neural representations of loudness and tones in the hierarchical processing of speech. When two features change simultaneously in the *CD* condition, three scenarios of interaction could occur (Fig. 1B). In the *linear combination* hypothesis, two single-feature MMNs do not interact but linearly combine across the activation duration, yielding the amplitude of the combined MMN as the sum of two single-feature MMNs and peak latency between that of two single-feature MMNs. In the *feature domination* hypothesis, one feature overwhelms another and makes the combined MMN have a similar response pattern and latency as the MMN of the dominant feature. In the *temporal-shift combination* hypothesis, the process of one feature shifts in time to temporally align with the process of another feature for facilitating integration (Treisman 1996). For example, in the visual field, a face is represented by neurons with small receptive fields in the early processing stages. Different assemblies of neurons respond to different local elements of the face, and the activity of these distributed populations of neurons in lower-level areas synchronizes temporally to construct the complete face representation (Rossion 2013). We propose that a similar temporal synchronization can occur in the auditory domain when multiple auditory features are present. Because of the faster processing speed of low-level features, the processing of high-level features could be facilitated. The temporally shifted responses combine with the processing of another feature, resulting in the combined MMN amplitude that linearly sums the two single-feature MMNs. This linear sum occurs at a latency consistent with the MMN of the nonshifted feature.

Materials and methods

Experimental model and subject details

Based on previous similar studies (Horváth et al. 2008) in which the power of MMNs that was observed in different sequence lengths, the target sample size is calculated ranging from 14 to 22 by using G*Power (Faul et al. 2009). Combine with the length of our auditory sequence, we take a moderate sample size of 20. So, 20 participants (four males; mean age: 22.56; range: 19–27) from East China Normal University were recruited for this study. Monetary incentives were provided for their participation. All participants had normal hearing without any neurological deficits (self-reported). Protocols were approved by the institutional review board at New York University Shanghai, which followed the Declaration of Helsinki as a statement of ethical principles concerning human testing. Written and informed consent was obtained for every participant.

Method details

Stimuli

Two humming sounds of Chinese lexical tones (first level tone and second rising tone) were used in our experiment, due to the perception of simple frequency manipulations, such as pure tone sinewaves are arguably not a categorization process, which is not optimal to address our research question of segregation. The lexical tone is a categorical perception (Si et al. 2017). Although the processing of pitch level and pitch contour, as two dimensions of lexical tones, can lateralize to different hemispheres (Wang et al. 2013), both dimensions cooccur and change simultaneously in Mandarin Chinese—they are processed as a whole and are perceived holistically (Lee 2000; Xu and Emily 2001; Wang et al. 2013). Therefore, we took advantage of the categorical and interactive features and used lexical tones as stimuli. They were recorded by a female speaker with a sampling frequency

Downloaded from https://academic.oup.com/cercor/article/33/16/9542/7204411 by New York University Libraries user on 04 September 2023

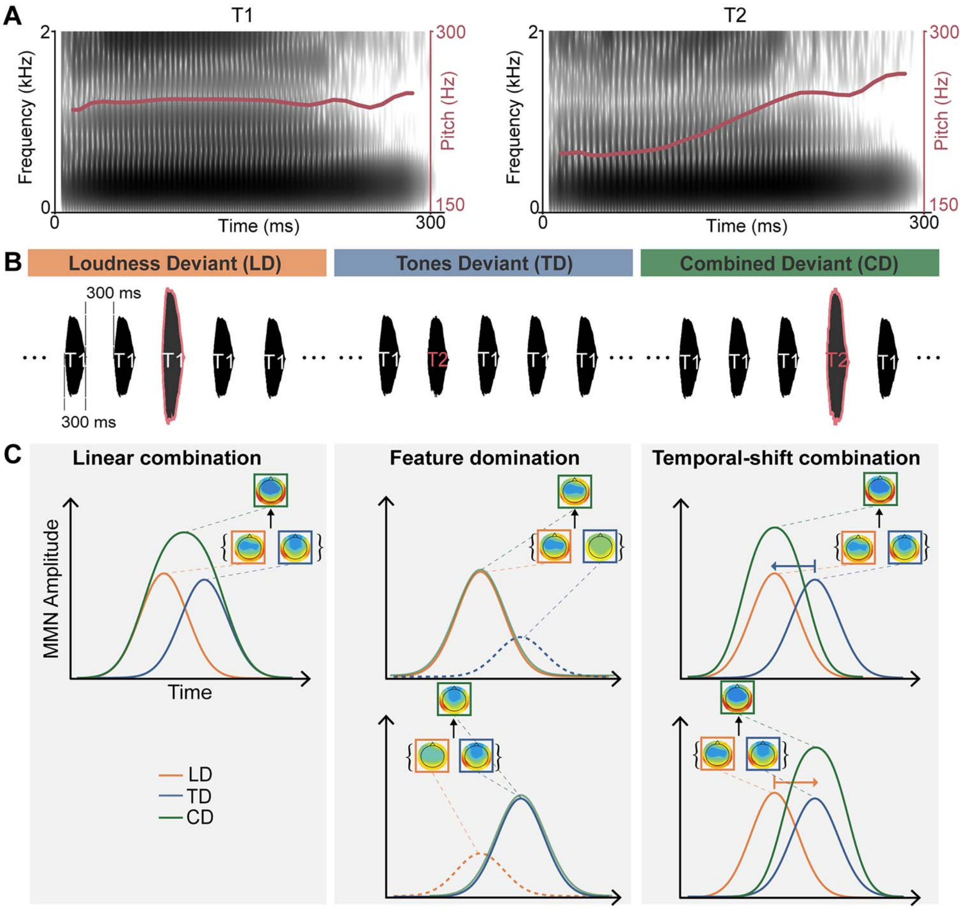

Fig. 1. Schematics of experimental procedures and hypotheses. (A) Experimental procedures of a multilevel oddball paradigm. Sequences of five auditory waveforms are depicted for illustration purposes. Each auditory stimulus has a duration of 300 ms with an ISI of 300 ms. The left plot is the LD condition in which the deviant sound has larger intensity (elongated amplitude highlighted with red contour) than the more frequently presented standard sounds. Lexical tone (T1, first level tone, as an example) is the same in all stimuli in this condition. The middle plot depicts the TD condition. The deviant stimulus (T2, second rising tone, denoted in red) differs from other standard stimuli, whereas the intensity of all stimuli is the same. The right plot is the CD condition where both the intensity and lexical tone of the deviant are different from the standards (see *Materials and methods* for details.) (B) Hypotheses regarding the potential interactive mechanism among multilevel deviant monitoring. Lines in each plot represent the hypothetical deviant neural responses (resembling the mismatch negative, MMN) for each condition. A presupposition is that processes of deviant monitoring for features at distinct levels are temporally isolated, as the peak latency for the basic attribute of *LD* (orange line) is earlier than that of tone deviant (blue line). Moreover, the neural sources of deviant monitoring for both levels are presumably independent, as reflected in distinct topographic patterns in noninvasive scalp electrophysiological recordings (hypothetical EEG topographic patterns are depicted for each condition). Three sets of hypotheses for possible interaction between multilevel deviant monitoring are available. First, the sources of LD and TD monitoring could remain independent both spatially and temporally, whereas the EEG measures linearly sum the activity from the monitoring at both levels when the two deviants are presented simultaneously (green line in the left panel). This linearly combination hypothesis predicts that the peak latency of the CD condition should be in the middle between the peak latencies of LD and TD conditions; the topography of the CD condition should be different from either single deviant condition. Second, the deviant monitoring in the CD condition could be dominated by one of the features (feature domination hypothesis, middle panel), where one feature deviant monitoring overwhelms (solid line) and makes the other absent (dashed line). The deviant monitoring responses in the CD condition would resemble the responses in the dominant feature condition. The upper plot illustrates the situation of loudness dominant and the lower plot for tone dominant. The responses to the CD are lifted for a better depiction. Third, the interaction could occur in the temporal domain as one of the monitoring processes shifts in time (temporal-shift combination hypothesis, right panel). That is, the processing dynamics is interactive—The process of deviant monitoring at one feature would alter the timing of the other. The upper plot demonstrates that the TD monitoring would be pulled forward in time toward the LD monitoring, whereas the lower plot illustrates the opposite (the corresponding color arrows indicate the direction of temporal shift). While the neural sources remain independent, the EEG measures linearly sum the activity from the monitoring at both levels after the temporal shift (green lines). This temporal-shift combination hypothesis predicts that the peak latency of the CD condition should be consistent with the feature that is not shifted; the topography of CD should also be different from either single deviant condition.

of 44.1 kHz and a duration of 300 ms (https://osf.io/mhgbj/files/osfstorage/Stimuli). And the intensity sound pressure level (SPL) of both sounds was adjusted to two levels—62 dB SPL (soft) and 81 dB SPL (loud) using Praat (Boersma and Weenink 2021), yielding four auditory stimuli—soft first tone (sT1), soft second tone (sT2), loud first tone (lT1), and loud second tone (lT2). Stimuli were presented binaurally via plastic air tubes connected to foam earplugs (ER-3C Insert Earphones; Etymotic Research).

