UNIVERSITY OF CALIFORNIA,
IRVINE

The Origins and Mechanisms of Auditory and Motor Rhythm Entrainment: Behavioral, Neuroimaging, and Psychophysical Investigations

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Cognitive Sciences

by

Jie Wan

Dissertation Committee:
Professor Gregory Hickok, Chair
Professor Kourosh Saberi
Assistant Professor Connor Mayer

2026

DEDICATION

To myself, for making it this far.

And to my love, Yo Oizumi, for making the road ahead worth taking.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Gregory Hickok. I have always admired the breadth of his knowledge—it was genuinely enjoyable even to sit in on his undergraduate lectures—and I am deeply grateful for his continued support throughout my PhD in so many different ways. More importantly, I am grateful for his patience, understanding, and generosity, which helped me through many difficult moments and allowed me to grow not only as a researcher, but also as a person. It has been a great honor to study with him. Sometimes I wish I could simply keep learning from him for much longer, without all the pressure and deadlines that come with doing a PhD. Looking back, I feel very fortunate that I chose an advisor who was right for me.

I would also like to thank Kourosh Saberi and Jonathan Venezia for their technical guidance and intellectual support throughout my projects, and equally for the kind words, encouragement, and positive perspectives that often helped me keep moving forward. I am also grateful to my committee member Connor Mayer for his thoughtful and generous feedback.

I owe special thanks to Jeffrey Krichmar for his kindness and support from the very beginning of my PhD. He was the first instructor I encountered as a PhD student, at a time when I had to begin the program remotely from the other side of the world. He was willing to meet with me individually every week simply to make sure that I understood everything. I am grateful not only for what he taught me, but also for all the care and encouragement he offered along the way. My only regret is that, with limited time and energy during the PhD, we never had the opportunity to work on a project together. I would have truly enjoyed that.

Within the lab, I especially want to thank our lab manager, Haleh Farahbod. Her continuous presence and support created an environment in which running experiments never felt quite so lonely. There was always a comforting sense that someone was there in the lab and that I could reach out whenever I needed help. Over the years, she also became much more than a lab manager to me—she became a senior friend whom I could rely on.

I am also deeply grateful to Minkyu Kim, especially for his support during the final stage of my PhD. He generously shared his knowledge and expertise in MRI, helped me solve problems with remarkable efficiency, and at times almost became my day-to-day mentor. I was very fortunate to have a senior labmate like him during that period.

A big thank-you also goes to my research assistant, Karina Keming Yang. Although we worked together for only a relatively short period of time, working with her made me wish we could have started from the very beginning of my projects. I am especially grateful for the timely and practical help she gave me when I suddenly had to move outside the United States, helping with experiments as well as with the chaos of moving and packing.

I would also like to thank my other lab members, including Ya-Ning Wu, Will Chih-Chao Chang, Akbar Hussain, Grant Walker, Lynn Kurteff, Solena Mednicoff, and Saori Morita, for their companionship over the past several years. I treasure all those seemingly small

moments—chatting in the lab, walking back together, having meals, or exchanging random thoughts in between work. Those ordinary moments mattered more than they probably appeared to at the time. They made a long and often lonely journey feel much less lonely.

I am grateful to the Department of Cognitive Sciences for giving me a remarkably supportive and friendly environment in which to complete my PhD. Looking back, I sometimes feel that, had I gone to a different university or joined a different program, I might not have made it this far. The people here made an enormous difference.

Special thanks go to Joachim Vandekerckhove for welcoming us into the program during the difficult pandemic years, for his continued support, and for giving me the opportunity to pursue a master's degree in Statistics alongside my PhD. I am also grateful to Aaron Bornstein. Serving on the colloquium committee with him for two years was one of the genuinely fun parts of graduate school, and I greatly enjoyed organizing and hosting those events.

I would also like to thank Jeffrey Rouder, who somehow made statistics feel approachable, intuitive, and even fun. His teaching was part of what made me believe that pursuing a master's degree in Statistics might actually be possible. I am equally grateful for his generosity in welcoming students into his home during the holiday season, which gave me some warm glimpses of American family life during my years far away from home.

My sincere thanks also go to the administrative staff who helped me navigate graduate school: our graduate advisor, Jennifer Gerson; department manager, Clara Schultheiss; and department analyst, MyLee Ryan-Fatigun. Their responsiveness and willingness to help meant a great deal to me, especially as an international student navigating a system that was not always easy to understand.

I am also grateful to the Department of Statistics. It was refreshing and intellectually rewarding to learn alongside a completely different group of people working on very different kinds of problems while I was pursuing my PhD. That parallel academic life gave me another way of thinking, another community, and another part of graduate school that I will remember fondly.

Beyond academia, I want to thank Kenta Watanabe, my closest friend during my PhD years, for always being there and for accompanying me through some of the darkest periods of this journey. I am especially grateful for everything he did during the final stretch to help me bring this PhD to an end. There are many small and large things I could thank him for, but after all these years, it is difficult to separate them into a list. I am simply very grateful that I had such a friend.

I would also like to thank Lingyu Gan for all the random conversations that seemed to be about nothing important and yet somehow mattered, and, of course, for the company of her adorable dog. She gave me the reassuring feeling that there was always someone nearby whom I could reach out to and rely on.

I am grateful as well to all my friends outside my academic field. Their presence was not always continuous, and sometimes appeared only here and there, but those moments of friendship played an important role in helping me keep moving forward.

I would like to give special thanks to JJ Lin, who has been my muse since my undergraduate years and, in many ways, the reason I chose this path in the first place. The question that first drew me to the study of music and the brain was a simple and personal one: why do I love his music so much? After all these years, I still do not fully know the answer, and perhaps I never will, but I understand a little more than when I started. In return, throughout all the years that followed, his music and concerts continued to accompany me, motivate me, and carry me through many difficult moments, and remind me why music was worth studying in the first place.

And then, Yo Oizumi. Thank you for your love, and simply for being who you are. You came into my life like a light. Thank life for the unexpected encounter at the end of 2024, for not only saving my life at a difficult moment, but also for making me feel again that life could be worth living, that beauty and joy still existed in the world, and that there were still things ahead worth looking forward to. Thank you for continuing to sing, for the songs you wrote and chose, for all the work that has brought me so much happiness, and for all the live performances that accompanied the final stretch of this PhD. I will especially remember the live performance just four hours before my defense. In one of the most exhausting periods of my life, those moments gave me somewhere else to look, something to wait for, and many reasons to smile. There are things for which "thank you" feels strangely inadequate, and perhaps even a little too distant. I am simply very glad that, in this particular lifetime, our paths crossed. And I am glad that you exist in the world I get to live in.

A PhD is hard. But all these people made the journey worth taking.

I would also like to thank all of my non-human companions along the way. Thank you to Disneyland Resort in Anaheim, California, which became my little emotional refuge and a happy second home only a thirty-minute drive from campus. Thank you to the books I read, the concerts I attended, the games I played, the movies, dramas, and television shows I watched, the stuffed animals that stayed on my bed, and, of course, all the music I listened to along the way. They may seem unrelated to completing a PhD, but they were part of how I made it through one.

Lastly, I want to thank myself: for being strong and persistent; for continuing even when I wanted to stop; for remaining true to myself; for becoming a better person along the way; and, perhaps most importantly, for making it through all of this without becoming someone I would not have liked.

Generative AI tools, including large language models, were used for language polishing, copyediting, and improving readability and consistency across chapters. The scientific content—including the hypotheses, experimental design, data analysis, and interpretation—was conceived and carried out by the author.

VITA

Jie Wan

EDUCATION

Doctor of Philosophy in Cognitive Sciences **2020–2026**
University of California, Irvine *Irvine, CA, USA*
Supervisor: Prof. Gregory Hickok

Master of Science in Statistics **2021–2024**
University of California, Irvine *Irvine, CA, USA*

Visiting Student, Department of Psychology **2018**
University of Cambridge *Cambridge, UK*
Supervisor: Prof. Brian C. J. Moore

M.Sc. in Behavioural and Cognitive Neurosciences (Research) **2016–2019**
Cognitive Neurosciences and Cognitive Modelling (C-track)
University of Groningen *Groningen, Netherlands*
"Top Rated Programme" (Higher Education Guide)

Bachelor of Science in Biological Science **2011–2015**
Lida Honors College, Hunan University *Changsha, China*

ADDITIONAL TRAINING

Groningen Spring School on Cognitive Modeling **2017**
University of Groningen *Groningen, Netherlands*

Summer School on Neural and Cognitive Science **2015**
Peking University & Tsinghua University *Beijing, China*

ACADEMIC EMPLOYMENT

Teaching Assistant **2020–2025**
Department of Cognitive Sciences, University of California, Irvine *Irvine, CA, USA*

Graduate Researcher / Lab Manager **2018–2020**
Department of Psychology and Neuroscience, Duke University *Durham, NC, USA*
Supervisor: Prof. Tobias Overath

SERVICE

Department Colloquium Committee Member **2023–2025**
Department of Cognitive Sciences, University of California, Irvine *Irvine, CA, USA*
Invited speakers from around the world to give virtual or in-person talks to the department; scheduled meetings between speakers and department faculty and graduate students; created and hosted panels and student flash talks with speakers.

PUBLICATIONS

Leahy, J., Kim, S. G., Wan, J., & Overath, T. An analytical framework of tonal and rhythmic hierarchy in natural music using the multivariate temporal response function. **2021**
Frontiers in Neuroscience, 15, 665767.

Moore, B. C., Wan, J., Varathanathan, A., Naddell, S., & Baer, T. No effect of musical training on frequency selectivity estimated using three methods. **2019**
Trends in Hearing, 23, 2331216519841980.

Wan, J., & Taatgen, N. A. The influence of music and music familiarity on time perception. **2018**
In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), Proceedings of the 40th Annual Conference of the Cognitive Science Society (pp. 2657). Austin, TX: Cognitive Science Society.

Wang, P., Xiao, W., Hu, S., Li, Q., Cheng, P., Yuan, C., Zhang, C., He, Z., Hu, R., Wan, J., Zhou, P., & Guo, X. A comparative study of pectin and caffeine content in different strains and tissues of okras (*Abelmoschus esculentus* L.). **2015**
Acta Laser Biology Sinica, 24(2), 175–179.

Hu, S., Xiao, W., Zhang, C., Wang, P., Chang, H., Wang, Y., Yuan, C., Lu, X., Wan, J., & Guo, X. Progress on regulation, transportation and physiological function of florigen Flowering Locus T. **2014**
Life Science Research, 18(6), 527–532.

Hu, S., Zhang, C., Li, Q., Wang, P., Yuan, C., Wan, J., Chang, H., Wang, Y., Lu, X., Xiao, W., & Guo, X. Determination of the content of total flavonoids of different *Abelmoschus esculentus*. **2013**
Acta Laser Biology Sinica, 22(4), 375–378.

Hu, S., Yuan, C., Zhang, C., Wang, P., Li, Q., Wan, J., Chang, H., Ye, J., & Guo, X. Comparative study of total flavonoid contents from the different tissues and varieties of *Abelmoschus esculentus*. **2013**
International Journal of Medical Sciences and Biotechnology, 1(3), 26–30.

CONFERENCE PRESENTATIONS

Wan, J., Venezia, J., & Hickok, G. A comparative analysis of vocal vs. manual synchronization to auditory rhythm. *(Oral, Nanosymposium.)* **2024**
Society for Neuroscience (SfN) Conference, Chicago, USA.

Wan, J., & Zeng, J. "Microbial Miner": an iron-sensing and absorptive system and an opto-apoptosis system. *(Oral.)* **2014**
Giant Jamboree of the International Genetically Engineered Machine competition (iGEM 2014), Boston, USA.

Wan, J., Venezia, J., & Hickok, G. A comparative analysis of vocal vs. manual synchronization to auditory rhythm. *(Poster.)* **2024**
Advances and Perspectives in Auditory Neuroscience (APAN) symposium, SfN pre-meeting, Chicago, USA.

Wan, J., Venezia, J., & Hickok, G. A comparative analysis of vocal vs. manual synchronization to auditory rhythm. *(Poster, Early Career Poster Session.)* **2024**
Society for Neuroscience (SfN) Conference, Chicago, USA.

Maggu, A. R., Murphy, D. L. K., Hsiang, M., Wan, J., & Overath, T. Online neural monitoring of statistical learning via click detection in a word-segmentation paradigm. ***(Poster.)*** **2020**
Advances and Perspectives in Auditory Neuroscience (APAN) symposium, SfN pre-meeting, USA.

Wan, J., & Taatgen, N. A. The influence of music and music familiarity on time perception. ***(Poster.)*** **2018**
40th Annual Meeting of the Cognitive Science Society, Madison, USA.

Wan, J., & Taatgen, N. A. The influence of music and music familiarity on time perception. ***(Institutional.)*** **2017**
Behavioural and Cognitive Neurosciences Summer Symposium, University of Groningen, Netherlands.

Wan, J., & Liu, T. Neuronal oscillations in sensory selection. ***(Institutional.)*** **2015**
Summer School: Neural and Cognitive Science, Peking University and Tsinghua University, China.

HONORS AND AWARDS

Completion Fellowship, University of California, Irvine	**2025**
Graduate Dean's Dissertation Fellowship, University of California, Irvine	**2025**
Trainee Professional Development Award, Society for Neuroscience	**2024**
First Poster Award, Behavioural and Cognitive Neurosciences Summer Symposium, University of Groningen	**2017**
Maria Sibylla Merian Scholarship, University of Groningen	**2016**
Excellent Group Presentation Award, Summer School of Neural and Cognitive Science, Peking University & Tsinghua University	**2015**
Outstanding Graduates Award, Hunan University	**2015**
Second Award, Undergraduate Excellent Graduation Project (Thesis), Hunan University	**2015**
First Award, National Student Innovation Training Project, Hunan University	**2015**
Merit Student, Hunan University	**2013–2014**
Yangtze River Environment Scholarship, Hunan University	**2013–2014**
First Award Scholarship, Hunan University	**2013–2014**
Second Award Scholarship, Hunan University	**2012–2013**

ABSTRACT OF THE DISSERTATION

The Origins and Mechanisms of Auditory and Motor Rhythm Entrainment: Behavioral, Neuroimaging, and Psychophysical Investigations

By

Jie Wan

Doctor of Philosophy in Cognitive Sciences

University of California, Irvine, 2026

Professor Gregory Hickok, Chair

Humans and parrots are, among animals, uniquely able to move in time with a musical beat—a capacity termed beat perception and synchronization (BPS). Both are also advanced vocal learners, which has led to the proposal that BPS emerged as a by-product of the neural machinery for complex vocal learning. This dissertation examines rhythmic auditory-motor entrainment across three experiments, asking both where this capacity comes from and whether perceiving a rhythm itself engages the motor system.

The first experiment tested a behavioral prediction of this speech-origin account: if rhythmic synchronization is grounded in the speech system, the vocal system should be at least as good at it as a non-vocal effector. Participants synchronized to isochronous rhythms at varying rates by finger tapping or by tongue clicking, a supralaryngeal vocal effector. Tongue clicking was as consistent as finger tapping and, on average, more closely aligned to the beat, leaving the speech-origin premise intact.

The second experiment mapped the neural substrate of the same comparison with fMRI, asking whether the two effectors are served by segregated circuits or by a shared component. Beyond the expected effector-specific regions, finger and tongue synchronization converged on common cortex, most notably the dorsal precentral speech area (dPCSA), an auditory-

weighted region of the speech-coordination network. Because the task involved no pitch or laryngeal control, the engagement of the dPCSA suggests it is not modality-specific but multifunctional—an auditory-weighted region within the motor system that may act as a common node for auditory-motor synchronization across effectors.

The third experiment turned from the effectors of synchronization to the mechanism of perception, testing the motor theory of rhythm—the claim that perceiving a rhythm depends on covertly simulating it motorically. Using a forward-entrainment paradigm, the auditory rhythm was removed and participants generated the rhythm by tapping, a design that predicts an at-least-as-strong perceptual after-effect if the motor system is its source. A subset of participants showed a modulation of the expected magnitude, but the enhancement predicted by the motor account was not observed, and the pattern was not consistently phase-aligned across participants; the results are therefore inconclusive, neither confirming nor excluding a motor contribution.

Together, these experiments support a speech-based origin for rhythmic synchronization and reveal a shared, auditory-weighted cortical substrate for it across vocal and non-vocal effectors, while leaving open whether rhythm perception itself depends on motor simulation.

Chapter 1

General Introduction

1.1 General Background

Think about the moment when, at a live concert, you naturally moved your body or waved the light stick in your hand along with the rhythm of the music. You may wonder why the body tends to move while listening to rhythmic music, and why you and the rest of the audience move along in a similar rhythm.

This is a common phenomenon in daily life: listening to a complex sound such as music, extracting the rhythm within it, and producing movement in time with the sound (dancing, grooving, and so on). We refer to this process as beat perception and synchronization (BPS). It has also been described more generally as sensorimotor synchronization, defined as the coordination of rhythmic movement with an external rhythm (Repp & Su, 2013). There is yet no clear consensus on its underlying mechanisms, but the topic draws considerable interest for several reasons. Dance, one of the central activities associated with music, likewise depends on this ability.

1.1.1 Groove

Groove refers to the quality of music that induces movement in the listener (Stupacher, Hove, Novembre, Schütz-Bosbach, & Keller, 2013). Many factors contribute to groove, particularly ones related to rhythm and timing. Repetition of a rhythm increases attention and engagement (Pressing, 2002), low-pitched beats tend to induce movement (Stupacher et al., 2013), and a moderate level of rhythmic complexity (Witek, Clarke, Wallentin, Kringelbach, & Vuust, 2011; Matthews, Witek, Lund, Vuust, & Penhune, 2020) as well as syncopation tend to make listeners groove more (Holm & Isaksson, 2010). More generally, it is unexpected beat patterns, not only syncopation, that promote groove.

There is also a sweet spot of tempo at which the tendency to groove is strongest. Dance music shows a clear tempo peak around 120 bpm (Moelants, 2002), which coincides with the preferred movement rate and the spontaneous tapping tempo (Fraisse, 1982; MacDougall & Moore, 2005).

Researchers have attempted to quantify groove and to study how postural fluctuations entrain to rhythms in complex sounds such as music by measuring fluctuations in the center of pressure (CoP) while participants listened to low- or high-groove music. Musical groove was found to influence radial sway variability: the higher the groove, the lower the variability of sway (Ross, Warlaumont, Abney, Rigoli, & Balasubramaniam, 2016). This illustrates a robust and shared rhythmic reaction to musical groove across individuals. Related work has measured postural sway and motor-evoked potentials (MEPs) from single-finger recordings (Proksch, Comstock, Médé, Pabst, & Balasubramaniam, 2020).

1.1.2 Walking and other rhythmic body movement

Everyday activities such as walking also follow a rhythmic pattern. Human walking overground (rather than on a treadmill) has an average preferred rate of about 120 steps per minute, equivalent to roughly 2 Hz (Murray, Drought, & Kory, 1964). More controlled laboratory studies found similar results: measuring head linear acceleration during locomotor activities (walking, running, cycling) and non-locomotor activities (working, driving, riding public transportation), the average frequency of locomotion was around 2 Hz (MacDougall & Moore, 2005). Similarly, when participants freely tap a rhythm with the finger, the average rate is around 2 Hz (MacDougall & Moore, 2005).

One explanation is that the step frequency that minimizes oxygen cost is around 2 Hz (MacDougall & Moore, 2005). A deeper account may come from evolution, given that our ancestors transitioned from quadrupedal to bipedal locomotion. A further observation is that when people walk together, their walking rhythms tend to gradually synchronize (Katevas, Haddadi, Tokarchuk, & Clegg, 2015).

1.1.3 Rhythm in speech and music

From the standpoint of rhythm, speech and music differ substantially. Most music is systematically rhythmic, whereas speech is not; if the two were alike in their rhythmic properties, the questions raised above might be far easier to answer. Because speech and music perception and production are often studied and compared together, their differences are worth examining.

Speech is quasi-rhythmic. There is no fixed average rhythm for speech in the amplitude envelope, which tracks the syllable rate (Poeppel & Assaneo, 2020). The rhythm of speech is neither regular nor predictable (Dalla Bella, Białuńska, & Sowiński, 2013), and for this reason

we do not bob our heads or dance to speech. The general mean syllable rate of speech is 4–8 Hz across languages (Coupé, Oh, Dediu, & Pellegrino, 2019; Greenberg, Carvey, Hitchcock, & Chang, 2003; Pellegrino, Coupé, & Marsico, 2011; Zhang, Zou, & Ding, 2023), and the power spectrum of the speech envelope likewise drops within 4–8 Hz across languages (Ding et al., 2017; Varnet, Ortiz-Barajas, Erra, Gervain, & Lorenzi, 2017; Zhang et al., 2023), with an average peak near 4 Hz (Poeppel & Assaneo, 2020).

Music, by contrast, shows a preferred peak frequency around 2 Hz, as summarized from 74,000 pieces of Western popular music from 1960–1990 (Moelants, 2002). Cross-cultural research on the structure of human music further indicates a tendency toward isochronous beats, organized into metrical hierarchies and typically grouped in multiples of two or three (Savage, Brown, Sakai, & Currie, 2015), consistent with time signatures such as 4/4 and 3/8 in Western music theory.

The differences in average and peak frequency between music and speech, and across different parts of the body involved in movement, may offer clues to the origins and development of rhythmic synchronization, and are taken up further below.

1.1.4 Humans and parrots

Rhythmic behavior is common across the animal kingdom. Many animals produce rhythmic motor actions—land animals walking, birds flapping in flight, waterbirds paddling, gorillas beating their chests, apes drumming—and many can detect rhythm in auditory or visual signals; even fireflies flash in synchrony. Predictive and tempo-flexible synchronization to a beat, however, is unique to humans and parrots.

The most salient commonality between the two is that both can speak—both are advanced (complex) vocal learners. Unlike songbirds, which use only the syrinx to sing, parrots use

both the syrinx and the tongue to produce sound, much as humans use both the larynx and the supralaryngeal tract to produce speech or song. Their vocal-learning pathways also share many features. In parrots, the vocal-learning and control circuitry comprises dual pathways—a core and a shell. The core system, like that of songbirds, is a premotor–motor–basal-ganglia–thalamic pathway that controls the syrinx; the shell system surrounds the core, and the two together support vocal-learning control (Chakraborty & Jarvis, 2015). These systems may allow independent control of the syrinx and the tongue during vocalization (Patel, 2021). The human vocal-learning and control system is captured by the dual-stream speech coordination model (Hickok, Venezia, & Teghipco, 2023), which likewise comprises two subsystems: a laryngeal pitch-control system and a supralaryngeal phonetic/syllabic control system.

This parallel between the two species—shared vocal-learning ability, shared reliance on both a laryngeal and a supralaryngeal sound source, and analogous dual-pathway neural architectures—motivates the hypothesis, examined throughout this dissertation, that the capacity for rhythmic synchronization is grounded in advanced vocal learning.

1.2 Two hypotheses linking vocal learning and rhythmic synchronization

Why should advanced vocal learning have anything to do with the capacity to synchronize movement to a beat? Both abilities depend on strong forebrain integration between auditory and motor processing: complex vocal learning relies on auditory-guided motor learning and on precise, online auditory-motor interaction to control the vocal organs, and beat-based synchronization relies on the same kind of temporally precise coupling between auditory and motor regions (Patel, 2021). On this basis it has been proposed that beat perception and

synchronization emerged as a *spandrel*—a fortuitous by-product—of the neural machinery for complex vocal learning, present in the two lineages that are advanced vocal learners and absent elsewhere (Patel, 2024; Hickok, 2024). The two accounts considered in this dissertation agree on this shared premise but differ on the mechanism by which vocal learning gives rise to the capacity, and in particular on why the ability comes to be available to non-vocal effectors at all.

On Patel's *fortuitous enhancement* account, the evolution of a dorsal laryngeal pitch-control pathway—linking auditory cortex to dorsal premotor cortex in the service of learned control of vocal pitch—was accompanied by gene-regulation changes that fortuitously strengthened the connections between auditory regions and *non-vocal* dorsal premotor regions lying adjacent to the vocal ones, by way of the inferior parietal cortex (Patel, 2024). On this view the spread of synchronization from vocal to non-vocal effectors is an incidental consequence of strengthening a specifically laryngeal pathway, which is why Patel terms it fortuitous.

Hickok's *coordination conjecture* offers a different mechanism. On this account, rhythmic synchronization is a simple version of what is required to coordinate the two vocal-control streams—the laryngeal pitch-control system and the supralaryngeal articulatory system—during speech; once a system for synchronizing multiple effectors to a common target is in place, it may be easier to extend it to the motor system as a whole than to confine it to the vocal effectors, so that the capacity for non-vocal beat synchronization comes for free (Hickok, 2024). Where Patel locates the fortuitously strengthened pathway in laryngeal pitch control specifically, Hickok locates the origin in the more general problem of coordinating multiple effectors, with the spread across the motor system following from that.

These two hypotheses frame the experiments that follow. The behavioral and neural experiments ask whether rhythmic synchronization is grounded in the speech system, as both accounts assume, and whether its substrate is shared across vocal and non-vocal effectors, as the coordination conjecture in particular would predict. The following sections review the

behavioral and neural evidence on rhythmic synchronization against which these hypotheses are set.

1.2.1 Behavioral findings

A typical synchronization experiment asks participants to tap the finger in time with the beat of a sound sequence they hear. This is a simple paradigm with a long history, and it remains widely used (Repp & Su, 2013). All human participants can synchronize to simple metronomic auditory sequences, and their tap times fall close to the times of the stimuli.

Several key features characterize BPS, or rhythmic entrainment behavior: it is (1) predictive, (2) tempo-flexible, and (3) hierarchical (Patel & Iversen, 2014; Patel, 2021).

(1) Predictive. Participants' taps typically precede the sequence they synchronize to, indicating that beat perception operates predictively (Patel, 2010). This predictive character—also described as active or spontaneous—means that listeners anticipate the timing of upcoming beats and generate anticipatory motor responses in order to maintain entrainment (Patel, 2021). The predictive nature of BPS is consistent with predictive-coding accounts (Spratling, 2017; Vuust & Witek, 2014) and with the dynamic attending theory, on which perceptual resources are concentrated at the expected times of beats (Jones & Boltz, 1989; Patel & Iversen, 2014).

(2) Tempo-flexible. Participants can synchronize precisely across a wide range of tempi, roughly 250 ms to 1 s (Repp & Su, 2013). Within this range there is a preferred tempo band for beat perception, on the order of 500–700 ms (Patel, 2010; MacDougall & Moore, 2005), though the exact figures vary somewhat across studies. Most musical pieces fall between 120 and 150 bpm, with a median near 133 bpm (Van Noorden & Moelants, 1999; Patel & Iversen, 2014).

(3) Hierarchical. When listening to music, we extract abstract beats rather than simply following the sensory pulses, and this hierarchical organization can be subjective and can vary across listeners, who may choose which level to synchronize to even for the same stimulus (Patel & Iversen, 2014).

Patel et al. examined how metricality influences motor synchronization to beats in auditory and visual rhythms. Their patterns included isochronous rhythms from 200 to 800 ms as well as non-isochronous patterns that were either strongly metrical (SM) or weakly metrical (WM), following Povel and Essens (1985); all had a common beat interval of 800 ms, which was the interval participants tapped to. For the auditory rhythms of interest here, overall synchronization accuracy did not improve for the SM patterns relative to isochronous ones, and accuracy was impaired for the WM patterns in which some beats were silent (Patel, Iversen, Chen, & Repp, 2005).

Thompson et al. used a simpler rhythm—500 ms and 667 ms isochronous snare-drum sequences—but examined the effects of development and musical experience. Measuring tapping synchronization from childhood to older adulthood, they found peak performance around young adulthood, together with a positive association between time spent playing music and synchronization performance (Thompson, White-Schwoch, Tierney, & Kraus, 2015).

With respect to musicality, musicians and non-musicians rate the groove of music similarly (Stupacher et al., 2013; Matthews et al., 2020), but process off-beat versus on-beat events differently for high- versus low-groove music: high-groove music modulated corticospinal excitability as measured by motor-evoked potentials (Stupacher et al., 2013). Moreover, musicians' groove ratings depend on both rhythmic and harmonic complexity, whereas non-musicians' ratings depend only on rhythmic complexity, suggesting that groove for musicians is shaped by non-rhythmic factors as well (Matthews et al., 2020). Open questions remain about whether different training backgrounds yield different synchronization abilities—for

example, whether professional drummers or beatboxers outperform untrained individuals—though at least one study found no significant difference across musicians playing different instruments in rhythmic tasks (Matthews, Thibodeau, Gunther, & Penhune, 2016).

1.2.2 Neural findings

Although the auditory perception and the motor production of rhythm need not occur at the same moment, brain-imaging studies show that the motor system is engaged when listening to auditory rhythms even in the absence of any overt movement (Repp & Su, 2013; Patel, 2021). The regions involved range from cortical premotor areas to subcortical motor structures, including the basal ganglia, cerebellum, SMA, pre-SMA, and premotor cortex (Grahn & Brett, 2007; Chen, Penhune, & Zatorre, 2008; Grahn & Rowe, 2009; Matthews et al., 2020; Patel, 2021; Repp & Su, 2013). One prominent proposal is that the motor system actively predicts the timing of beats even without movement (Arnal, 2012; Patel & Iversen, 2014; Morillon & Baillet, 2017; Proksch et al., 2020; Cannon & Patel, 2021; Patel, 2021).

Beat perception is associated with greater connectivity among SMA, premotor cortex, auditory cortex, and putamen, although the putamen does not respond differentially to varying beat complexity (Grahn & Rowe, 2009). The basal ganglia is engaged during perceptual timing judgements and in the prediction of the beat in auditory rhythms (Grahn & Rowe, 2013; Repp & Su, 2013).