Procedures

A passive listening oddball paradigm was implemented in this study. Participants were instructed to watch a silent landscape documentary film while sequences of lexical tones were presented with an inter-stimulus interval (ISI) of 300 ms (Fig. 1). The short ISI was determined by previous studies in which shorter ISI induced stronger MMN effects (Schröger and Winkler 1995; Escera et al. 2000). Participants were required to answer questions about the content of the film at the end of the experiment.

Downloaded from https://academic.oup.com/cercor/article/33/16/9542/7204411 by New York University Libraries user on 04 September 2023

Table 1. Information regarding the experimental design. The number of each stimulus is demonstrated in parentheses.

Conditions	Session 1	Session 2	Session 3	Session 4
Standard	sT1 (972)	lT1 (972)	sT2 (972)	lT2(972)
Loudness Deviant	lT1 (81)	sT1 (81)	lT2 (81)	sT2 (81)
Tones Deviant	sT2 (81)	lT2 (81)	sT1 (81)	lT1 (81)
Combined Deviant	lT2 (81)	sT2 (81)	lT1 (81)	sT1 (81)

Downloaded from https://academic.oup.com/cercor/article/33/16/9542/7204411 by New York University Libraries user on 04 September 2023

The ratio between standard and deviant stimuli was 80:20 in this experiment. To eliminate experimental stimuli bias, the experiment was divided into four sessions in which each type of four auditory sound was utilized as standard and the other three as deviants. The deviant stimuli were equally separated in three conditions: (i) *LD*, (ii) *TD*, and (iii) *CD* (Fig. 1). The number of trials in each condition was summarized in Table 1.

Taking Session 1 as an example, the soft first tone (sT1) was selected as standard (972 trials in total); the remaining three auditory stimuli were used as deviants (81 trials each, 243 trials in total). Specifically, one session was divided into three blocks. In each block, lT1, sT2, or lT2 was selected to be the *LD* (81 trials), *TD* (81 trials), or *CD* (81 trials) as compared to the standard of sT1 (324 trials). In each block, the deviants were presented in a pseudorandomized order in a way that deviants never follow each other and never appeared at the last place in the sequence. The entire experiment included 4,860 trials in total and lasted $\sim$50 min.

Quantification and statistical analysis

EEG recording and preprocessing

EEG signals were recorded in an electromagnetically insulated and sound-proof room using a 32-channel active electrode system (Brain Vision actiCHamp; Brain Products) with a 1,000 Hz sampling rate. Electrodes were mounted on an EasyCap that had electrode holders set according to the 10–20 international electrode system. Each electrode's impedance was kept below 10 kΩ. The data were online referenced to the electrode of Cz, and offline re-referenced to the average of all electrodes according to a recent paper that investigated the effect of EEG referencing methods on auditory MMN (Mahajan et al. 2017) demonstrating that using the average re-reference yielded similar results in data with different EEG electrode setups. Moreover, similar results were obtained by using different referencing procedures. Therefore, we employed the common average referencing. Two additional electrooculogram (EOG) electrodes (horizontal EOG, HEOG and vertical EOG, VEOG) were attached for monitoring ocular activity. The EEG data were collected using Brain Vision PyCoder software (http://www.brainvision.com/pycorder.html) and filtered online between DC and 200 Hz with a notch filter at 50 Hz.

Customized Python codes (all scripts can be seen at https://osf.io/mhgbj), MNE-python (Gramfort et al. 2014), Easy-EEG (Yang et al. 2018), and Topography-based Temporal-analysis (TTT) toolboxes (Wang et al. 2019) were used to process and analyze EEG data. The continuous EEG dataset was initially subject to a band-pass filter with cut-off frequencies set to 0.1 and 30 Hz, using finite impulse response (FIR) filtering parameters, zero-phase (two-pass forward and reverse) noncausal filter with Hamming window with 0.0194 passband ripple and 53 dB stopband attenuation, the filter length is 33,001 samples (33.001 s). The filtered data was divided into epochs spanning from -100 to 400 ms relative to the onset of the auditory stimuli. The 100 ms prestimulus period was used for baseline correction. Epochs containing artifacts related to eye blinks and head movement were manually rejected. A total of $\sim$27.93% of epochs were rejected on average across participants. To balance the power across conditions, a method in the MNE toolbox was applied to equalize the number of trials in different scenarios. This trial-number-equalization method selected epochs for each condition according to the condition that had the smallest number of remaining trials, while reducing time-varying noise by selecting epochs in different conditions that occurred close in time. A total of $\sim$47 trials (32.19% of the total epochs after noise rejection and trial-number equalization) for each type of stimuli were included in the following analyses.

EEG data analysis

Temporal cluster analysis on ERPs in one MMN representative channel

To test the validity of the data, we first followed the standard analysis of MMN by performing event-related potential (ERP) analysis on the data in one representative channel Fz to demonstrate the MMN effects (Näätänen et al. 2007). A temporal clustering permutation test (Maris and Oostenveld 2007) was subject to the ERP responses at the channel of Fz, separately for *LD*, *TD*, and *CD* conditions. Specifically, a paired t-test was applied to the deviant and standard ERPs at each time point. Temporal clusters were defined by successive time points (more than two adjacent time points) that exceeded the precluster threshold of 0.05. All the t-values within a temporal cluster were summed to obtain the summary empirical statistics for this temporal cluster. To obtain a null distribution, after shuffling the condition labels, similar paired t-tests were performed and the temporal cluster with maximum sum t-values was selected. The shuffling was repeated for 1,000 times and the selected maximum sum t-values in each shuffle formed a null distribution. Last, the summary empirical statistics of each temporal cluster calculated in original data was tested in the null distribution with a cluster-level threshold of 0.05.

Spatiotemporal cluster analysis

To further test the spatial distribution of the effect, we carried out a spatiotemporal clustering permutation test with data in all electrodes. The spatiotemporal clustering permutation test was similar to the temporal clustering permutation test, except the clusters were determined by successive time points that were significant at the precluster level of 0.05 in spatially adjacent electrodes. Two-tail paired t-test at each time point in each electrode was performed to take account of the polarity of ERP across response dynamics. A similar permutation procedure of shuffling condition labels was performed to form a null distribution. The significance of spatiotemporal clusters was tested by comparing the empirical summary statistics with the cluster-level threshold of 0.05 obtained in the null distribution. The spatiotemporal clustering permutation test was carried out separately for *LD*, *TD*, and *CD* conditions.