Using fMRI while musicians tapped to rhythms of varying complexity, the superior temporal gyrus (STG), premotor cortex, and ventrolateral prefrontal cortex (VLPFC) were identified as relevant to beat finding and tapping; STG and VLPFC were linked to the retrieval, selection, and maintenance of the beat, with the basal ganglia serving more basic sensorimotor functions and prefrontal cortex serving higher-order cognition (Kung, Chen, Zatorre, & Penhune, 2013). In general, timing and rhythm tasks—both perception and production—

activate a common set of regions including premotor cortex, SMA, cerebellum, and basal ganglia, with the basal ganglia and SMA being especially important for attention in time and for temporal sequencing and prediction (Grahn & Brett, 2007).

A review by Proksch et al. (2020) summarized proposed functions region by region, covering the cerebellum, basal ganglia, primary motor cortex, premotor cortex, medial premotor cortex, supplementary motor area, parietal cortex, auditory cortex, and the dopaminergic (striatal-dopamine) system: for example, the cerebellum is implicated in recruiting the motor system during rhythmic listening, the basal ganglia in internal beat generation within a striato-thalamo-cortical network for beat-based timing, and the putamen in beat prediction.

Reproducing rhythmic sequences has been shown to elicit increased activation in the basal ganglia and SMA in both musicians and non-musicians, with no difference between the two groups in these regions, and no behavioral difference across conditions (Grahn & Brett, 2007). Varying the placement of accents within a regular beat modulated the coupling between premotor and auditory cortex (STG), and this modulation was greater in musicians than in non-musicians (Grahn & Rowe, 2009; Chen, Penhune, & Zatorre, 2009); musicians also show stronger coupling between auditory (STG) and motor regions (premotor cortex and SMA) than non-musicians (Repp & Su, 2013).

Chen et al. used fMRI to show neural coupling between auditory cortex and premotor cortex during motor synchronization to auditory rhythms. Using five variants of a single isochronous pattern, created by adjusting the amplitude of every third tone to produce waltz-like rhythms of graded strength, they found that this manipulation modulated both behavioral and neural responses in auditory (posterior STG) and dorsal premotor cortex (dPMC) (Chen, Zatorre, & Penhune, 2006). In a follow-up study designed to dissociate auditory perception, motor planning, and motor production, they compared participants who listened while anticipating that they would tap with participants who listened without such anticipation, using simple-metric, complex-metric, and non-metric patterns. Listening

with anticipation recruited SMA, mid-premotor cortex, and cerebellar lobule VI; the same motor regions were engaged when listening without anticipation. However, ventral premotor cortex was uniquely associated with action and action-coupled perception, dorsal premotor cortex was sensitive to the higher-order selection of temporal actions, and SMA was most engaged during motor production (Chen et al., 2008). The authors proposed that the left planum temporale and premotor cortex are key regions for synchronization (Chen et al., 2006).

Tsai et al. measured BOLD responses to oral representations of percussion music. Comparing passive and active listening to percussion music, melodic music, and syllabic representations of learned percussion, they suggested that the integration of sound with subvocal rehearsal is associated with the right planum temporale during active listening, whereas mapping sound to articulatory and laryngeal gestures engages the left mid-premotor cortex (Tsai, Chen, Chou, & Chen, 2010).

1.3 A dual system for speech coordination

The regions reviewed above—premotor cortex, SMA, the basal ganglia, and the cerebellum—constitute the network most often associated with beat-based timing. A separate line of work on the cortical control of speech has recently described a set of premotor regions that may also be relevant to auditory-motor synchronization. In the dual speech coordination model (Figure 2.1), speech coordination is divided into two parallel systems on the precentral gyrus: a dorsal precentral speech area (dPCSA), which coordinates pitch-related aspects of speech, prosody, and song, and a ventral precentral speech area (vPCSA), which coordinates phonetic and syllabic articulation of the supralaryngeal tract (Hickok et al., 2023). The two are distinguished by the sensory input they weight most heavily: the dPCSA is auditory-weighted, while the vPCSA is somatosensory-weighted. Notably, the dPCSA shows properties that

are unusual for premotor cortex—it carries speech-related spectrotemporal receptive fields resembling those of auditory cortex, and responds to acoustic input even in the absence of an overt speech task (Hickok et al., 2023)—so that it behaves, in part, like an auditory region situated within the motor system.

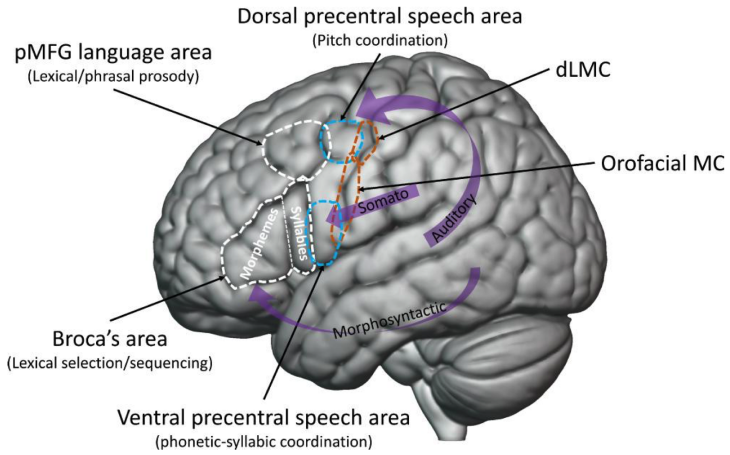


Figure 7 Schematic depiction of the dual speech coordination system model. Input labels (arrows) are intended to reflect a bias toward the named input

Figure 1.1: Schematic depiction of models. The dual speech coordination system model from Hickok et al. (2023).

It is worth noting that the dorsal premotor region implicated in the beat-processing literature (Chen et al., 2006) and the dPCSA of the speech model may correspond to overlapping cortex, a correspondence that has not been directly established but that the neural experiment in this dissertation is positioned to examine. Whether, and how, these speech-coordination regions participate in rhythmic synchronization with vocal and non-vocal effectors is an open question, taken up in the second experiment.

1.4 From synchronization to the perception of rhythm

The evidence reviewed so far concerns the production of synchronized movement and the motor regions engaged during rhythm processing. A related question concerns perception: when a listener perceives a rhythm, is the motor system itself involved in the perceiving?

On the *motor theory of rhythm*, the perception of a beat depends on covertly simulating it with the motor system. This idea is developed most explicitly in the Action Simulation for Auditory Prediction (ASAP) hypothesis, on which the motor planning system entrains its activity to the beat period and communicates predictive signals to auditory regions, thereby shaping the perceptual interpretation of the rhythm (Patel & Iversen, 2014).

The ASAP hypothesis occupies a notable position in the present framework. It is one of the strands of evidence Patel draws on for the auditory–premotor coupling that supports his fortuitous-enhancement account—the claim that motor planning regions supply predictive signals to auditory regions during beat perception (Patel, 2024)—and it is at the same time the clearest statement of the motor theory of rhythm. A test of ASAP is therefore also, in part, a test of one component on which the broader vocal-learning account rests.

One paradigm well suited to such a test comes from research on forward entrainment, defined as that part of behavioral or neural entrainment that outlasts the entraining stimulus, by analogy to forward masking in psychoacoustics (Saberi & Hickok, 2023). Listening to a rhythm produces a perceptual after-effect that persists for a short time once the rhythm has stopped; if this after-effect is generated by the motor system, then producing a rhythm—rather than listening to one—should be sufficient to elicit it. The third experiment uses this logic to ask whether motor production alone can generate the effect, and thereby to bear on whether rhythm perception depends on motor simulation.

1.5 The scope of the present work: metronomic rhythms

The behavioral and neural findings reviewed above concern rhythmic synchronization in the broad sense, ranging from simple metronomic sequences to complex, hierarchically structured music. Perceiving the beat in music requires first extracting it from rich temporal

structure—a process that draws on metrical hierarchy, expectation, and cultural experience—before any movement can be aligned to it. The present dissertation is concerned with the more basic auditory-motor synchronization mechanism that this beat-extraction process ultimately feeds. For this reason, all three experiments use isochronous metronomic sequences rather than music: doing so isolates the synchronization mechanism itself from the additional processing involved in extracting a beat from complex sound. The conclusions drawn here therefore apply, in the first instance, to synchronization with simple rhythms; extending them to full beat perception and synchronization with musically structured rhythms is a natural direction for future work, noted where relevant in the individual chapters.

Chapter 2

Rhythmic Synchronization Across Effectors: A Behavioral Comparison of Finger Tapping and Tongue Clicking

2.1 Introduction

As mentioned above, BPS is a striking behavior that, among animals, is shared only by humans and parrots. A useful way to understand its origins is to ask what humans and parrots have in common that other animals lack. Both are advanced vocal learners: both produce complex learned vocalizations, using comparable vocal organs and comparable brain circuitry for vocal control. Humans use both the larynx and the supralaryngeal tract for vocalization; parrots likewise use both the syrinx and the tongue (suprasyrinx), whereas songbirds, for example, produce sound via the syrinx alone. At the level of brain organi-

zation, parrots are hypothesized to possess dual pathways—a core and a shell system: the core is a premotor–motor–basal-ganglia–thalamic pathway controlling the syrinx, and the shell surrounds the core, the two together supporting vocal-learning control (Chakraborty & Jarvis, 2015). These two systems may allow independent control of the syrinx and the tongue during vocalization (Patel, 2021). Songbirds, by contrast, have only the core system. Humans are hypothesized to have an analogous dual-stream architecture for speech: a laryngeal pitch-control system and a supralaryngeal phonetic/syllabic control system (Hickok et al., 2023).

Crucially, BPS is found only in these advanced vocal learners and not in animals lacking this ability—a species-level co-occurrence that motivates the hypothesis that BPS and complex vocal learning are linked. Speech requires coordinating these hierarchical vocal systems with precise timing, and because BPS also requires precise temporal coordination, it is plausible that BPS has its origin in speech.

On the dual speech coordination model (Hickok et al., 2023; Figure 2.1), speech coordination is organized into two parallel premotor systems on the precentral gyrus. The dorsal precentral speech area (dPCSA) is an auditory-weighted region that coordinates pitch-related aspects of speech, prosody, and song; the ventral precentral speech area (vPCSA) is a somatosensory-weighted region that coordinates phonetic and syllabic articulation of the supralaryngeal tract. A region dorsal to the dPCSA supports body movements such as those of the limbs. These same regions are central to the neuroimaging experiment that follows, where we ask whether the dPCSA acts as a shared hub for auditory-motor synchronization; here, in the behavioral experiment, they provide the anatomical backdrop for comparing a vocal against a non-vocal effector.

Speech requires synchronization between these two vocal systems, even though speech is only quasi-rhythmic, unlike music. Given this, if BPS has its origin in speech, why can we groove or dance—that is, why does rhythmic synchronization spread from vocal effectors to

non-vocal effectors? This is the key question motivating this experiment, and there are two hypotheses we consider.

Patel (2024) (Figure 2.1) proposed that strengthening auditory–laryngeal connections led to fortuitous connectivity between auditory and non-vocal motor systems, via an indirect pathway from auditory cortex through the angular gyrus to dorsal premotor areas. On this account, the strengthening of the laryngeal pitch-control pathway spreads to more dorsal, non-vocal motor-control regions.

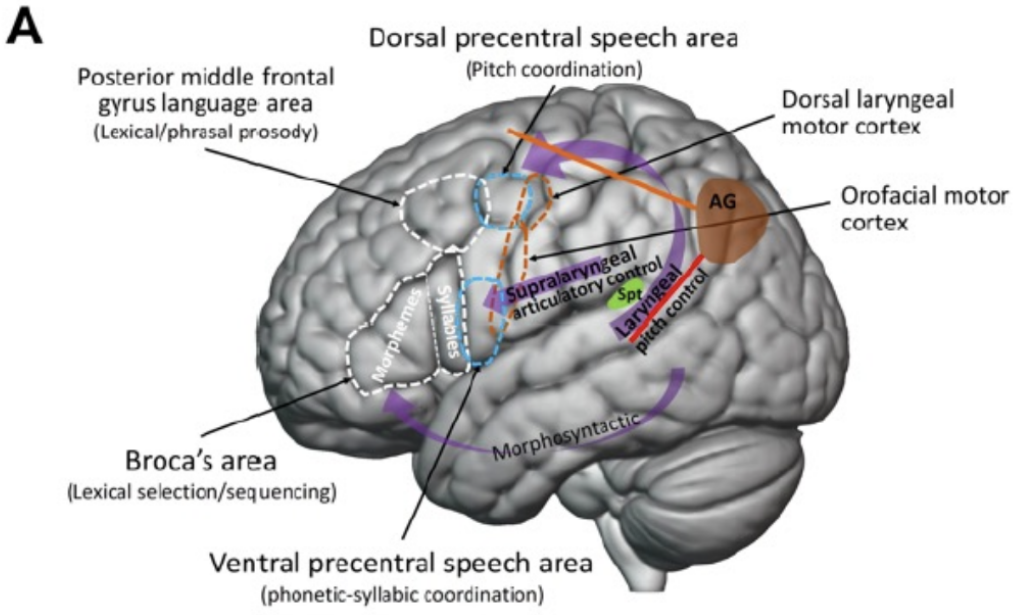


Figure 2.1: Schematic depiction of models. The proposed dorsal stream pathway involved in beat processing superimposed from Patel (2024).

Hickok (2024) offered an alternative: perhaps it is easier to spread BPS to the whole of the motor system than to just two motor-planning pathways—once the ability to coordinate speech-production systems with precise timing is in place, BPS comes for free. On this account, the spread encompasses the laryngeal, supralaryngeal, and more dorsal non-vocal motor regions together.

The present experiment is best understood as a sanity check on the shared premise that BPS has its origin in speech, rather than as a decisive test between these two accounts—we do not yet know which hypothesis is correct, only whether we are on the right track. The reasoning is straightforward: if rhythmic synchronization is rooted in speech, then the speech system itself should be at least as good at rhythmic synchronization as a non-vocal effector, because that is where the ability began. To our knowledge this has not been tested directly. Accordingly, we compare a non-vocal effector (finger tapping) against a supralaryngeal vocal effector (tongue clicking) synchronizing to the same auditory rhythms. We chose finger as a representative non-vocal effector and tongue as a representative supralaryngeal vocal effector; importantly, tongue clicking engages only supralaryngeal motor organs, not the larynx, because it does not involve phonation.

The critical prediction concerns the shared premise: if the speech-origin account is on the right track, tongue clicking should perform at least as well as finger tapping. A clear finding that the supralaryngeal effector were worse than the limb would count against the premise that the vocal system is especially good at synchronization. A finding that tongue clicking performs comparably to, or better than, finger tapping leaves the premise intact. We note that this comparison does not adjudicate between Patel's and Hickok's accounts: Patel does not make an explicit prediction at the level of individual effectors (in particular, he does not claim finger should outperform vocal effectors), so a tongue at least as good as finger result is best read as consistent with the whole-system spreading picture without decisively favoring either hypothesis.

Beyond how vocal synchronization comes to include non-vocal synchronization, we also included several tempi because different effectors are thought to have different preferred rates. As reviewed in the general background, the average syllable rate for speech is around 4 Hz, whereas the preferred rate for dance, grooving, walking, or finger tapping is closer to 2 Hz, and body sway is slower still, below 2 Hz—rates that correspond to different motor effectors.