Latency analysis

The temporal variance among EEG channels makes the analysis of latency in single channels unreliable. Moreover, selecting channels may be subject to subjective bias and individual differences among participants (Tian and Huber 2008; Tian et al. 2011; Yang et al. 2018). Therefore, to test the hypothesis regarding response latency, we calculated the global field power (GFP), an omnibus measure that summarizes responses in all 32 EEG channels. The GFP is derived mathematically based on the standard deviation across all channels at each time point. It represents the sum of power from all channels varying across time, which minimizes the temporal variance among channels to derive a consistent measure of latency. Individual peak latencies of MMN components were identified in the difference waveforms obtained by subtracting GFP responses to standards from those of deviants, separately for *LD*, *TD*, and *CD* conditions. The latency identification was carried out semi-automatically using the TTT toolbox (Wang et al. 2019) in predetermined time windows based on the findings of prior permutation tests. The identified peaks were visually checked to verify that they were within the correct time ranges for each participant. Repeated measures Analysis of Variance (ANOVA) and post-hoc *t*-tests were subject to the MMN peak latencies among *LD*, *TD*, and *CD* conditions.

Topographic dissimilarity analysis

The topography of ERP responses represents the underlying neural sources. Pattern similarity/dissimilarity between topographies can assess the relations between underlying neural sources among conditions (Tian and Huber 2008; Tian et al. 2011). To test the hypothesis that the processing of *CD*s was the result of two independent deviant monitoring of loudness and tones, topographic ANOVA (TANOVA) analysis was carried out using EasyEEG (Yang et al. 2018). Specifically, a topography at each time point was a high-dimensional vector (32 dimensions in this experiment as 32 channels in EEG recordings). The angle (θ) between two vectors of topographies in two conditions represented the pattern similarity between them. The cosine value of the angle, cosθ, ranging from [−1, 1] was obtained, where 1 was a perfect match between two topographies and −1 was completely opposite patterns. To test the hypothesis that the neural sources of MMN in the *CD* condition were different from those in *LD* and *TD* conditions, a topographic dissimilarity index was derived by taking 1 − cosθ. Therefore, the topographic dissimilarity index ranged from [0, 2] where 2 was the maximum dissimilarity between two topographies. The topographic dissimilarity index was calculated between a pair of conditions in a time window of 5 ms, consecutively from −100 to 400 ms. To determine the statistical significance, a nonparametric permutation test was implemented. We first generated the null distribution of the topographic dissimilarity index. To keep the subject's information when permutating data, we put each participant' data into one pool regardless of experimental conditions, then shuffled the pool and randomly relabel the condition for each trial within each subject. After that, topographic dissimilarity indices between new group-averaged ERPs were calculated. And this permutation repeated 1,000 times to form a null distribution to correct multiple comparison problem for a set of consecutive significant time points. Finally, the significance of clusters was tested by comparing the empirical topographic dissimilarity indices with the cluster-level threshold of 0.05 obtained in the null distribution.

Simulation analysis

We further carried out a simulation using empirical data to directly test the *temporal-shift combination* hypothesis. The *temporal-shift combination* hypothesis assumes that the MMN in *CD* is a linear combination of MMN in *LD* and temporally advanced MMN in *TD* (Fig. 1B). Therefore, we compared the empirical MMN in *CD* with simulated MMN responses that were derived by adding a temporally shifted MMN in *TD* to MMN in *LD*. Specifically, a time window of 100 ms in duration was used to select data for simulation. First, the data from 100 ms to 200 ms were extracted from the individual MMN waveforms responses in all conditions. The epochs of *LD* and *TD* were summed at each corresponding time point and yielded a simulated epoch. The topographic similarity (cosθ) was calculated between the simulated epoch and the extracted epoch from empirical *CD* at each corresponding time point. The cosine values before and after 15 ms centered around the individual peak latency of empirical *CD* were averaged to index the topographic similarity between the simulated and empirical *CD*. The higher the index of topographic similarity represents more similar topographic patterns between the simulated *CD* and empirical *CD*.

Next, the 100 ms time window was moved left or right in a step of 1 ms to extract epochs in the MMN waveform responses only for the *TD* condition. The maximum moving range was 100 ms in either direction. That is, 200 epochs with each epoch of 100 ms in duration were extracted from *TD.* These epochs were [0100] ms to [200300] ms with an increment of 1 ms in the empirical *TD*. Each of the extracted epochs of *TD* was summed with the epoch of *LD* that was still extracted from 100 ms to 200 ms to yield one simulated epoch for each moving step. Noted that the simulated epochs yielded by the procedure of moving window and summation are equivalent to the results of moving the *TD* epoch in an opposite direction. For example, moving the time window to the *right* by 1 ms resulted in extracting the *TD* epoch of [101201] ms. The summation of this epoch with the *LD* epoch of [100200] ms is the same as if moving the *TD* epoch of [101201] ms to the *left* by 1 ms and then taking the summation. Therefore, for clarity, we define a parameter, Δt, to index the moving distance and direction of the *TD* epoch. The negative value of Δt represents the move of the extracted *TD* epoch to the left (e.g. moving the time window to the right to extract a later epoch).

After obtaining the 200 simulated epochs of *CD* by moving the responses of *TD*, the topographies in each simulated epoch were compared with the empirical *CD* epoch of [100200] ms, and the index of topographic similarity between the simulated and empirical *CD* was obtained for each moving distance. That is, a line of topographic similarity index as a function of Δt was obtained. To statistically determine the significant period of moving distance of *TD*, a bootstrap approach was employed to generalize a null distribution of topographic similarity and the significance threshold. Specifically, 1,000 samples were randomly sampled with replacement out of the 100 topographic similarity indices in the range of negative Δt values (*TD* moved to the left). The samples were averaged to form one data point in the distribution. Only the data in the range of negative Δt values were used for sampling because they more likely came from the same distribution (*TD* moved to the left as the MMN latency of *TD* lagged that of *LD* and *CD*) and hence more conservative and could efficiently reduce the type I error (topographic similarity indices in the range of negative Δt values were larger than those in the range

Downloaded from https://academic.oup.com/cercor/article/33/16/9542/7204411 by New York University Libraries user on 04 September 2023

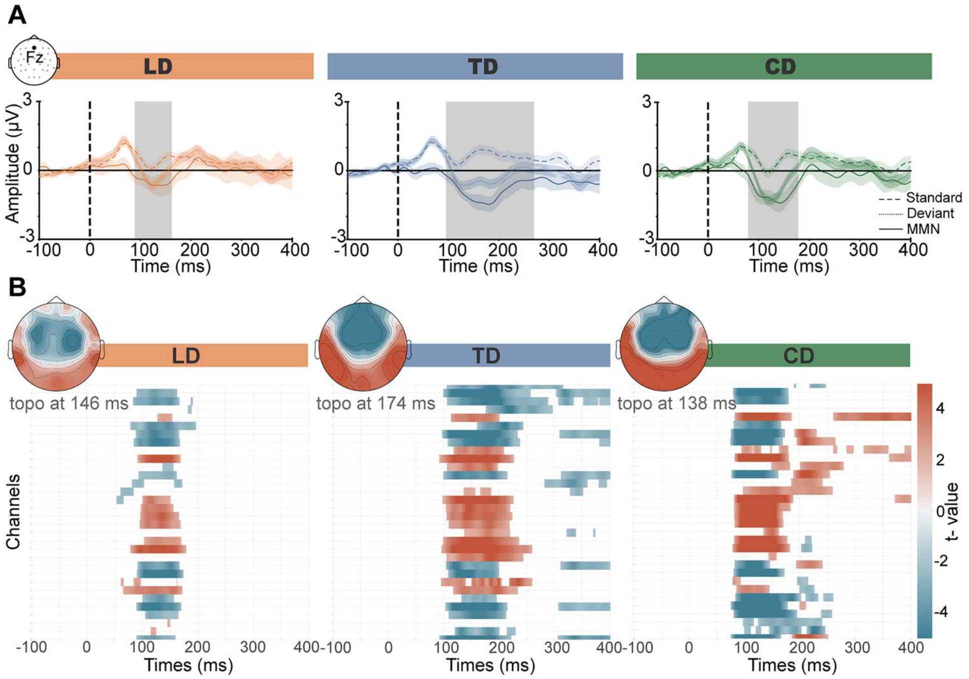

Fig. 2. (A) The ERP results of MMN in the representative channel of Fz. The solid and dashed lines in each plot represent the ERP responses to deviant and standard sounds in LD (left), TD (middle), and CD (right) conditions. The shaded areas indicate the significant period ($P < 0.05$, cluster level) in which the ERP amplitude of deviant sounds is more negative than that of standard sounds (MMN). The significant period in the TD condition is longer than that in LD and CD conditions, whereas the significant periods in LD and CD conditions are similar. (B) The spatiotemporal characteristics of MMN responses in all channels. Two significant clusters were observed in the spatiotemporal clustering permutation test in LD (left), TD (middle), and CD (right) conditions ($P < 0.05$, cluster corrected). The blue and red in each plot represent the time windows and spatial distributions of the two significant clusters in negative and positive directions of amplitude differences (deviant minus standard). A topographic display of the significant clusters at the centroid time point in each condition is illustrated at the top.

of positive Δt values). This bootstrap process was repeated 1,000 times. The value at the 95th percentile in the distribution was selected as the significance threshold. Finally, the middle point and its corresponding Δt in the period of significant points (points above the threshold) were identified to represent the temporal shift of responses to the tone deviant in the *CD* condition when the deviants of loudness and tones were available simultaneously.