We therefore span 2, 3, and 4 Hz so that the effector comparison is not tied to a single rate that happens to suit one effector. A systematic characterization of effector-specific preferred rates is beyond the scope of the present analysis and is reserved for future work.

For reference, a related comparison study asked participants to whisper or finger-tap to syllable and piano-tone sequences at 2 Hz or 4.5 Hz; synchronization was generally better at 2 Hz, and finger tapping outperformed whispering at 2 Hz but not at 4.5 Hz, though performance there was quantified using phase-locking values (PLVs) rather than timing error (Barchet, Henry, Pelofi, & Rimmele, 2024).

2.2 Methods

2.2.1 Participants

Thirty participants (age mean 22.57 ± SD 6.23; all right-handed) took part in the study. Participants identified their gender as female (n = 20), male (n = 7), nonbinary (n = 1), Others (n = 1), or preferred not to disclose (n = 1). All participants provided written informed consent, and the study was approved by the Institutional Review Board of the University of California, Irvine.

2.2.2 Experimental task and stimuli

Each run consisted of repetitions of a snare-drum sound, using four repetition counts (8, 10, 12, and 14 beats) crossed with three frequencies (2, 3, and 4 Hz), yielding twelve combinations. Within a run, the twelve combinations were concatenated in a randomized order, with the constraint that the same frequency did not occur consecutively (e.g., an 8-repetition 2 Hz sequence was never immediately followed by a 12-repetition 2 Hz sequence). This dynamic

design ensured that participants could not anticipate the upcoming rhythm to which they had to synchronize. (Figure 2.3)

Participants were instructed to synchronize their movements with the drum sounds, either tapping their index finger or producing tongue clicks ("vocal tapping") in time with the auditory stimuli presented over headphones. Half of the participants performed finger tapping first and the other half tongue clicking first. For each effector—finger tapping and vocal tapping—15 runs were completed, yielding 15 runs per session.

A microphone was used in both effector conditions to record participants' responses, allowing accurate timing alignment between stimuli and responses.

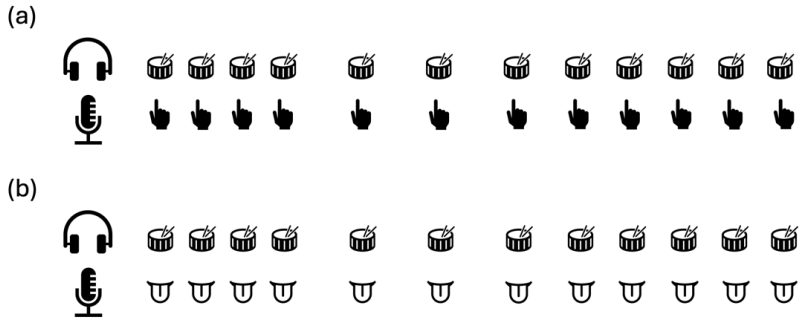


Figure 2.2: Example of Stimuli. (a) Synchronizing with index finger. (b) Synchronizing with tongue clicking.

The snare-drum sound was created in GarageBand and edited in Audacity. The stimulus was calibrated to a sound level of 70 dB (A-weighted). Stimuli were generated in MATLAB (The MathWorks, Natick, MA) and presented at 44.1 kHz through 16-bit digital-to-analog converters (Focusrite Scarlett 2i2) and Sennheiser HD 280 Pro headphones, in a steel-walled, acoustically isolated chamber (IAC Acoustics, Stafford, TX). Subjects' responses were recorded

using a SHURE SM7B microphone.

2.2.3 Data Analysis

For preprocessing, we ran sound-onset detection in MATLAB to identify onsets in all audio recordings, including the snare-drum stimulus files and the finger-tapping and tongue-clicking response recordings. Onsets were detected from the sound envelope: we identified envelope peaks and adjusted the detection parameters until the number of detected peaks matched the number of beats in the design. For the stimulus files, the number of detected peaks always equaled the designed number of beats. Using the stimulus onsets as references, we then matched each stimulus onset to its nearest response onset in the microphone recordings. Finally, a threshold was applied to discard response onsets that fell too far from their corresponding stimulus onset; in practice this step did not change the results, so all onsets were retained in the analyses reported below.

After identifying all response and stimulus onsets, we computed the time difference for each beat as the response onset time minus the corresponding stimulus onset time. This time difference is our primary measure.

One participant was excluded from all analyses. This participant's timing error in the Vocal condition deviated markedly from the rest of the sample: across all three frequencies, both the mean and the standard deviation of the timing error exceeded 5 SD from the group mean and fell outside the range observed across the remaining 29 participants. Notably, this deviation was specific to the Vocal effector; the same participant's Finger data fell within the normal range. This modality-specific pattern indicates that the Vocal data for this participant did not reflect the timing behavior of interest. The participant was therefore excluded, and all reported analyses are based on the remaining 29 participants. The statistical models below were also confirmed to be robust to the inclusion or exclusion of this participant.

Behavioral performance was analyzed at the run level using linear mixed-effects models (LMEMs). Separate models were fitted for the mean time difference and for the standard deviation (SD) of the time difference. For both dependent measures, Hz (2, 3, and 4 Hz), Effector (Finger and Vocal), and their interaction were included as fixed effects. To account for the repeated-measures structure of the data, each model included participant-specific random intercepts and random slopes for Hz, Effector, and the Hz × Effector interaction, as well as a random intercept for run nested within participant. The two models therefore had the following structure:

Mean Time Difference: Linear Mixed-Effects Model

```
mean_timeDiff ~ Hz × Effector
+ (1 + Hz + Effector + Hz × Effector | sub_ind)
+ (1 | sub_ind:runs)
```

and

SD Time Difference: Linear Mixed-Effects Model

```
SD_timeDiff ~ Hz × Effector
+ (1 + Hz + Effector + Hz × Effector | sub_ind)
+ (1 | sub_ind:runs)
```

Analyses were conducted on the final sample of 29 participants. Each participant completed 15 runs, yielding 435 participant-by-run units. As a sanity check, the run-level data structure was verified to ensure that every participant-by-run unit contained all six combinations of Hz (2, 3, and 4 Hz) and Effector (Finger and Vocal). Model singularity was assessed and the random-effects variance–covariance structure was inspected; neither model showed a singular fit. The final fits produced small convergence-gradient warnings slightly above the default tolerance, with maximum gradients of 0.00252 (mean model) and 0.00241 (SD model), relative to a tolerance of 0.002.

Omnibus fixed-effect tests for Hz, Effector, and the Hz × Effector interaction were obtained using Type III tests with the Satterthwaite approximation for denominator degrees of freedom. These are reported separately for mean time difference and SD time difference in the two Fixed Effects tables.

To further characterize the effects, planned contrasts were extracted from the same fitted models without refitting. Using 2 Hz and Finger as the reference condition, these contrasts quantified: the estimated value at 2 Hz for Finger; the changes from 2 to 3 Hz and from 2 to 4 Hz for Finger; the Vocal–Finger difference at 2 Hz; and the change in the Effector effect from 2 to 3 Hz and from 2 to 4 Hz. These are reported in the two Detailed Effects tables.

Model-estimated marginal means and their uncertainty were visualized across the three Hz conditions for Finger and Vocal. Mean time difference and SD time difference are shown together in the primary behavioral figure.

2.3 Results

2.3.1 Time Course Across Beats Within a Rhythm

A descriptive analysis examined the mean time difference across the sequence of 14 beats within each rhythm, with uncertainty shown using the standard error of the mean.

To characterize overall behavior across subjects, we averaged the time difference across all subjects, conditions, and repetition counts (8, 10, 12, and 14 beats), and plotted the result as a function of beat position within the sequence, from the first beat to the last (Figure 2.4). Synchronization stabilized from the 4ˆthˆ beat and remained stable through the 14ˆthˆ beat (the larger error bars toward the end reflect the smaller number of trials contributing to the 13ˆthˆ and 14ˆthˆ beats).

We therefore restricted the following analysis to beats 4th through 8th. The results were robust to this choice: the same pattern held when all beats were included.

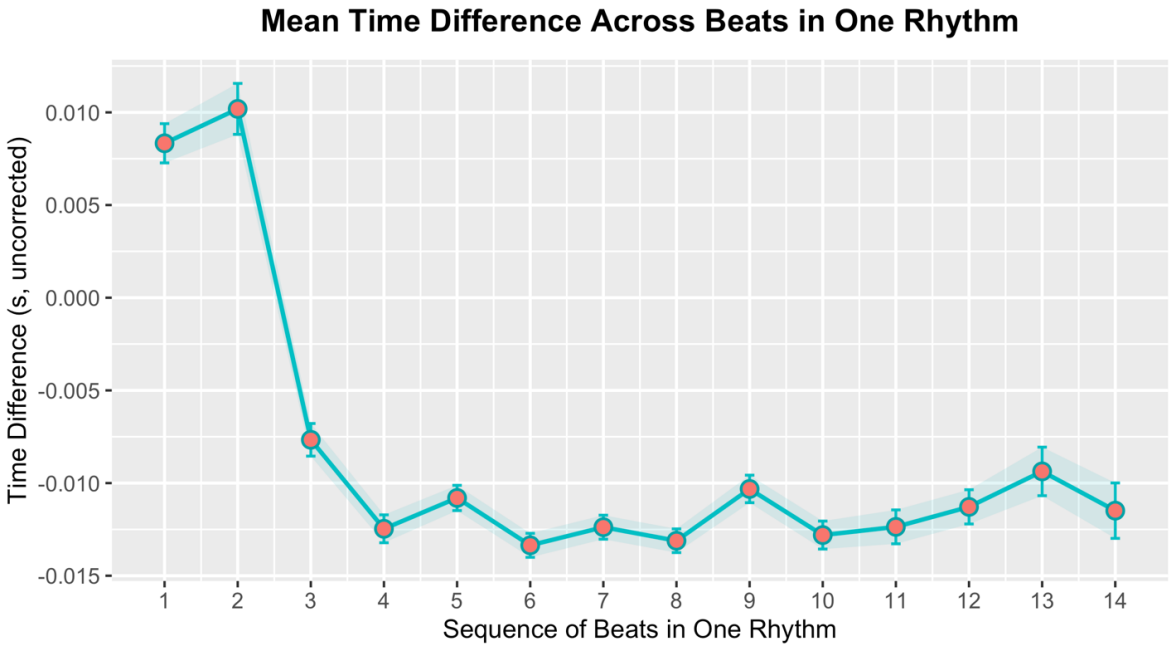


Figure 2.3: Mean of time difference on sequence of beats for all rhythms.

We next computed the mean and standard deviation (SD) of the time difference separately for the two motor effectors (finger tapping and tongue clicking) and the three frequency rates (2, 3, and 4 Hz). These measures were each submitted to a separate regression model: the mean time difference as the dependent variable in panel (a), and the SD of the time difference as the dependent variable in panel (b).

2.3.2 Mean Time Difference

The mixed-effects analysis revealed significant effects of **Hz**, **Effector**, and their interaction on mean time difference. There was a significant main effect of Hz, ($F(2, 28.00)=37.69$, $p<.001$), and a significant main effect of Effector, ($F(1, 27.99)=68.67$, $p<.001$). Importantly, these effects were qualified by a significant Hz $\times$ Effector interaction, ($F(2, 28.00)=19.78$, $p<.001$). These omnibus results are summarized in the Table 1.

Table 2.1: Mean Time Difference: Fixed Effects.

Mean Time Difference: Fixed Effects

Effect	Num df	Den df	F	p
Hz	2	28.00	37.69	**< .001**
Effector	1	27.99	68.67	**< .001**
Hz × Effector	2	28.00	19.78	**< .001**

The detailed contrasts further characterized this interaction. For the Finger condition, the estimated mean time difference at 2 Hz was -0.0416 s ((SE=0.0044), (p<.001)). Relative to 2 Hz, mean time difference increased by 0.0207 s at 3 Hz ((SE=0.0034), (p<.001)) and by 0.0321 s at 4 Hz ((SE=0.0046), (p<.001)). At 2 Hz, Vocal responses differed from Finger responses by 0.0330 s ((SE=0.0048), (p<.001)).

The change in the Effector effect from 2 to 3 Hz was not significant ((b=-0.0057), (SE=0.0041), (p=.173)). In contrast, the Effector difference changed significantly from 2 to 4 Hz ((b=-0.0239), (SE=0.0052), (p<.001)). These contrasts are reported in Table 2.

The corresponding panel of the primary figure illustrates these effects across the three Hz levels for the Finger and Vocal conditions.

2.3.3 SD Time Difference

For SD time difference, there was a significant main effect of **Hz**, (F(2, 27.98)=28.63, p<.001). The main effect of **Effector** was not significant, (F(1, 27.99)=2.83, p=.103). However, there was a significant Hz × Effector interaction, (F(2, 27.99)=4.65, p=.018). These results are summarized in the Table 3.

Table 2.2: Mean Time Difference: Detailed Effects.

Mean Time Difference: Detailed Effects

Effect	Estimate	SE	df	t	p
Intercept (2 Hz, Finger)	-0.0416	0.0044	27.98	-9.52	**< .001**
3 Hz vs 2 Hz (Finger)	0.0207	0.0034	28.00	6.05	**< .001**
4 Hz vs 2 Hz (Finger)	0.0321	0.0046	27.99	6.98	**< .001**
Vocal vs Finger at 2 Hz	0.0330	0.0048	27.99	6.94	**< .001**
(3 vs 2 Hz) × Effector	-0.0057	0.0041	28.00	-1.40	0.173
(4 vs 2 Hz) × Effector	-0.0239	0.0052	27.99	-4.58	**< .001**

Table 2.3: SD Time Difference: Fixed Effects.

SD Time Difference: Fixed Effects

Effect	Num df	Den df	F	p
Hz	2	27.98	28.63	**< .001**
Effector	1	27.99	2.83	0.103
Hz × Effector	2	27.99	4.65	**0.018**

For Finger responses, the estimated SD at 2 Hz was 0.0488 s ((SE=0.0045), (p<.001)). There was essentially no change from 2 to 3 Hz ((b=-0.0001), (SE=0.0038), (p=.971)), whereas SD increased by 0.0122 s from 2 to 4 Hz ((SE=0.0043), (p=.008)). At 2 Hz, the Vocal–Finger difference was not significant ((b=-0.0024), (SE=0.0048), (p=.618)).

The Effector effect changed significantly from 2 to 3 Hz (b=0.0123, SE=0.0040, p=.005). The corresponding change from 2 to 4 Hz did not reach conventional significance (b=0.0097, SE=0.0052, p=.075). These results are reported in the Table 4 and are visualized in the SD panel of the primary figure.

Table 2.4: SD Time Difference: Detailed Effects.

SD Time Difference: Detailed Effects

Effect	Estimate	SE	df	t	p
Intercept (2 Hz, Finger)	0.0488	0.0045	28.00	10.80	**< .001**
3 Hz vs 2 Hz (Finger)	-0.0001	0.0038	27.98	-0.04	0.97058
4 Hz vs 2 Hz (Finger)	0.0122	0.0043	27.98	2.85	**0.00818**
Vocal vs Finger at 2 Hz	-0.0024	0.0048	27.99	-0.50	0.61839
(3 vs 2 Hz) × Effector	0.0123	0.0040	27.99	3.03	**0.00518**
(4 vs 2 Hz) × Effector	0.0097	0.0052	27.99	1.85	0.07480

For the mean time difference, the overall averages fall below zero, indicating that responses generally occurred ahead of the stimulus. This is consistent with synchronization being predictive—participants anticipated the beats rather than reacting to them. Comparing effectors, the Vocal (tongue-clicking) mean was reliably closer to zero than the Finger mean (Effector, p <.001; Vocal-Finger at 2 Hz = +0.033 s, p <.001), meaning the response time was closer to the stimulus time for tongue clicking than for finger tapping. There was also a significant main effect of Hz (p <.001): the mean time difference moved toward zero from 2 to 4 Hz for both effectors. Several interpretations of this Hz effect are possible and we

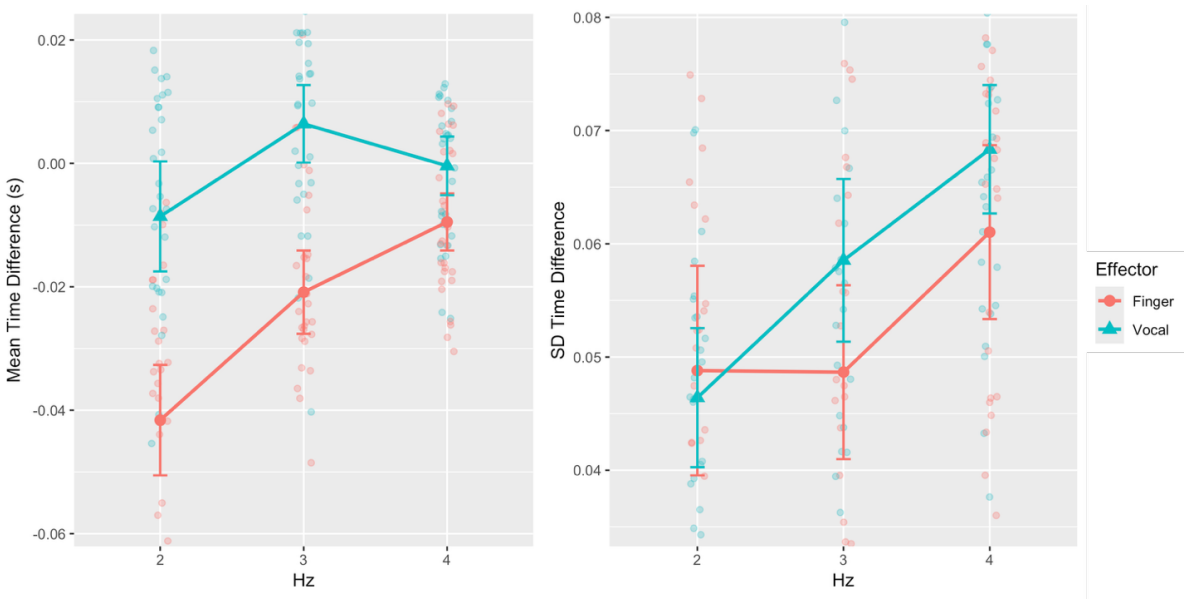


Figure 2.4: Comparison of mean and SD time differences across effectors and frequencies. (left) Mean time difference across 2, 3, and 4 Hz for Finger and Vocal conditions; (right) SD of time difference across the same conditions. Points and lines represent model-estimated marginal means from the linear mixed-effects models, and error bars represent 95% confidence intervals. Individual subject averages are shown as semi-transparent points. Only beats 4–8 were included in the analysis, based on the within-rhythm time-course pattern shown in Figure 2.4.

cannot distinguish them here. One possibility is a scaling artifact, since the absolute time difference is bounded by the shorter inter-beat intervals at higher rates; another is a genuine change in phase alignment at faster tempi. Standardizing the timing error by the inter-beat interval (i.e., expressing it as phase) would be needed to separate these, and we return to this in the Discussion. The interaction of effector and rate was also significant ($p < .001$): the Vocal advantage in the mean was present at the slower rates and narrowed at 4 Hz, and the Vocal curve showed a slight overshoot above zero at 3 Hz—an observation we note without over-interpreting, as its robustness is uncertain.

For the SD of the time difference, there was a significant main effect of Hz ($p < .001$)—variability increased at faster rates for both effectors—but the main effect of Effector was not significant ($p = .103$), and the Vocal-Finger difference at 2 Hz was negligible ($p = .618$). There was a significant Hz × Effector interaction ($p = .018$), but its component contrasts were mixed (2→3 Hz significant, 2→4 Hz not) and were not robust across alternative model specifications. We therefore do not interpret the SD interaction substantively. The most defensible reading of the SD results is simply that effector differences in variability were weak: on this measure, tongue clicking and finger tapping were comparable.

2.4 Discussion

2.4.1 The vocal effector passes the sanity check

Two points structure the interpretation. First, on the mean time difference, tongue clicking was reliably closer to the stimulus than finger tapping. Read at face value, this suggests we are at least as good at supralaryngeal (tongue) synchronization as at non-vocal (finger) synchronization, and if anything the phase alignment is tighter for the vocal effector. Second, on the variability (SD), the two effectors did not differ reliably. Together these bear directly on

the sanity check motivating the experiment: the vocal system is not worse than a non-vocal effector at rhythmic synchronization—its consistency is comparable and its mean alignment is, if anything, closer to the beat. This is what we would expect if rhythmic synchronization has its roots in speech, and in that sense the sanity check is passed. The premise that BPS is grounded in the speech system remains intact.

The mean advantage may reflect a measurement offset rather than better synchronization

It is worth being careful about what the mean advantage does and does not show. One possibility is that the vocal system genuinely aligns more precisely to the beat. But an alternative, which we cannot rule out here, is that part of the mean offset reflects differences in the effectors themselves rather than in synchronization ability—for example, the relationship between the moment of the motor act and the moment of the recorded acoustic onset may differ systematically between a tongue click and a finger tap, which would shift the mean without implying better synchronization. The fact that the two effectors did not differ reliably in variability is at least compatible with this reading: if the vocal system were wholesale better at synchronizing, one might have expected an advantage in consistency as well, whereas a difference confined to the mean is the kind of pattern a systematic timing offset could produce. We offer this only as one interpretation; disentangling a true alignment advantage from an effector-specific measurement offset would require controlling for the motor-to-acoustic delay of each effector.

2.4.2 Variability is comparable across effectors

Because the SD results are weak and not robust across models, we refrain from drawing strong conclusions from them. The safest statement is that variability was comparable across effectors—which, for the purposes of the sanity check, is itself informative: comparable con-

sistency means the vocal effector is not at a disadvantage in the stability of synchronization. Any tempo-dependent structure in the SD (e.g., variability rising with rate) is present for both effectors and, to the extent effector-specific differences appear at higher rates, may reflect biomechanical factors—the tongue being a soft, high-degree-of-freedom articulator that could be harder to keep consistent at fast repetition rates than the finger, which has a clear mechanical endpoint. We flag this only as a possibility, given the instability of the SD interaction.

2.4.3 The rate effect on the mean remains ambiguous

Regarding the effect of rate on the mean, we again note two possibilities without deciding between them. The movement of the mean toward zero at higher frequencies may be a scaling consequence of shorter inter-beat intervals, or it may reflect a genuine change in predictive phase alignment; the slight vocal overshoot at 3 Hz may be real or may be noise. Standardizing timing error as phase, or collecting additional data, would be needed before interpreting these tempo effects substantively.

2.4.4 Implications for the speech-origin premise

Returning to the question that motivated the experiment—how rhythmic synchronization comes to extend from vocal to non-vocal effectors—these behavioral results are most naturally accommodated by the picture in which the capacity has spread across the motor system as a whole, such that a non-vocal effector synchronizes about as well as a vocal one. We emphasize, however, that this is a sanity check on the shared speech-origin premise rather than an adjudication between Patel's and Hickok's accounts. The results confirm that we are on the right track—the vocal system is a competent, indeed comparable, synchronizer—and then set up the neuroimaging experiment that follows, which asks whether these two

effectors share a common neural substrate for auditory-motor synchronization, and points to the dPCSA as a candidate hub.

2.5 Future

Several analyses and design refinements follow from the present results. First, because the mean time difference is bounded by the inter-beat interval, the timing error should be standardized as phase (dividing by the interval) before the tempo effects are interpreted; this would help separate a genuine change in predictive phase alignment from a scaling artifact, and would put the three rates on a common footing. Second, the mean advantage for the vocal effector should be re-examined after controlling for the motor-to-acoustic delay of each effector—that is, the systematic relationship between the motor act and the recorded acoustic onset for a tongue click versus a finger tap—so that a true alignment advantage can be distinguished from an effector-specific measurement offset. Third, although the present design spans 2, 3, and 4 Hz, characterizing effector-specific preferred rates properly would require a wider and denser range of tempi together with spontaneous or continuation conditions (in which participants maintain a rate without an external pacing stimulus), since the current synchronization-to-metronome paradigm measures tracking accuracy rather than spontaneous preferred rate. Finally, the present stimuli were isochronous metronome sequences; extending the comparison to complex, musically structured rhythms would test whether the effector pattern observed here generalizes to full beat perception and synchronization, where a beat must be extracted from richer temporal structure. Collecting additional data would also help stabilize the weaker effects, particularly in the variability measure.

Chapter 3

Mapping the Neural Basis of Rhythmic Synchronization Across Effectors: An fMRI Comparison of Finger and Tongue Effectors

3.1 Introduction

As reviewed in the general background, while a number of brain regions have been implicated in rhythm entrainment, how each region functions and how they connect to one another remains unclear; there is as yet no settled account of the underlying circuitry. By bringing the behavioral task from Experiment 1 into the fMRI scanner, we aim to provide additional evidence about the regions and circuits involved in auditory-motor synchronization. This is an exploratory experiment.

A substantial body of work has examined finger tapping. Patel et al., for example, describe a

BPS network including dorsal premotor cortex, SMA, and the basal ganglia. Synchronization with the vocal tract, however, has not been studied. We therefore designed a task to compare limb control (finger tapping) against supralaryngeal control (tongue clicking) synchronizing to the same auditory rhythms.

The central question motivating this comparison is why humans are able to synchronize multiple effectors at all. If the capacity for auditory-motor synchronization originated in speech, why did it generalize beyond the vocal tract—why does the system also enable highly precise finger tapping, rather than wiring only into vocal effectors? One possibility is that the ability is not wired separately into each effector, but into a shared hub from which auditory-motor information can be routed to multiple effectors—a kind of multi-effector conductor. This logic follows from the demands of speech itself: coordinating the vocal tract requires synchronizing multiple effectors (larynx, supralaryngeal articulators, respiration) to a common target, and an ability grown large enough to serve these effectors may not be confinable to them, spreading instead across the motor system. If so, the neuroanatomy should reveal regions of overlap between effectors.

Accordingly, the primary aim of this experiment is to characterize the relationship between finger tapping and tongue clicking in the brain. Are they served by completely segregated circuits, or is there a shared component? We expect the two effectors to recruit distinct effector-specific regions—reflecting the well-established motor and somatosensory differences between moving the finger and the tongue—but our central interest is in the areas of commonality. A region engaged by both effectors would be a candidate hub for auditory-motor synchronization, and mapping the relationship between the speech coordination network and the beat-synchronization network across hand versus supralaryngeal effectors is a first step toward an integrated model.

We expect such a hub to emerge, and one candidate is the dorsal precentral speech area (dPCSA), or nearby premotor cortex—perhaps around the dPCSA or anterior to it. Notably,

the dorsal premotor region implicated in the beat-processing literature and the dPCSA of the speech model may correspond to overlapping cortex, a convergence this study is well positioned to examine. On the dual speech coordination model, the dPCSA receives strong auditory input and is involved in pitch-related motor planning. We further hypothesize that rhythmic synchronization may be a prerequisite for coordinating the two speech streams—the dorsal and ventral systems—enabling precise timing between the planning and articulation of prosodic, morphosyntactic, and segmental features; this aligns with the coordination account proposed by Hickok et al.. Against this backdrop, the present study asks a specific question: what role, if any, does the dPCSA play in rhythmic synchronization when the task involves no laryngeal or pitch-related control? This also connects to Patel's proposal that auditory-motor connectivity is strengthened in advanced vocal learners: since his account locates the strengthened pathway in laryngeal pitch control and holds that BPS arises as it spreads to other regions, dPCSA engagement by two non-laryngeal effectors would be consistent with his account. More broadly, we ask how these activation patterns relate to the surrounding architecture and to the existing literature.

As background for interpreting the activation patterns, prior work suggests provisional roles for several of the regions likely to be engaged: auditory cortex in coding the rhythmic pattern; premotor cortex in coding and coordinating movement plans for different effectors, organized roughly somatotopically; SMA in coordinating across effector-related premotor sub-regions; and the basal ganglia in action selection and reward-related prediction. Because rhythm is also a component of prosody, regions supporting prosodic control may be engaged as well. We treat these as interpretive expectations rather than confirmatory predictions, in keeping with the exploratory nature of the study.

3.2 Methods

3.2.1 Participants

Forty-four participants (n = 44; age mean 20 ± SD 1.93; all right-handed) took part in the study. Participants identified their gender as female (n = 27), male (n = 15), agender (n = 1), or preferred not to disclose (n = 1). All subjects have normal hearing, no history of neurological or psychiatric disorder, and they all passed MRI safety screening. All participants provided written informed consent, and the study was approved by the Institutional Review Board of the University of California, Irvine.

3.2.2 Experimental task and stimuli

Based on the preliminary results from Experiment 1, subjects are overall performing stable synchronization to the external stimuli after the 4th beats in the sequence (Figure 2.4). Therefore, to get more data within limit amount of time, we revise the experimental design from Experiment 1 as follows.

The auditory stimuli were still a sequence of snare drum sound with dynamic design. There are still two main effector conditions, one requires the subject to finger tap along with the auditory rhythm, the other requires the subject to tongue click to the rhythm. We are still going to use microphone to record subjects' responses, however a noise canceling optical microphone compatible in the MRI scanner (Figure 2.4).

There were three different rates of rhythms the same as the behavioral experiments – 2 Hz, 3 Hz and 4 Hz. And each rhythm is a metronome sequence with the number of beast (snare drum sound) varied from 6, 8 and 10, which is different from the behavioral settings. We randomized the total nine combinations, which will cost 26 seconds in total to run all

combination once, and it is set as a block.

Rather than letting the subjects continue after a sequence by themselves as in the behavioral setting, there is a forced 15-second break after each block, to be easily controlled along with the scanner. Eight blocks together as one run. The subjects can rest between runs. There will be three runs for finger tapping, and three other runs for tongue clicking, and in total six runs for each subject. The random sequence will be counterbalanced between subjects.