Results

Distinct deviant monitoring of loudness and tones

To examine the temporal dynamics of MMN, we carried out a nonparametric temporal cluster analysis on the ERP responses to standard and deviant sounds in one representative channel of Fz, separately for *LD*, *TD,* and *CD* conditions. Significant differences between standard and deviant (MMN) were observed in all three conditions (Fig. 2A) and followed by the reporting specification (Sassenhagen and Draschkow 2019). The time window of MMN differed between *LD* condition (duration of 73 ms, the cluster in the observed data extended from 89 to 162 ms) and *TD* condition (duration of 175 ms, the cluster in the observed data extended from 95 to 270 ms), whereas the time window of MMN in *CD* condition (duration of 99 ms, the cluster in the observed data extended from 79 to 178 ms) was comparable to that of *LD* and shorter than that of *TD*. The MMN in *TD* lasted substantially longer (from 95 to 270 ms) than the MMN in *LD* (from 89 to 162 ms) and *CD* (from 79 to 178 ms) in the representative channel of Fz.

To quantify the spatial distribution of MMN in addition to its temporal characteristics, a nonparametric spatiotemporal cluster analysis was carried out. The results in Fig. 2B revealed distinct distributions of mismatch responses to deviants among three conditions (including positive and negative differences). Specifically, more channels were observed in the spatial cluster that showed significant MMN effects in *TD* than in *LD* (heatmaps in Fig. 2B). Moreover, the distribution in *LD* was bilateral, whereas the distribution in *TD* was more central (inserted topographies in Fig. 2B). The temporal window for the MMN spatial cluster was longer in *TD* than that in *LD*, which was consistent with the ERP results (Fig. 2A). The temporal and spatial differences in the MMN between *LD* and *TD* suggested that distinct processes of deviant monitoring for loudness and tones. In the following sessions, we further statistically tested the temporal and spatial differences between the MMN in *LD* and *TD*.

Moreover, in the *CD* condition, the distribution of channels in the spatial cluster was similar to that in *TD* (Fig. 2B). However, the temporal characteristics of MMN in *CD* were similar to that in *LD* (Fig. 2). The dissociation in temporal and spatial characteristics regarding how *CD* related to *LD* and *TD* indicated that the simultaneous deviant monitoring of two features was not a simple combination of two single deviant detection processes. In the following sessions, we quantitatively tested how the single deviant detection processes interact when deviants in multiple features occurred simultaneously.

Interactive process in simultaneous deviant monitoring of loudness and tone

To further investigate how the deviant monitoring operated when two features changed simultaneously, the dynamics of MMN responses in *LD*, *TD,* and *CD* conditions were compared. Specifically, a GFP waveform of MMN response was calculated in each condition (Fig. 3A) and the peak latency was identified based on the topographic patterns (see *Materials and methods* for details). According to different hypotheses (Fig. 1), the peak

Downloaded from https://academic.oup.com/cercor/article/33/16/9542/7204411 by New York University Libraries user on 04 September 2023

latency in the *CD* condition would show distinct relation to that in *LD* and *TD* conditions. Repeated measures one-way ANOVA on the peak latencies revealed a significant main effect of condition [$F(2, 38) = 27.1$, $P < 0.01$, $\eta_p^2 = 0.588$] (Fig. 3A, right inserted plot). Further pairwise *t*-test showed that the peak latency of MMN in *LD* was significantly earlier than that in *TD* [$t(19) = -5.042$, $P < 0.001$]. The peak latency in the *CD* condition was also significantly earlier than that in *TD* [$t(19) = -7.726$, $P < 0.001$]. However, the peak latencies in *LD* and *CD* were not different [$t(19) = -0.768$, $P = 0.452$]. The onset and offset latencies of the MMN component were not different between *LD* and *CD* either [for onset, $t(19) = 0.041$, $P = 0.968$; for offset, $t(19) = -0.373$, $P = 0.713$]; however, they are significantly different between *LD* and *TD* [for onset, $t(19) = -3.552$, $P = 0.002$; for offset, $t(19) = -5.503$, $P < 0.001$] and between *CD* and *TD* [for onset, $t(19) = -3.568$, $P = 0.002$; for offset, $t(19) = -4.946$, $P < 0.001$]. That is, in terms of processing dynamics, *LD* may dominate the deviant detection process even with the existence of lexical tones feature in *CD*. That peak latency of *TD* differed from that in *LD* and *CD* statistically supported the separation of monitoring processes of loudness and tone deviants, and the results of similar timing between *LD* and *CD* were consistent with the *feature domination* and *temporal-shift combination* hypotheses (Fig. 1).

To differentiate the *feature domination* and *temporal-shift combination* hypotheses, the response magnitude was further tested. Because of the similar latencies of MMN in *LD* and *CD* conditions (mean peak latency: 128 ms for *CD* and 133 ms for *LD*), a temporal clustering permutation test was applied to the GFP waveform responses of these two conditions. The analysis revealed a temporal window (from 91 to 146 ms) around their peak latencies that showed significant response magnitude differences between the MMNs of *LD* and *CD* conditions (highlighted in the shadowed rectangle area in Fig. 3A). A further component analysis using a paired *t*-test on individual MMN peak of GFPs showed that *CD* (mean: 1.470 μV) induced significantly a larger response magnitude than that in *LD* (mean: 1.318 μV) [$t(19) = 2.383$, $P = 0.028$]. These results were consistent with the *temporal-shift combination* hypotheses that predicted the combinatory nature of *CD*.

Independent neural source patterns when detecting combined feature changes

To further investigate how the deviant monitoring processes were combined, topographic analyses were used to investigate the neural dynamics of underlying sources. TANOVA revealed significantly different topographic patterns among MMNs in *LD*, *TD*, and *CD* conditions (Fig. 3B). For the conditions that only include single feature deviant, the topographic patterns of MMN responses in *TD* started to show significant differences from those in *LD* ~130 ms when the deviant responses of *TD* emerged. The pattern dissimilarity increased along the time approaching the processing peak of *LD* as shown in Fig. 3A. These results were consistent with the separation of deviant monitoring for *LD* and *TD* (Figs. 2 and 3A).

Moreover, the single feature deviant conditions (*LD* and *TD*) were compared with the *CD* condition to probe the possible dynamic changes when two levels of deviants occurred simultaneously. The topographic patterns of MMN in *CD* differed from MMN in *LD* starting ~110 ms and sustained to 210 ms. This starting time of dissimilarity was similar to the onset time of MMN in *TD* (101 ms), suggesting the topographic dissimilarity between *CD* and *LD* was likely due to the onset of the process for tone deviant that was available in CD. Whereas *CD* differed from *TD* starting as early as ~85 ms (similar to the onset time of MMN in *LD*, 82 ms) and suddenly became indistinguishable ~100 ms. These results suggested that the *LD* was first processed (*LD* was available in *CD* but not in *TD* and hence induced the response pattern differences between *CD* and *TD* at the beginning), then the *TD* process joined in (the additional tone deviant process in *CD* may minimize the response pattern differences between *CD* and *TD*). These results suggested that the deviant monitoring of loudness and tones remained independent when two types of deviants presented simultaneously—MMN of *CD* could be the combination of two independent MMNs (*LD* and *TD*), which yielded the results that topographic patterns in CD did not resemble either *LD* or *TD*.