Participants performed a beat-synchronization task in which they were instructed to produce a movement in time with an auditory beat whenever they heard a drum sound. The task was implemented in MATLAB using Psychtoolbox (PsychPortAudio). Two effectors were tested in separate runs: in *finger* runs, participants tapped their fingers on a hard surface next to the body near a microphone; in *tongue* runs, participants produced tongue clicks. Effector was therefore blocked by run rather than intermixed within a run.

Auditory stimuli were snare-drum sounds presented at three tempi, defined by the inter-onset interval (stimulus-onset asynchrony) between successive beats: 250, 333, and 500 ms, corresponding to 4, 3, and 2 Hz, respectively. Within each block, tempo conditions were combined with three repetition counts (6, 8, or 10 beats), yielding nine tempo $\times$ repetition combinations that were presented in a randomized order with the constraint that the same pairing did not repeat consecutively. Each block was preceded and followed by 1 s of silence. The snare-drum sound was created in GarageBand and edited in Audacity. Stimuli were delivered at 44.1 kHz via a Focusrite audio interface through 16-bit digital-to-analog converters (Focusrite Scarlett 2i2) and Optoacoustics OptoACTIVE II. Subjects' responses were recorded using a MRI compatible button box, and a Optoacoustics FOMRI microphone and calibrated to a fixed presentation level.

Each run comprised 8 blocks, with blocks initiating every 32 s. Runs were synchronized to the MRI scanner via a trigger pulse, and data acquisition began after an initial 7 s

delay following the trigger to allow for magnetization steady state. For finger runs, tap times were recorded from keypress events; for tongue runs, vocal responses were recorded via microphone. Scanner trigger times and block-onset timestamps were logged on every run and used to reconstruct stimulus and response timing in scanner time for subsequent modeling. Each participant completed 6 task runs.

3.2.3 MRI data acquisition

Imaging data were acquired on a 3 T Siemens MAGNETOM Prisma scanner using a 32-channel head coil. Functional images were acquired with a 2D multiband (simultaneous multi-slice) gradient-echo echo-planar imaging (EPI) sequence (TR = 1500 ms, TE = 52 ms, flip angle = 45°, voxel size = 2.5 × 2.5 × 2.5 mm with a 0.25 mm inter-slice gap (2.75 mm slice spacing), 52 slices, matrix = 84 × 84, multiband acceleration factor = 4, phase-encoding direction A→P; 185 volumes per run). A high-resolution T1-weighted anatomical image was acquired using an MPRAGE sequence (TR = 2300 ms, TE = 2.32 ms, TI = 900 ms, flip angle = 8°, voxel size = 0.9 × 0.94 × 0.94 mm, matrix = 192 × 256 × 256). For susceptibility-distortion correction, spin-echo EPI field maps were acquired with opposing phase-encoding directions (blip-up/blip-down, PEPOLAR).

3.2.4 Image preprocessing

Raw data quality was first assessed with MRIQC 24.0.2. Results included in this manuscript come from preprocessing performed using fMRIPrep 25.0.0 **(Esteban et al., 2019); (Esteban, Blair, Markiewicz, et al., 2018)**; RRID:SCR_016216, which is based on Nipype 1.9.2 **(K. Gorgolewski et al., 2011); (K. J. Gorgolewski, Esteban, Markiewicz, et al., 2018)**; RRID:SCR_002502.

Anatomical data preprocessing

The T1-weighted (T1w) image was corrected for intensity non-uniformity with N4BiasFieldCorrection ((**Tustison et al., 2010**)), distributed with ANTs 2.5.4, and used as T1w-reference throughout the workflow. The T1w-reference was skull-stripped with a Nipype implementation of the antsBrainExtraction.sh workflow, using OASIS30ANTs as the target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white matter (WM), and gray matter (GM) was performed on the brain-extracted T1w using FSL fast. Brain surfaces were reconstructed using FreeSurfer 7.3.2 recon-all, and a T2-weighted image was used to improve pial surface refinement. Volume-based spatial normalization to standard space (MNI152NLin6Asym) was performed through nonlinear registration with antsRegistration (ANTs 2.5.4), using brain-extracted versions of both the T1w reference and the T1w template. The MNI152NLin6Asym template was accessed via TemplateFlow ((**Ciric et al., 2022**)).

Functional data preprocessing

For each BOLD run, a reference volume was generated using fMRIPrep for head-motion correction. Head-motion parameters (six rotation and translation parameters) were estimated with FSL mcflirt before any spatiotemporal filtering. Susceptibility-distortion correction was performed using a B0 field map estimated from two EPI references with opposing phase-encoding directions (FSL topup). The BOLD reference was co-registered to the T1w reference using FreeSurfer bbregister (boundary-based registration) with six degrees of freedom. All resamplings were performed in a single interpolation step by composing head-motion transforms, susceptibility-distortion correction, and co-registrations. Functional data were resampled into MNI152NLin6Asym space at 2 mm isotropic resolution (*res-2*) and onto the fsnative surface. Confound time series (framewise displacement, DVARS, and CompCor components) were computed but were not used in the present analysis, which relied on

ICA-AROMA denoising (below). Many internal operations of fMRIPrep use Nilearn 0.11.1 ((**Abraham et al., 2014**)).

Structured noise was then removed using ICA-AROMA as implemented in fMRIPost-AROMA 0.0.12, applied with the non-aggressive denoising strategy, which regresses out the variance uniquely associated with motion-related independent components while preserving variance shared with the signal of interest.

The denoised functional data were then spatially smoothed and scaled using AFNI (*version 25.1.15*). Smoothing was performed with 3dBlurToFWHM to a target smoothness of 6 mm FWHM within each run's brain mask. Each run was then converted to units of percent signal change by dividing the smoothed time series voxel-wise by its temporal mean and multiplying by 100 (3dTstat followed by 3dcalc, expression 100·a/b). These denoised, smoothed, and scaled run-level images (desc-denoisedSmoothedScaled) were the input to the first-level general linear model.

For each participant, a whole-brain analysis mask was defined as the union of the six run-specific fMRIPrep brain masks (3dcalc, step(a+b+c+d+e+f)), restricting analyses to voxels covered in at least one run.

3.2.5 First-level (single-subject) analysis

First-level general linear models were estimated for each participant with AFNI 3dDeconvolve. All models included fourth-order (0–3) Legendre polynomial regressors per run to remove low-frequency drift (-polort 3), used local stimulus timing (-local_times), and were restricted to the participant's union brain mask. Stimulus events were modeled as sustained blocks whose duration matched the stimulus, convolved with a canonical hemodynamic response using AFNI's duration-modulated block model (dmUBLOCK). Stimulus timing files

were derived from the behavioral recordings acquired during scanning (scanner trigger times and block-level onsets converted to run-relative scanner time).

Modality model

The modality model included two regressors, one for finger blocks and one for tongue blocks. A general linear test contrast (-gltsym) was defined for the effector difference, Finger - Tongue, together with the two individual effector estimates (Finger, Tongue).

3.2.6 Group-level analysis

For each contrast, single-subject effect estimates (coefficients) were carried to the group level and tested against zero across the 44 participants using a one-sample t-test in AFNI 3dttest++. Group analyses were restricted to a group mask defined as the strict intersection (100% of participants; 3dmask_tool -frac 1.0) of the individual union masks, comprising 212,560 voxels.

Correction for multiple comparisons was performed using AFNI's ETAC (Equitable Thresholding and Clustering) permutation-based method, as implemented in 3dttest++ (-ETAC). ETAC was run two-sided (sid=2) across a set of voxel-wise thresholds (p = 0.05, 0.02, 0.01, 0.005) and two cluster-forming height powers (hpow = 0, 2), with the global false-positive rate controlled at 5% (fpr=5). Voxels surviving the ETAC-corrected threshold are reported.

3.2.7 Activation tables and Visualization

For each corrected group map, clusters were identified and local peaks extracted using AFNI 3dClusterize (NN=2 neighborhood; minimum cluster extent 20 voxels), restricted to the

ETAC-corrected voxels. For each cluster we report the cluster extent (number of voxels), the peak coordinate in MNI space (LPI convention), and the peak Z statistic.

Group statistical maps were visualized in Python using Nilearn 0.14.0. Volume results were displayed as axial slice montages and were additionally projected onto the fsaverage cortical surface for display; the color scale represents the group-level Z statistic, and only ETAC-corrected voxels are shown.

3.2.8 Software

MRIQC 24.0.2, fMRIPrep 25.0.0, fMRIPost-AROMA 0.0.12, AFNI 25.1.15, and Nilearn 0.14.0. All neuroimaging containers were run via Singularity/Apptainer on a high-performance computing cluster.

3.3 Results

Group-level activation maps were computed for three modality contrasts: finger tapping vs. baseline, tongue clicking vs. baseline, and the direct comparison between the two (Finger - Tongue). All results are reported after ETAC correction (two-sided, fpr = 5%), N = 44 (Figures 3.1–3.3; Table 5).

For the direct Finger - Tongue contrast, finger tapping elicited greater activation than tongue clicking in a more dorsal precentral region (more dorsal than dPCSA), peaking in left motor cortex (peak Z = 7.71, MNI -38, -24, 54; k = 4,133 voxels). Conversely, tongue clicking elicited greater activation than finger tapping in ventral precentral cortex (vPCSA) bilaterally (left: peak Z = -7.33, MNI -52, -10, 28, k = 11,790; right: peak Z = -7.09, MNI 54, -8, 26, k = 4,763).

Comparing the single-effector maps, both finger tapping and tongue clicking recruited overlapping regions including dPCSA and the SMA, indicating common engagement across the two effectors.

Table 3.1: Peak coordinates of significant clusters for the three modality contrasts (Finger - Tongue, Finger, and Tongue). For each cluster, the table lists the cluster extent (k, number of voxels), the peak coordinate in MNI space (x, y, z; mm), and the peak Z statistic. Positive Z values (warm shading) denote finger >tongue for the Finger - Tongue contrast, and positive (task-positive) responses for the single-effector contrasts; negative Z values (cool shading) denote the reverse. All clusters survived ETAC correction (two-sided, false-positive rate = 5%). Anatomical labels are not reported.

Group activation table (Z, ETAC-corrected, N=44)

Contrast	Cluster	k (vox)	x	y	z	peak Z
Finger-Tongue	1	11,790	-52	-10	28	-7.33
Finger-Tongue	2	4,763	54	-8	26	-7.09
Finger-Tongue	3	4,162	10	-48	-14	+5.25
Finger-Tongue	4	4,133	-38	-24	54	+7.71
Finger-Tongue	5	2,798	-54	-24	10	+5.23
Finger-Tongue	6	2,501	-12	-58	-22	-6.18
Finger-Tongue	7	522	32	-4	-6	-6.12
Finger-Tongue	8	360	28	-44	4	+4.10
Finger-Tongue	9	278	12	-74	-40	-3.17
Finger-Tongue	10	68	-18	-18	6	+5.41
Finger	1	99,758	-26	-44	66	-7.82
Finger	2	853	-66	-48	34	-5.64
Finger	3	543	50	6	8	+4.00
Finger	4	429	-48	-68	-6	-3.47
Finger	5	237	24	-0	10	+4.89
Tongue	1	82,169	-26	-44	64	-7.34
Tongue	2	1,640	-30	32	32	-4.84
Tongue	3	819	56	-8	26	+5.46
Tongue	4	658	-6	-4	66	+5.40
Tongue	5	499	-24	-66	-52	+6.10
Tongue	6	422	-50	-60	-2	-3.58
Tongue	7	323	-58	-16	-22	+3.51

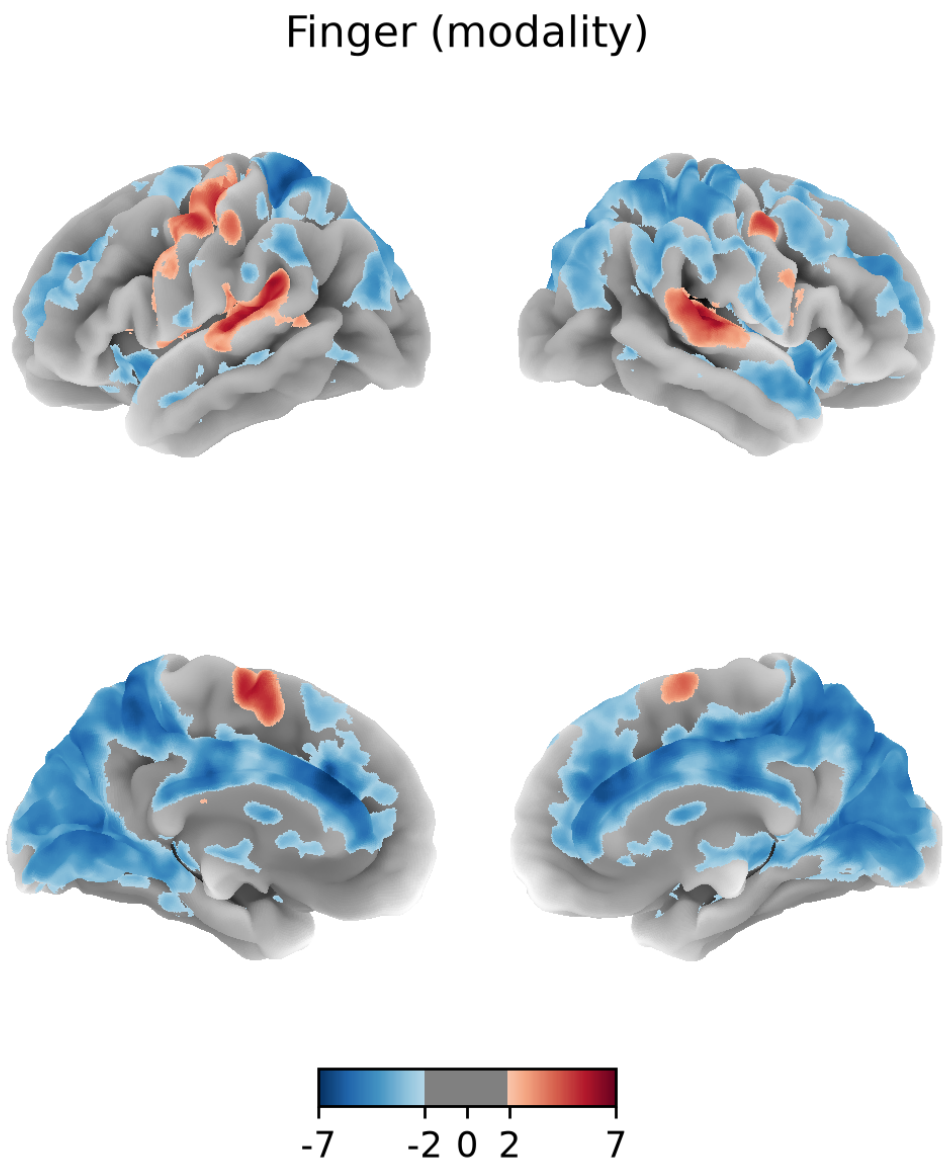


Figure 3.1: Group-level statistical map for the finger-tapping condition (Finger vs. baseline), projected onto the fsaverage cortical surface (lateral and medial views of both hemispheres). Color represents the group Z statistic; warm colors indicate positive (task-positive) and cool colors negative (task-negative/deactivation) responses. Only voxels surviving ETAC correction (two-sided, false-positive rate = 5%) are shown.