Interestingly, the significant pattern dissimilarity between *CD* and *TD* emerged again after 175 ms (similar to the peak latency of *TD*). The significant dissimilarity between *CD* and *TD* after the latency of *TD* suggested that the processing of tone deviant could be shifted forward in time when presented simultaneously with *LD* in the *CD* condition—the deviant monitoring for tones started and ended earlier in the *CD* condition. This temporal shift of deviant monitoring for tones could yield different topographies in the later time window when deviant monitoring for tones was still processing in the *TD* condition but not in the *CD* condition. To quantitatively test the details of this temporal shift mechanism, we carried out a computational simulation using the empirical data in the next session.

Computational simulations to test the *temporal-shift combination* hypothesis for simultaneous monitoring of multilevel deviants

The *temporal-shift combination* hypothesis (Fig. 4A, selectively highlighted from Fig. 1) was supported by the observations of two independent monitoring processes that combined when deviants of multiple features were presented simultaneously (Figs. 2 and 3). The temporal distance between the MMN in *CD* and *TD* (Fig. 4B, selectively highlighted from Fig. 3A) was qualitatively analogs to the temporal shift in the hypothesis (Fig. 4A). To quantitatively test the hypothesis as well as to probe the details of the temporal shift mechanism, a computational simulation was carried out to reproduce the observed neural responses in *CD* by temporally manipulating and combining neural responses in *LD* and *TD*. Specifically, the time series of MMN responses in *TD* was temporally shifted by a parameter of Δt and linearly added to the MMN responses in *LD* to create a time series of *simulated CD* (Fig. 4C, see *Materials and methods* for details). The topographic similarity between the *simulated CD* and *empirical CD* as a function of temporal shift Δt was obtained (Fig. 4D).

The *simulated CD* was significantly similar to the *empirical CD* in the range of Δt from −70 to 10 ms (negative Δt values mean moving *TD* forward in time). The middle point of the significant period was −30 ms, which was not different from the peak latency difference between the MMN responses in *LD* and *TD* (mean peak latency difference of −39.5 ms) [$t(19) = -1.40$, $P = 0.178$]. The simulation results were consistent with the *temporal-shift combination* hypothesis and suggested that the dynamics of deviant monitoring for tones was accelerated and temporally aligned with the process of loudness when the deviant monitoring for loudness co-occurred.

Discussion

Using a novel multifeature oddball paradigm, we found complicated interaction between the neural processes of two sensory features. The EEG results showed MMN latency differences (Figs. 2 and 3A) and topographic differences (Fig. 3B) between

Downloaded from https://academic.oup.com/cercor/article/33/16/9542/7204411 by New York University Libraries user on 04 September 2023

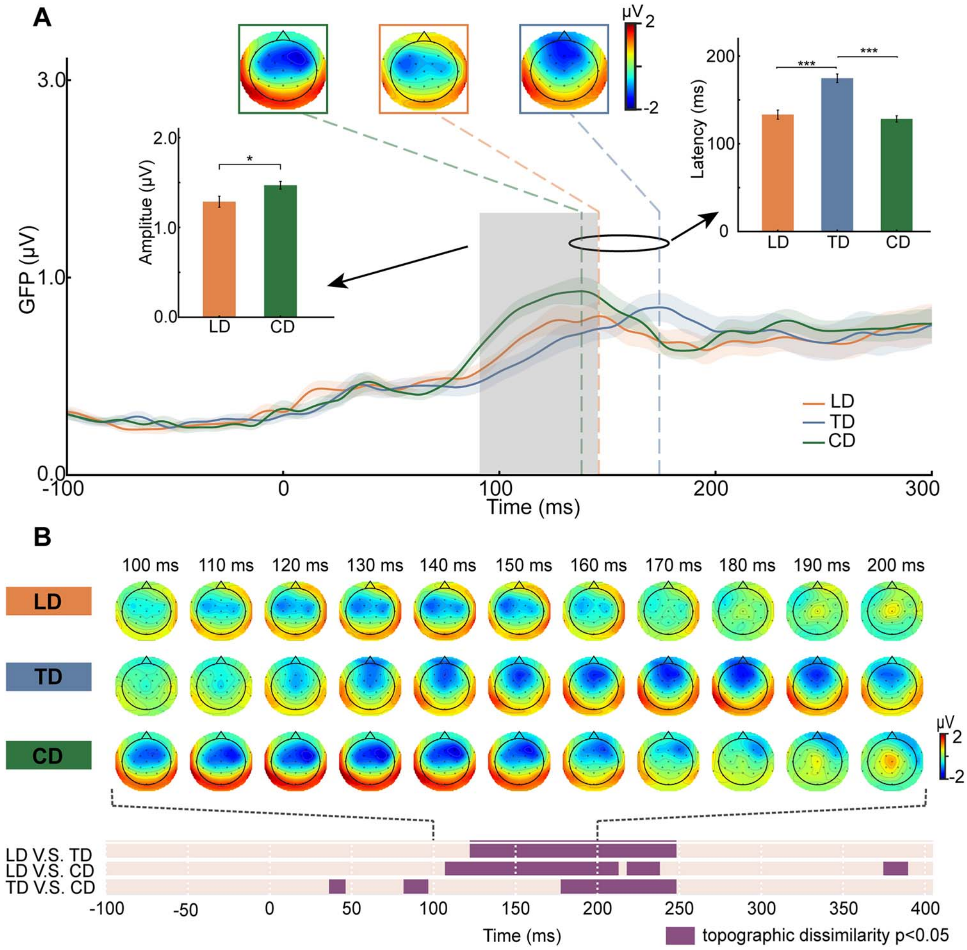

Fig. 3. (A) GFP results of MMN latency and amplitude among three conditions. The global measures of response dynamics of MMN are represented as the GFP waveforms, separately for each condition in different colors. The topographies of MMN responses are depicted in corresponding color boxes, while their latencies are indicated by dashed lines. The inserted bar plot on the right shows that the peak latency in TD is significantly later than those in LD and CD. The shadowed rectangle area across the GFP waveforms indicates the significant temporal cluster that shows amplitude differences between LD and CD. The left-inserted plot shows the results of component analysis of MMN amplitude differences that are consistent with the temporal cluster results. * indicates $P < 0.05$, *** indicates that $P < 0.001$. (B) Results of topographic dissimilarity among three conditions. The index of topographic dissimilarity was computed in each paired comparison between the two conditions. The heatmap at the bottom shows the significant results in the topographic dissimilarity analysis using the permutation TANOVA test (see *Materials and methods* for details). The significant clusters are indicated by purple (cluster-level threshold of 0.05). Topographic patterns start to be significantly different at ∼125 ms between LD and TD; whereas the topographic pattern differences emerge earlier in the comparisons between LD and CD (∼110 ms), as well as a cluster of dissimilarity patterns between TD and CD (∼85–95 ms). The dissimilarity between TD and CD becomes nonsignificant at a period approximately from 100 ms to 175 ms, followed by the significant dissimilarity reappearing again starting at ∼ 180 ms. On the top, grand averaged topographies from 100 ms to 200 ms are depicted to visualize the evolution and pattern differences among the three conditions.

deviant detection of loudness and tones, suggesting independent and sequential responses that mediate separate analyses of each feature in arguably hierarchical processing to achieve loudness invariance and abstract representation of lexical tones. Whereas complex interaction occurred when multiple acoustic features change simultaneously. The latency of tones was temporally shifted forward and aligned with the earlier process of loudness (Figs. 3A and 4). The neural sources of the two processes remained independent (Fig. 3B) and linearly combined, manifested in the response magnitude of combined topographies (Figs. 3A and 4). These consistent results support the *temporal-shift combination* hypothesis and suggest that temporally shifting and aligning independent neural representations are potential mechanisms for balancing the computational demands between achieving invariance and feature integration.

We observed independent neural sources operating at distinct latencies that mediate deviant detection of loudness and tones (Figs. 2 and 3B). These results suggest that although the featural attributes of intensity and frequency are available in the acoustic signals and presumably analyzed in the cochlea, the representation and hence the perception of loudness and pitch that associate with the two attributes are established independently in a serial manner. We found that the latency of LD was earlier than that of tone deviant (Fig. 3A). These results are consistent with previous findings that demonstrate the loudness perception can be established as early as in the subcortical auditory pathway (Sun et al. 2021). Whereas the neural representation of pitch mostly emerges at the auditory cortical level (Bendor and Wang 2005), consistent with our observations of later latency of tone deviant.

The sequential manner of separately establishing loudness and tones representation potentially offers a conceivable neural procedure to achieve loudness invariance. The early loudness process presumably establishes the loudness perception in an independent neural representation and removes the information

Downloaded from https://academic.oup.com/cercor/article/33/16/9542/7204411 by New York University Libraries user on 04 September 2023

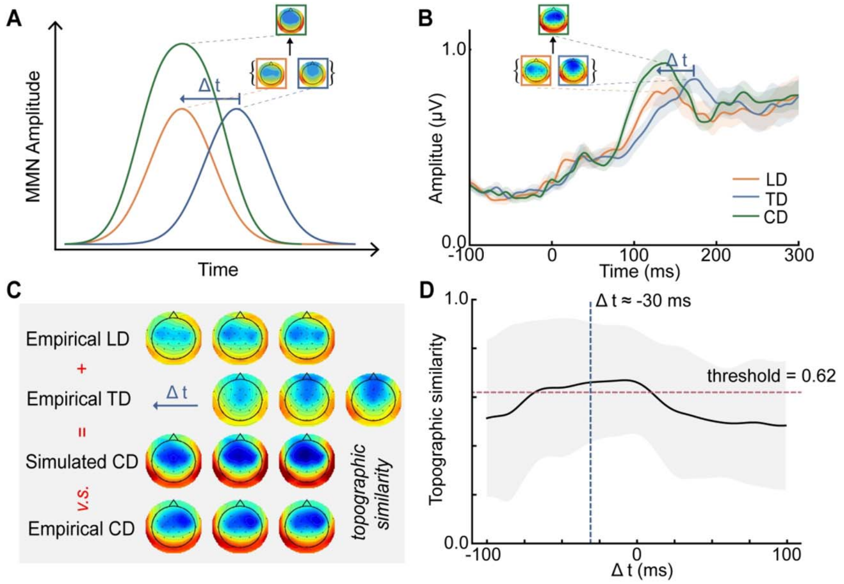

Fig. 4. Results of computational simulations using empirical data on CD monitoring. (A) The temporal-shift hypothesis, drawn and highlighted from Fig. 1, indicates a temporal-shifted linear combination process when loudness and lexical tones change simultaneously. (B) The empirical data from the EEG experiment, are qualitatively similar to the hypothesis in (A). Arrows in (A) and (B) indicate the moving direction of lexical tones MMN, with the parameter Δt referring to the shifted time. (C) Demonstration of the computational simulation. The empirical LD responses were linearly added to the temporally shifted (stepwise with the parameter Δt) empirical TD responses, yielding the simulated CD responses. The simulated CD responses were compared with the empirical CD responses. Topographic similarity measures were obtained between the empirical and simulated CD responses at each step of Δt. D) the results of topographic similarity between simulated and empirical CD responses. The parameter Δt on the x-axis indicates the moving direction and amount of the empirical TD responses when creating the simulated CD, with negative indicating moving forward in time. The topographic similarity exceeds the threshold (red dashed line) from −70 ms to 10 ms, suggesting that the topographies of simulated and empirical CD are significantly similar. The blue dashed line indicates the middle point of the significant period, approximating the shifted time of TD when detecting the combined loudness and lexical tones feature in CD.

about intensity in later processes for creating abstract representations of lexical tones. This sequential procedure could be a general mechanism for achieving invariance in hierarchical processing—the lower basic features induce independent representation in an early stage, whereas the variance in the lower features can be normalized and hence abstract away, only leaving the essential information for establishing abstract representations of higher-level perception. That is, the invariance can be achieved in a serial process via independent representations in distinct processing latencies.

The independent neural representations with distinct latencies in serial processing increase the computational demands when integrating multiple features for establishing a coherent percept. This computational demand can potentially be fulfilled by temporally shifting and aligning the two independent representations. We observed that when the deviants of loudness and tones occurred simultaneously, two neural representations remained independent yet the processing time changed (Fig. 3B). Consistent with the *temporal-shift combination* hypothesis, the responses to *TD* were shifted forward in time to a similar latency as the responses to *LD* in the CD condition when two types of deviants occurred simultaneously (Figs. 3 and 4). That is, when two features changed together, potential interactions facilitate the processing of lexical tones and make it temporally align with the early processing of loudness for integration across these two features.

This temporal-shifting mechanism is consistent with the feature synchronization of binding theory (Treisman 1996). To solve the binding problem, mechanisms of synchronized firing were proposed, in which the to-be-integration neural representations are firing in synchrony to label and integrate the features for a coherent percept. For example, four segments of contour lines in a rectangle shape would activate different orientation neurons in corresponding receptive fields; when the four neural ensembles fire in synchrony, the percept of a rectangle shape can be established. Our results are consistent with the synchronization for binding, indicated by the temporally shifting and alignment of processes of two features. Unlike in vision where basic features are processed in parallel, auditory features are processed mostly in serial. The synchronous firing for integration necessitates the alignment in time, presumably as manifested in the temporal shifting observed in our study.

The observed temporal shifting for integration can occur automatically without attention or feature-related task demands. As the current oddball procedure implemented a task that was irrelevant to the auditory features and deviants—participants were required to watch a silent movie and to answer questions about what they saw, the temporal-shifting effects were observed for the auditory features.

The main findings of multisensory integration studies indicate two distinct manifestations of multimodal integrative responses. The integrated responses are commonly larger than the sum of unimodal responses—a property known as "super-additivity." For example, such supra-additivity effects were found in audio–visual speech compared to unimodal speech (Calvert et al. 2000), suggesting a possible integrated representation. However, sub-additivity effects—the integrated responses are smaller than the sum of unimodal responses—were also observed in multisensory integration (Gu et al. 2008), suggesting that integrating between different modalities could reduce variance (Knill and Pouget 2004). Regardless of whether the effects are supra- or sub-additivity in multisensory integration, observations of temporal facilitation are commonly observed. For example, recent multisensory

Downloaded from https://academic.oup.com/cercor/article/33/16/9542/7204411 by New York University Libraries user on 04 September 2023

integration studies suggest that somatosensory processes can be automatically speeded up by the co-occurrence of visual processes (Zheng et al. 2021). Synchronized firing across separate but interconnected cortical areas supports feature integration (Treisman 1996), while the neural processing for different levels of features can influence such synchronization, causing temporal shifts. In the speech domain, the synchronous presentation of multimodal stimuli can induce temporal facilitation in which visual speech speeds up the processing of auditory speech (van Wassenhove et al. 2005). The temporal facilitation in multisensory integration is consistent with our results of temporal shifting in the tonal process for aligning with loudness responses, suggesting that the synchrony of two independent sources may be a ubiquitous manipulation for integrating features either across sensory modalities or across hierarchies within a single modality. The observed linear addition is the manifestation of temporally aligned loudness and tone responses that are mediated by two separate neural sources, which is the initiation of integration followed by the results of integration in a later processing stage presumably in associate areas where supra- or sub-additivity effects could occur.

The underlying neural mechanisms for the observed temporal shifting could be an adjustment in the time constant in neural processing. Our results show that the advanced peak latency of *TD* is likely a consequence of faster accumulation (Fig. 3B)—the duration of raising from onset to peak was shorted in CD than that in TD. According to the predictive coding model (Bastos et al. 2012), deviant detection is manifested in an error term that is the difference between prediction and neural processes of stimuli. Considering the assumption that the error only needs to compare between the bottom-up and top-down neural signals, the changes in processing timing more likely exist in the layer of feedforward processes of sensory stimuli.