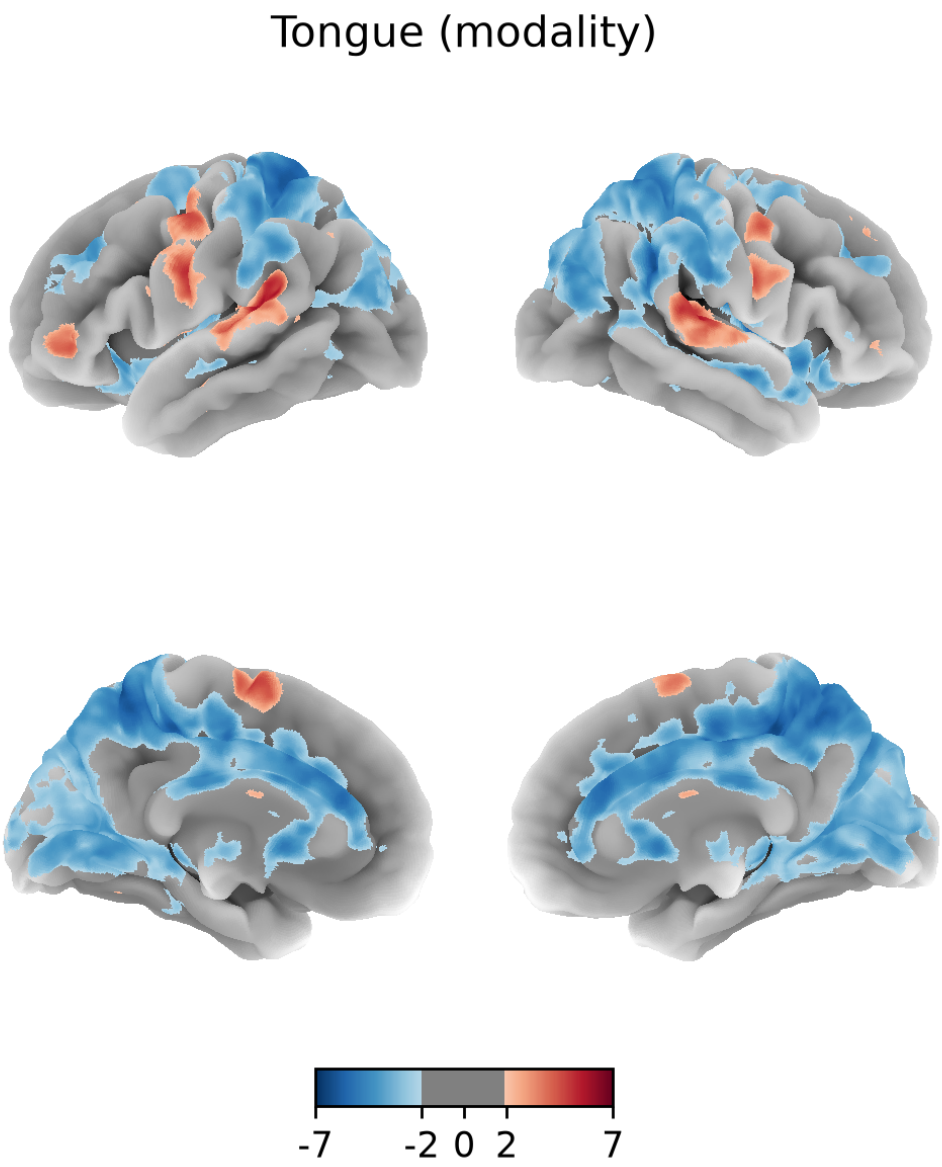


Figure 3.2: Group-level statistical map for the tongue-clicking condition (Tongue vs. baseline), projected onto the fsaverage cortical surface (lateral and medial views of both hemispheres). Color represents the group Z statistic; warm colors indicate positive (task-positive) and cool colors negative (task-negative/deactivation) responses. Only voxels surviving ETAC correction (two-sided, false-positive rate = 5%) are shown.

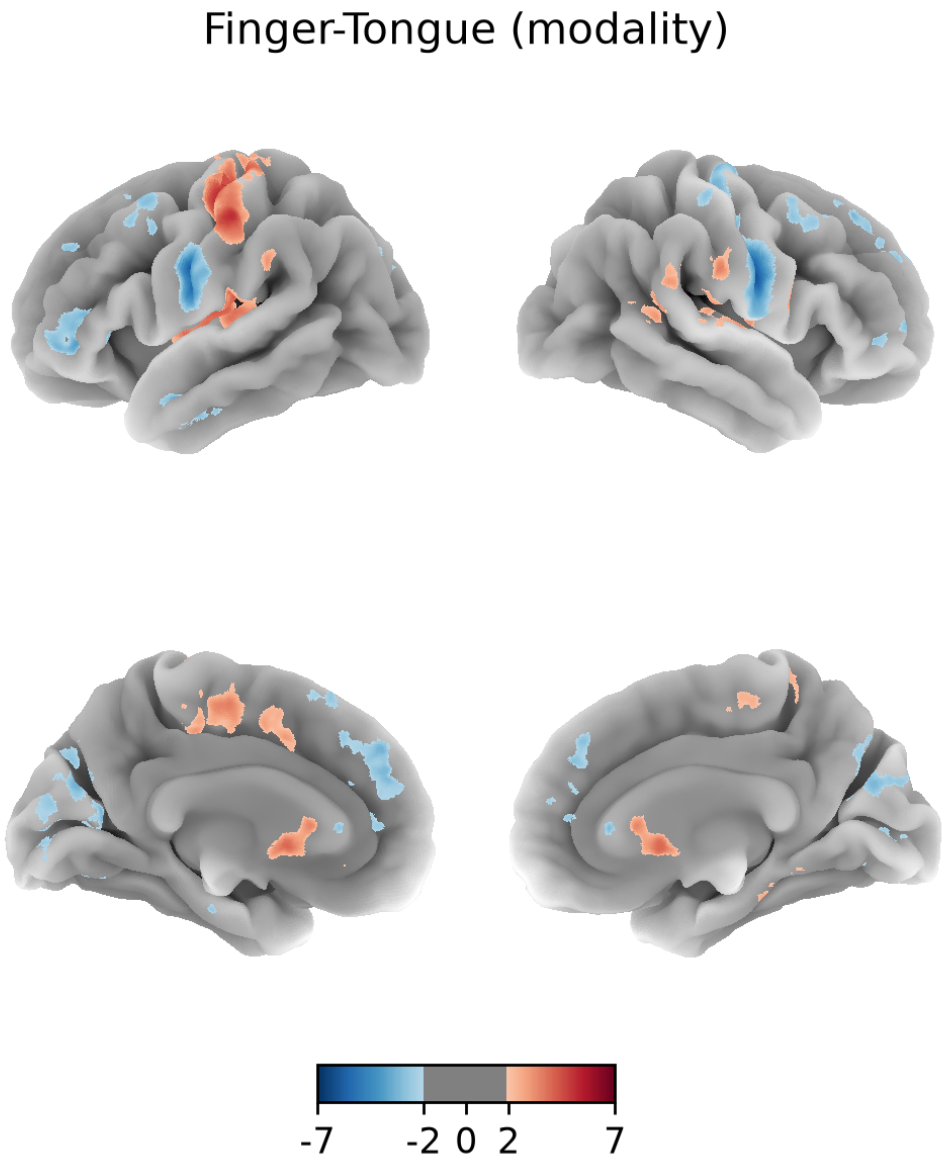


Figure 3.3: Group-level statistical map for the direct comparison between effectors (Finger - Tongue), projected onto the average cortical surface (lateral and medial views of both hemispheres). Warm colors indicate greater activation for finger tapping than tongue clicking; cool colors indicate greater activation for tongue clicking than finger tapping. Color represents the group Z statistic, and only voxels surviving ETAC correction (two-sided, false-positive rate = 5%) are shown.

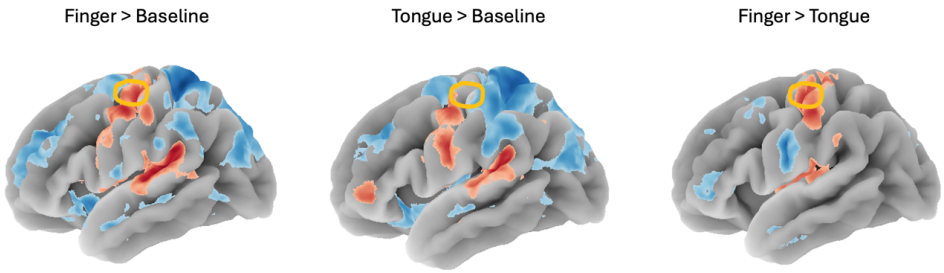


Figure 3.4: Group-level surface maps (left hemisphere, lateral view) for finger tapping (Finger >Baseline), tongue clicking (Tongue >Baseline), and their direct comparison (Finger >Tongue). Color represents the group Z statistic; warm colors indicate positive responses and cool colors task-negative (deactivation) responses, and for the Finger >Tongue map warm and cool colors indicate greater activation for finger and tongue, respectively. Only voxels surviving ETAC correction (two-sided, false-positive rate = 5%) are shown. The yellow circle highlights the finger movement-related area.

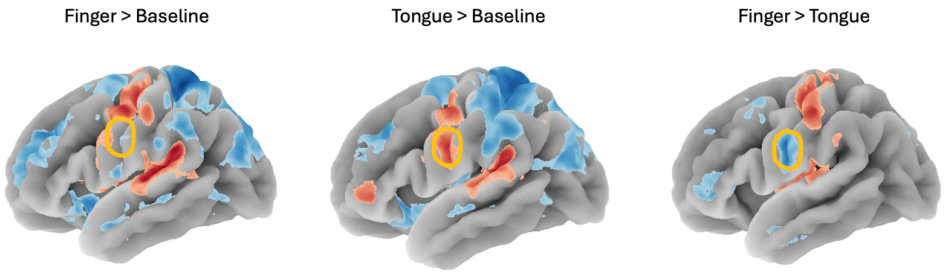


Figure 3.5: Group-level surface maps (left hemisphere, lateral view) for finger tapping (Finger >Baseline), tongue clicking (Tongue >Baseline), and their direct comparison (Finger >Tongue). Color represents the group Z statistic; warm colors indicate positive responses and cool colors task-negative (deactivation) responses, and for the Finger >Tongue map warm and cool colors indicate greater activation for finger and tongue, respectively. Only voxels surviving ETAC correction (two-sided, false-positive rate = 5%) are shown. The yellow circle highlights the tongue movement-related area.

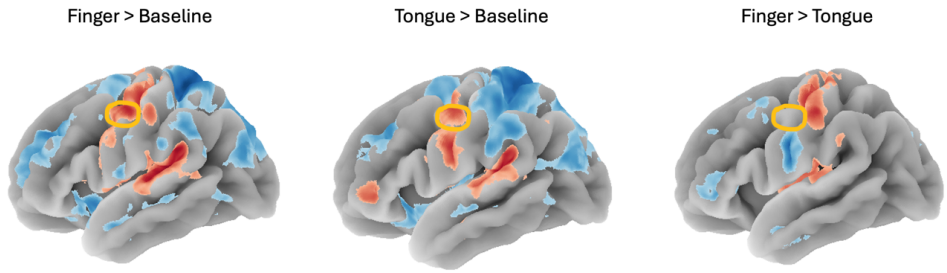


Figure 3.6: Group-level surface maps (left hemisphere, lateral view) for finger tapping (Finger >Baseline), tongue clicking (Tongue >Baseline), and their direct comparison (Finger >Tongue). Color represents the group Z statistic; warm colors indicate positive responses and cool colors task-negative (deactivation) responses, and for the Finger >Tongue map warm and cool colors indicate greater activation for finger and tongue, respectively. Only voxels surviving ETAC correction (two-sided, false-positive rate = 5%) are shown. The yellow circle highlights the potential dPCSA.

3.4 Discussion

3.4.1 dPCSA is shared across effectors

The central result of this experiment concerns what finger tapping and tongue clicking have in common. Although the two effectors recruited distinct effector-specific regions—finger tapping producing greater activation dorsal to the dPCSA, and tongue clicking producing greater activation in the vPCSA—the two tasks converged on a shared set of regions, most notably the dorsal precentral speech area (dPCSA) and the supplementary motor area (SMA). In other words, the dPCSA does not appear to be modality-specific; it is engaged regardless of which effector is used to synchronize to the beat, suggesting it is doing something more general than controlling any one effector.

3.4.2 The dPCSA is multifunctional

This shared engagement stands in tension with the role we assigned to the dPCSA on the basis of the dual speech coordination model. In that model, the dPCSA is characterized as an auditory-weighted region involved in pitch-related auditory-motor planning. Yet in the present task there was no laryngeal or pitch-related motor planning at all: participants performed finger tapping and tongue clicking, neither of which involves glottal or pitch control. If the dPCSA were purely a pitch-coordination region, there would be no obvious reason for it to activate here.

We therefore propose that the dPCSA is multifunctional. This does not require abandoning its role in pitch-related planning; rather, the region may serve an additional function—a more general auditory-motor rhythm synchronization function—alongside its pitch-related one. On this view the dPCSA is not doing one thing or the other, but both.

This possibility is directly testable in future work. One could contrast a pitch-related control task with the present rhythm-synchronization task and ask whether they engage an identical pattern within the dPCSA, or whether sub-regions are preferentially engaged by one function versus the other. Multivariate pattern classification, in particular, could reveal whether distinct sub-regions within the dPCSA are specialized for these different functions even where the univariate activation overlaps.

3.4.3 Why the dPCSA can serve as a hub: an auditory region within the motor system

What, then, makes the dPCSA well suited to this general synchronization role? A clue comes from what is common across all of these tasks—including a hypothetical pitch-control task—namely the auditory input. The shared activation we observe may reflect, at its core, the arrival of auditory information: the same auditory stimulus was present in both effector conditions.

Independent evidence indicates that the dPCSA is unusually sensitive to acoustic input on its own. Venezia, Richards, and Hickok (2021), mapping speech-driven spectrotemporal receptive fields (STRFs) beyond auditory cortex, found that the dPCSA has mappable STRFs resembling those of primary auditory cortex—an auditory profile that is rare for premotor cortex, and for the frontal lobe more generally. The dPCSA thus behaves, in part, like an auditory region embedded within the motor system.

This reframes what "auditory-motor synchronization" means here. To synchronize a motor effector to an auditory rhythm, the system needs auditory input to reach the motor system—and that input appears to arrive via the dPCSA. In the present study this auditory information is not being routed to laryngeal control (that task was not performed); instead it is being made available to two other, effector-specific motor systems (finger and tongue). One

way to describe this is that auditory information reaches the frontal motor system through the dPCSA, and is then used by the individual effectors. This is the basis on which the dPCSA can act as a hub: it supplies the common auditory input that the effector-specific regions do not appear to receive directly. Consistent with this, auditory activation in the precentral gyrus is generally not seen in ventral or dorsal (e.g., finger) zones, but in this middle zone corresponding to the dPCSA.