A recent study did not find significant differences between response latencies in Herschel's gyrus (c.f. primary auditory cortex) and superior temporal gyrus (c.f. secondary auditory cortex) when listening to speech (Hamilton et al. 2021). This absence of effects led the authors to speculate a parallel processing account in speech perception, contrasting with the classical view of serial processing. In the current study, we observed both the onset and peak latency differences between the *LD* and *TD* conditions. These significant latency differences suggest serial processing for multiple auditory features. More importantly, the co-occurrence of deviants in multiple features temporally shifted the response latencies, yielding the latencies for processing two features indistinguishable. And hence, the absence of response latency differences across auditory neural hierarchy (Hamilton et al. 2021) could due to the automatic alignment for integrating auditory features when processing speech. Therefore, our results are consistent with the classic view of serial processing in audition rather than the alternative of parallel processing.

The current study did not design gender as an independent variable. Previous MMN studies did not find the gender effects on the MMN amplitude in auditory and visual domains (e.g. Yang et al. 2016), nor on the peak latency of the MMN (Barrett and Fulfs 1998). Whether gender differences exist in the observed temporal shifts in multilevel features MMN is subject to further investigation.

Methodologically, our novel multifeature oddball paradigm advances the investigation of MMNs. Classical auditory MMN studies mainly concentrated on a single basic feature such as frequency, intensity, and locations (Näätänen et al. 2004, 2007), and single more complex and abstract features such as phonemes and syllables in alphabetic languages (Aaltonen et al. 1987; Sharma and Dorman 2000; Sussman et al. 2004) and lexical tones in Chinese (Si et al. 2017). Although some studies included multiple features, such as frequency and intensity changes in an auditory sequence (Gomes et al. 1997; Ruusuvirta et al. 2003), one feature change was still presented at a time. In this study, we simultaneously changed multiple features at one instance. The combinatory featural changes in a novel multifeature oddball paradigm facilitate the investigation of deviant monitoring from a new interactive perspective.

More generally, our results of different neural sources and latencies in MMN responses to deviants of distinct features hint that the deviant detection on different levels of features could be mediated by a canonical neural computation with similar neural computational structures (Bastos et al. 2012). Our results show the earliest latency for *LD* and a later latency for tone deviant. The list of latency changes added when the deviants on more abstract representation, such as the deviant of phonetic features for phoneme category with an MMN latency of 170 ms (Phillips et al. 2000), an N400 component for the semantic anomaly (Chwilla et al. 1995), a P600 component reflecting syntactic integration processes (Kaan et al. 2000), and even later error-related negativity when participants were aware of making an erroneous judgment (Shalgi et al. 2009). These progressive latencies in the detection of novelty or anomaly are associated with the timing of establishing the particular level of representation. The common point is the detection of changes in a given representation, but the difference is the level of analysis on which the computation of detection applies. These progressive latencies in distinct neural sources for detecting deviants at distinct levels of representation hint at a canonical neural computational structure that mediates the establishment of representation, receiving predictions, and comparing between the two and generating errors (Bastos et al. 2012).

The fact that the response to the TD moves forward to synchronize with LD in CD condition provides important evidence suggesting that the sequential dynamic processes can be modulated by the flexible demands to fulfill featural segregation and integration. Together, we can further explore the neural hierarchy of sequential processing when different levels of speech properties pour into the brain.

In sum, we implemented a novel multifeature oddball paradigm and found that the neural responses to deviants in the loudness and tone dimensions had distinct neural sources that activated at different latencies. Moreover, the responses to the tone deviant advanced in time and synchronized with the responses to the *LD* when two features altered simultaneously. This temporal facilitation and synchrony of independent neural representations are potential mechanisms for balancing the computational demands between achieving invariance and feature integration.

Downloaded from https://academic.oup.com/cercor/article/33/16/9542/7204411 by New York University Libraries user on 04 September 2023

Acknowledgments

We thank Siqi Li for her help in data collection, Xikang Luo for his help in data analysis.

Author contributions

Zhili Han (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing—original draft, Writing—review & editing), Hao Zhu (Formal analysis, Methodology, Software, Validation),

Yunyun Shen (Data curation, Resources, Validation, Visualization), and Xing Tian (Conceptualization, Funding acquisition, Project administration, Supervision, Writing—review & editing).

Funding

This study was supported by the National Natural Science Foundation of China (32071099 and 32271101); Natural Science Foundation of Shanghai (20ZR1472100); Program of Introducing Talents of Discipline to Universities (Base B16018) and NYU Shanghai Boost Fund.

Competing interests: The authors declare no competing interests.

Code and data availability

All data (including raw data, preprocessed data, event files, and stimuli) and codes are available at https://osf.io/mhgbj/.

References

Aaltonen O, Niemi P, Nyrke T, Tuhkanen M. Event-related brain potentials and the perception of a phonetic continuum. *Biol Psychol*. 1987:24(3):197–207.

Barbour DL. Intensity-invariant coding in the auditory system. *Neurosci Biobehav Rev*. 2011:35(10):2064–2072.

Barrett KA, Fulfs JM. Effect of gender on the mismatch negativity auditory evoked potential. *J Am Acad Audiol*. 1998:9(6):444–451.

Bastos AM, Usrey WM, Adams RA, Mangun GR, Fries P, Friston KJ. Canonical microcircuits for predictive coding. *Neuron*. 2012:76(4): 695–711.

Bendor D, Wang X. The neuronal representation of pitch in primate auditory cortex. *Nature*. 2005:436(7054):1161–1165 Nature Publishing Group.

Boersma P, Weenink D. Praat: doing phonetics by computer (Version 6.1.40) [Computer program: http://www.praat.org/]. 2021.

Calvert GA, Campbell R, Brammer MJ. Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Curr Biol*. 2000:10(11):649–657.

Chwilla DJ, Brown CM, Hagoort P. The N400 as a function of the level of processing. *Psychophysiology*. 1995:32(3):274–285.

Deco G, Tononi G, Boly M, Kringelbach ML. Rethinking segregation and integration: contributions of whole-brain modelling. *Nat Rev Neurosci*. 2015:16(7):430–439 Nature Publishing Group.

Duanmu S. *The phonology of standard Chinese*. 2nd ed. Oxford; New York: Oxford University Press; 2007

Escera C, Yago E, Polo MD, Grau C. The individual replicability of mismatch negativity at short and long inter-stimulus intervals. *Clin Neurophysiol*. 2000:111(3):546–551.

Faul F, Erdfelder E, Buchner A, Lang A-G. Statistical power analyses using G*power 3.1: tests for correlation and regression analyses. *Behav Res Methods*. 2009:41(4):1149–1160.

Fisher DJ, Grant B, Smith DM, Knott VJ. Effects of deviant probability on the 'optimal' multi-feature mismatch negativity (MMN) paradigm. *Int J Psychophysiol*. 2011:79(2):311–315.

Gomes H, Bernstein R, Ritter W Jr, Vaughanm HG, Miller J. Storage of feature conjunction in transient auditor memory. *Psychophysiology*. 1997:34(6):712–716.

Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, Parkkonen L, Hämäläinen MS. MNE software for processing MEG and EEG data. *NeuroImage*. 2014:86:446–460.

Gu Y, Angelaki DE, DeAngelis GC. Neural correlates of multisensory cue integration in macaque MSTd. *Nat Neurosci*. 2008:11(10): 1201–1210 Nature Publishing Group.

Hamilton LS, Oganian Y, Hall J, Chang EF. Parallel and distributed encoding of speech across human auditory cortex. *Cell*. 2021:184(18):4626–4639.e13.

Honbolygó F, Kolozsvári O, Csépe V. Processing of word stress related acoustic information: a multi-feature MMN study. *Int J Psychophysiol*. 2017:118:9–17.

Horváth J, Czigler I, Jacobsen T, Maess B, Schröger E, Winkler I. MMN or no MMN: no magnitude of deviance effect on the MMN amplitude. *Psychophysiology*. 2008:45(1):60–69.

Jääskeläinen IP, Ahveninen J, Bonmassar G, Dale AM, Ilmoniemi RJ, Levänen S, Lin F-H, May P, Melcher J, Stufflebeam S, et al. Human posterior auditory cortex gates novel sounds to consciousness. Proceedings of the National Academy of Sciences. *National Acad Sci*. 2004:101(17):6809–6814.

Jacobsen T, Schröger E. Is there pre-attentive memory-based comparison of pitch? Psychophysiology. 2001:38:723–727.

Jacobsen T, Horenkamp T, Schröger E. Preattentive memory-based comparison of sound intensity. *Audiol Neurotol*. 2003:8(6):338–346.

Kaan E, Harris A, Gibson E, Holcomb P. The P600 as an index of syntactic integration difficulty. *Lang Cogn Processes*. 2000:15(2): 159–201 Routledge.

Kell AJE, McDermott JH. Invariance to background noise as a signature of non-primary auditory cortex. *Nat Commun*. 2019:10(1):3958 Nature Publishing Group.

Knill DC, Pouget A. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci*. 2004:27(12): 712–719.

Lee C-Y. Lexical tone in spoken word recognition: A view from Mandarin Chinese. *J Acoust Soc Am*. 2000:108(5_Supplement):2480.

Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M. Perception of the speech code. *Psychol Rev*. 1967:74(6): 431.

Lovio R, Pakarinen S, Huotilainen M, Alku P, Silvennoinen S, Näätänen R, Kujala T. Auditory discrimination profiles of speech sound changes in 6-year-old children as determined with the multi-feature MMN paradigm. *Clin Neurophysiol*. 2009:120(5): 916–921.

Mahajan Y, Peter V, Sharma M. Effect of EEG referencing methods on auditory mismatch negativity. *Front Neurosci*. 2017:11:560.

Maris E, Oostenveld R. Nonparametric statistical testing of EEG- and MEG-data. *J Neurosci Methods*. 2007:164(1):177–190.

McPherson MJ, McDermott JH. Relative pitch representations and invariance to timbre. *Cognition*. 2023:232:105327.

Meng Q, Zheng N, Li X. Loudness contour can influence mandarin tone recognition: vocoder simulation and Cochlear implants. *IEEE Trans Neural Syst Rehabilitation Eng*. 2017:25(6):641–649.

Moore BCJ. *Hearing*. San Diego: Academic Press; 1995

Näätänen R. The mismatch negativity: a powerful tool for cognitive neuroscience. *Ear Hearing*. 1995:16(1):6–18.

Näätänen R, Pakarinen S, Rinne T, Takegata R. The mismatch negativity (MMN): towards the optimal paradigm. *Clin Neurophysiol*. 2004:115(1):140–144.

Näätänen R, Jacobsen T, Winkler I. Memory-based or afferent processes in mismatch negativity (MMN): a review of the evidence. *Psychophysiology*. 2005:42(1):25–32.

Näätänen R, Paavilainen P, Rinne T, Alho K. The mismatch negativity (MMN) in basic research of central auditory processing: a review. *Clin Neurophysiol*. 2007:118(12):2544–2590 Elsevier.

Neuhoff JG, McBeath MK, Wanzie WC. Dynamic frequency change influences loudness perception: a central, analytic process. *J*

Downloaded from https://academic.oup.com/cercor/article/33/16/9542/7204411 by New York University Libraries user on 04 September 2023

Experimental Psychol Human Percep Performance. 1999:25(4):1050 American Psychological Association.

Pakarinen S, Takegata R, Rinne T, Huotilainen M, Näätänen R. Measurement of extensive auditory discrimination profiles using the mismatch negativity (MMN) of the auditory event-related potential (ERP). *Clin Neurophysiol*. 2007:118(1): 177–185.

Pantev C, Hoke M, Lütkenhöner B, Lehnertz K. Tonotopic Organization of the Auditory Cortex: pitch versus frequency representation. *Science*. 1989:246(4929):486–488 American Association for the Advancement of Science.

Phillips C, Pellathy T, Marantz A, Yellin E, Wexler K, Poeppel D, McGinnis M, Roberts T. Auditory cortex accesses phonological categories: an MEG mismatch study. *J Cogn Neurosci*. 2000:12(6): 1038–1055.

Rossion B. The composite face illusion: a whole window into our understanding of holistic face perception. *Vis Cognition*. 2013:21(2):139–253 Routledge.

Ruusuvirta T, Huotilainen M, Fellman V, Näätänen R. The newborn human brain binds sound features together. *Neuroreport*. 2003:14(16):2117–2119.

Sassenhagen J, Draschkow D. Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location. *Psychophysiology*. 2019:56(6):e13335.

Schreiner CE, Malone BJ. Representation of loudness in the auditory cortex. In: *Handbook of clinical neurology*. Elsevier; 2015. pp. 73–84

Schröger E, Winkler I. Presentation rate and magnitude of stimulus deviance effects on human pre-attentive change detection. *Neurosci Lett*. 1995:193(3):185–188.

Shalgi S, Barkan I, Deouell LY. On the positive side of error processing: error-awareness positivity revisited. *Eur J Neurosci*. 2009:29(7): 1522–1532.

Sharma A, Dorman MF. Neurophysiologic correlates of cross-language phonetic perception. *J Acoust Soc Am*. 2000:107(5): 2697–2703 Acoustical Society of America.

Si X, Zhou W, Hong B. Cooperative cortical network for categorical processing of Chinese lexical tone. *Proc Natl Acad Sci USA*. 2017:114(46):12303–12308.

Sun J, Wang Z, Tian X. Manual gestures modulate early neural responses in loudness perception. *Front Neurosci*. 2021:15:634967.

Sussman E, Kujala T, Halmetoja J, Lyytinen H, Alku P, Näätänen R. Automatic and controlled processing of acoustic and phonetic contrasts. *Hear Res*. 2004:190(1):128–140.

Tian X, Huber DE. Measures of spatial similarity and response magnitude in MEG and Scalp EEG. *Brain Topogr*. 2008:20(3):131–141.

Tian X, Poeppel D, Huber DE. TopoToolbox: using sensor topography to calculate psychologically meaningful measures from event-related EEG/MEG. *Comput Intel Neurosci*. 2011:2011:1–8.

Treisman A. The binding problem. *Curr Opin Neurobiol*. 1996:6(2): 171–178.

Treisman A, Gelade G. A feature-integration theory of attention. *Cogn Psychol*. 1980:12(1):97–136.

Wang X-D, Wang M, Chen L. Hemispheric lateralization for early auditory processing of lexical tones: dependence on pitch level and pitch contour. *Neuropsychologia*. 2013:51(11): 2238–2244.

Wang X, Zhu H, Tian X. Revealing the temporal dynamics in non-invasive electrophysiological recordings with topography-based analyses. *bioRxiv*. 2019:779546.

van Wassenhove V, Grant KW, Poeppel D. Visual speech speeds up the neural processing of auditory speech. *Proc Nat Acad Sci*. 2005:102(4):1181–1186.

Xu Y, Emily WQ. Pitch targets and their realization: evidence from mandarin Chinese. *Speech Comm*. 2001:33(4):319–337.

Yang X, Yu Y, Chen L, Sun H, Qiao Z, Qiu X, Zhang C, Wang L, Zhu X, He J, et al. Gender differences in pre-attentive change detection for visual but not auditory stimuli. *Clin Neurophysiol*. 2016:127(1): 431–441.

Yang J, Zhu H, Tian X. Group-level multivariate analysis in EasyEEG toolbox: examining the temporal dynamics using topographic responses. *Front Neurosci*. 2018:12:468. *Frontiers*.

Zahorik P, Wightman FL. Loudness constancy with varying sound source distance. *Nat Neurosci*. 2001:4(1):78–83 Nature Publishing Group.

Zheng Q, Zhou L, Gu Y. Temporal synchrony effects of optic flow and vestibular inputs on multisensory heading perception. *Cell Rep*. 2021:37(7):109999.

Downloaded from https://academic.oup.com/cercor/article/33/16/9542/7204411 by New York University Libraries user on 04 September 2023