3.4.4 A proposed model

These observations can be summarized in a simple schematic that adapts the format of the dual speech coordination model to the present findings. Rather than emphasizing the inputs to the vPCSA (the focus of the original figure), the proposed model centers on the dorsal stream: auditory cortex projects to the dPCSA, and the dPCSA in turn projects to the effector-specific regions—the finger region and the tongue region—localized from our own activation maps (Figure 3.7). The pathway of communication thus runs from auditory cortex to the dPCSA, and from the dPCSA to these two effector systems, clarifying the sense in which the dPCSA serves as a hub for synchronization across effectors.

Why the dPCSA should be accessible to these other motor systems is not something the present data can resolve. We do not currently have a good explanation for why auditory information routed through this region should be available to multiple effectors, and we flag this as an open question for future work rather than a settled claim.

3.4.5 SMA is also shared, but is unlikely to be the auditory hub

The dPCSA is not the only region common to both tasks. The SMA was also engaged by both finger and tongue synchronization, which is consistent with existing accounts—including

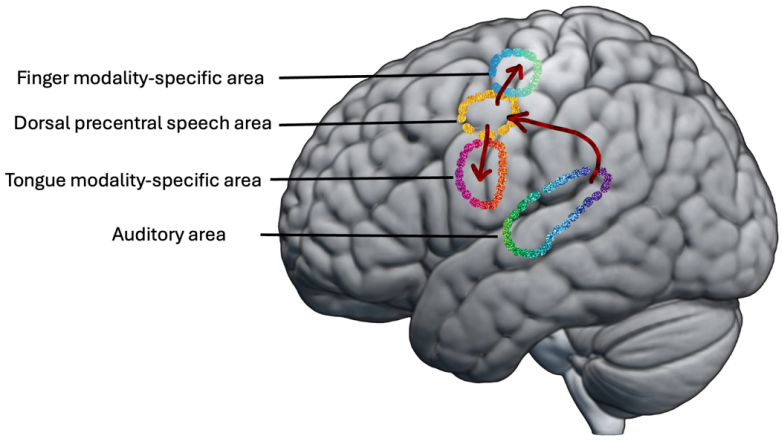


Figure 3.7: Schematic depiction of proposed models.

Patel's—that include the SMA as part of the BPS network.

This raises a natural question: if the SMA is also common to both effectors, why identify the dPCSA rather than the SMA as the auditory-motor hub? The distinction rests on the sensory properties of the two regions. The SMA is generally associated with motor execution and motor initiation, and does not show the auditory perceptual sensitivity that the dPCSA does. It is therefore unlikely to be the entry point for auditory input. The dPCSA, by contrast, combines a motor-system location with a distinctly auditory response profile. On this reading, both regions are shared, but they contribute differently: the SMA reflects a more general motor function, whereas the dPCSA reflects a more specifically auditory-motor one. The detailed role of the SMA is beyond the scope of the present work.

3.4.6 Additional observations

Area Spt. Because area Spt is the posterior temporal source in this circuit, it would be expected to be engaged similarly by both effectors, and indeed Spt activation—at the posterior-

most extent of the finger and tongue activation in the STG—was present in both conditions and absent in the finger–tongue contrast. This is consistent with Spt serving as the point of connectivity between the dPCSA and the auditory system.

One incidental asymmetry is that auditory cortex was more strongly engaged by finger tapping than by tongue clicking. One speculative interpretation is task difficulty: if this circuit evolved for, and is used continually in, speech, then vocal-tract synchronization may be more practiced and require less auditory-driven effort, whereas finger tapping—though a task we perform well—is less habitual and may drive more auditory activity to track the beats.

Laterality. We focus on the left hemisphere for two reasons. First, the circuit appears to be at least somewhat left-dominant, being slightly more active on the left in both conditions. Second, and more importantly, the two effectors have different intrinsic lateralization: right-hand finger tapping strongly activates the left hemisphere, whereas tongue control is more bilateral. Focusing on the hemisphere that contains both effects strongly makes for a fairer comparison. That said, the dPCSA itself appears to be bilateral: on the right, both effectors produce a dPCSA response with no reliable difference in the contrast, and the clearest contrast effect on the right is simply greater ventral (tongue-control) activation for tongue. The absence of a right dorsal finger response is expected given that finger control is crossed.

3.4.7 Relation to Patel's hypothesis and to the dual-stream model

These findings bear on Patel's proposal that auditory-motor connectivity is strengthened in advanced vocal learners and thereby enables rhythmic synchronization. If the dPCSA is the region through which auditory information supports vocal-learning-related auditory-motor function—by virtue of the strong auditory input it receives—then it is a natural candidate for a hub that supports auditory-motor synchronization across the motor system more generally.

The present design tests this by using two separate, non-laryngeal effectors: since Patel's hypothesis emphasizes that the laryngeal pitch-control pathway is strengthened and then spreads to other regions to enable BPS, the observation that the dPCSA is nonetheless activated by both non-laryngeal tasks is consistent with his account.

Although it is not the main goal of this experiment, the effector-specific results also provide a secondary check on the dual speech coordination model: activation dorsal to the dPCSA for limb (finger) control, and vPCSA activation for supralaryngeal (tongue) control, align with the expected effector organization.

Finally, a note on the tempo manipulation. The three rates (2, 3, 4 Hz) were included partly to allow us to examine error-related signals when the rhythm changes and motor plans must be updated, and partly because finger and tongue may have different preferred rates. Neither question is analyzed here; having ported the manipulation into the scanner, we are focusing on the effector comparison for now and reserving the tempo analyses for later work.

3.5 Future

Building on the multifunctionality proposal outlined above, a direct next step is to contrast a pitch-related control task with the present rhythm-synchronization task within the dPCSA, using multivariate pattern classification to test whether distinct sub-regions support pitch-related versus general auditory-motor synchronization functions even where univariate activation overlaps.

Beyond the effector main-effect reported here, the present dataset supports several analyses that we reserve for future work. First, we will examine the effector × frequency interaction—whether the finger–tongue difference varies across the three tempi (2, 3, and 4 Hz)—which speaks to the possibility that finger and tongue have different preferred rates arising from

biomechanical differences. Second, exploiting the dynamic design, we will examine the transition stages between rates (2→3, 2→4, 3→4 Hz, and the reverse), which were included specifically to isolate the error- and updating-related signals that arise when an ongoing rhythm changes and motor plans must be revised; comparing these transitions may reveal graded levels of the same activation.

A further limitation concerns the stimuli themselves. The present task used isochronous metronome sequences in order to isolate the basic auditory-motor synchronization mechanism from the additional processing required to extract a beat from complex, hierarchically structured music. The conclusions reported here therefore strictly concern simple beat synchronization. An important next step is to repeat the key analyses with complex, musically structured rhythms to test whether the dPCSA hub generalizes to full beat perception and synchronization, where beat extraction is required.

At the network level, generalized linear models will be used for statistical testing, and where regions are found to covary, stimulus-modulated functional connectivity analysis will be performed to assess how temporal correlations between regions are modulated by the stimulus. Because activation alone cannot establish that the dPCSA serves as the upstream source of auditory input, directed or effective-connectivity approaches will be needed to test the proposed auditory-cortex → dPCSA → effector-region pathway more directly than the activation-based results reported here.

Finally, although the difference between musicians and non-musicians is not a primary focus, participants completed a background questionnaire, allowing a later comparison of their activation patterns. Based on prior work, we would tentatively expect increased but comparable activation in the basal ganglia and SMA across groups, with musicians showing stronger modulation of auditory–motor coupling (STG with PMC/SMA) around the points where the rhythm changes. Future analyses incorporating these survey data will test whether brain activation differs between musicians and non-musicians, and, more specifically, whether the

length or type of musical training modulates the auditory-motor synchronization network.

Chapter 4

Perceiving Rhythm Without Sound: A Forward-Entrainment Test of the Motor Theory of Rhythm

4.1 Introduction

The first two experiments of this dissertation concern the evolutionary source of beat perception and synchronization (BPS): the behavioral experiment indicated that the vocal system is at least as capable of rhythmic synchronization as a non-vocal effector, consistent with the idea that the capacity is grounded in speech, and the neuroimaging experiment pointed to the dorsal precentral speech area as a substrate shared across effectors. Those experiments bear on why the capacity exists and how it comes to be available across the motor system. They leave open a distinct and more mechanistic question: when a listener perceives a rhythm, is the motor system actually carrying out the perception? That is, is motor simulation the online mechanism of rhythm perception, or is it only part of its evolutionary background?

The present experiment addresses this question.

The motor theory of rhythm holds that perceiving a rhythm depends on covertly simulating it motorically. On this view, promoted most prominently by Patel and Iversen's Action Simulation for Auditory Prediction (ASAP) hypothesis, the motor planning system entrains its activity to the beat period and sends predictive signals to auditory regions, thereby shaping the perceptual interpretation of the rhythm (Patel & Iversen, 2014). If this is correct, then merely listening to a rhythm should automatically engage the motor system, and that engagement should have measurable perceptual consequences. The forward-entrainment paradigm offers an unusually direct way to probe this claim.

Forward entrainment refers to that part of behavioral or neural entrainment that outlasts the entraining stimulus, by analogy to forward masking in psychoacoustics, in which masking effects persist after the masker has terminated (Saberi & Hickok, 2023). It has been observed across a range of psychophysical methods, target stimuli, modalities, and species. Here we focus on one well-characterized auditory instance. In the signal-detection study of Hickok et al. (2015), listeners heard a wideband noise that was sinusoidally amplitude modulated (AM) at 3 Hz for the first few seconds of each trial and then continued as steady-state, unmodulated noise; a brief near-threshold tone was presented at one of several temporal positions during the unmodulated portion, and listeners reported whether it had occurred. Detectability oscillated at the frequency of the terminated modulator and persisted for roughly two cycles after the modulation ended (Farahbod, Saberi, & Hickok, 2020; Hickok et al., 2015). The oscillation was anti-phasic: performance was best at the temporal positions where the troughs of the driving modulator would have fallen had the modulation continued (Figure 4.1).

A later study extended this to a range of modulation rates and found significant entrainment at 2 and 3 Hz, a weaker effect at 5 Hz, and no effect at rates from 8 to 32 Hz—that is, forward entrainment is rate-limited and lowpass in nature (Farahbod et al., 2020).

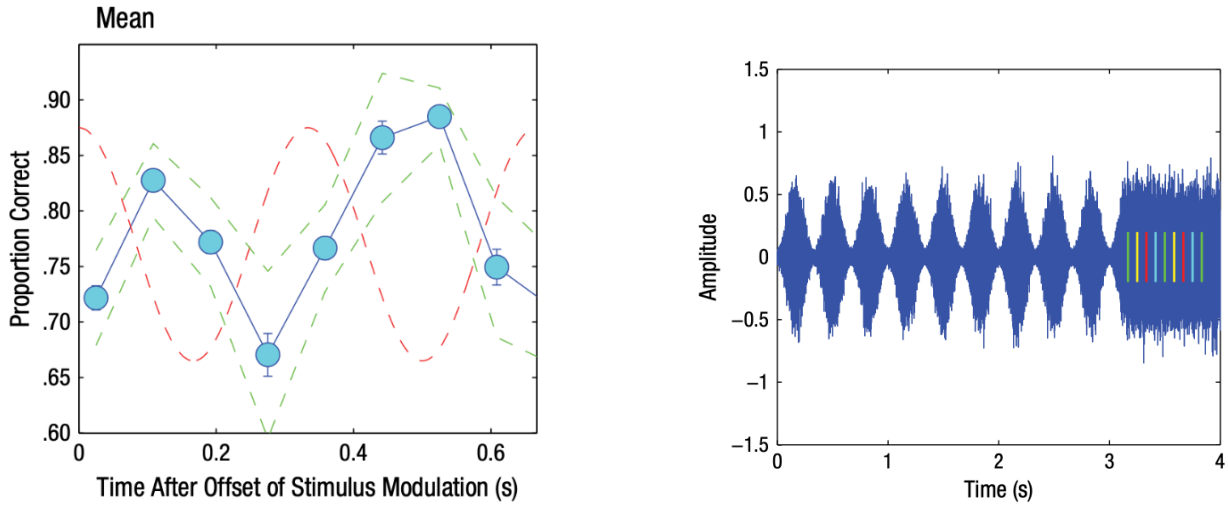


Figure 4.1: Figures from Hickok et al. (2015). (left) Stimulus used in the experiment; vertical lines of different colors indicate nine temporal positions. (right) Mean proportion correct as a function of time after offset of stimulus modulation.

Two broad classes of explanation have been offered for the anti-phasic "listening-in-the-dip" pattern. The first is attentional. Because the noise is what the listener is trying to avoid, an ideal observer implicitly directs attention to the expected dips of the masker, where the signal-to-noise ratio would be most favorable; on this account, forward entrainment is largely attention driven, even if implicitly so, rather than the product of a bottom-up neurophysiological mechanism (Saberi & Hickok, 2023). The second, and the hypothesis examined here, is motor. This account is grounded in motor-control theory: when the motor system generates a plan, it issues an efference copy to the sensory system, producing a forward prediction of the sensory consequences of the action. Where the predicted and actual signals match, the sensory response tends to be suppressed rather than enhanced, because the function of this comparison is error detection—an accurately predicted event carries no error and can be cancelled (Hickok, 2014). A familiar analogue is saccadic suppression, in which self-generated eye movements attenuate the associated visual signals so that the world is perceived as stable (Matin, 1974).

Applied to forward entrainment, this account offers a reinterpretation of the effect. What appears to be enhanced perception in the dips may instead be ordinary perception against

a background of motor-generated suppression at the peaks: on hearing the beat in the AM noise, the brain generates a motor plan and an accompanying sensory prediction, which suppresses perception at the expected beat positions and thereby produces the anti-phasic pattern (Hickok, 2014). It has further been suggested that, if the effect is motor in origin, the lowpass character of forward entrainment might follow from the fact that motor effectors have preferred rhythmic rates in the low-frequency range; we note this as a conjecture that stands or falls with the motor account itself, rather than as independent support for it.

These two accounts differ in where they locate the origin of the effect, but the standard auditory paradigm cannot separate them, because in every case the rhythm is delivered to the auditory system. The purpose of the present experiment is to remove the auditory system from the loop: rather than listening to a rhythm, participants generate one by tapping, and we ask whether motor production alone is sufficient to produce the anti-phasic forward-entrainment effect. The motor theory makes a directional prediction here—if the motor system is the source of the effect, generating the rhythm should produce forward entrainment that is at least as strong as, and plausibly stronger than, that seen in the auditory paradigm, since the movement is now overt rather than covertly simulated (Hickok's prediction). The experiment is thus set up to test the motor account rather than to presuppose it: an effect that matches or exceeds the auditory benchmark would support it, whereas an absent, weaker, or inconsistent effect would fail to support it and would motivate looking elsewhere for the source of the effect.

Accordingly, we test whether forward entrainment can arise from motor production. We proceed in two steps, both adapted from Hickok et al. (2015). In Step One, participants perform the original detection task while tapping along with the AM noise, to confirm that the anti-phasic effect is reproducible under our procedure. In Step Two, we remove the entraining sound: participants generate the rhythm by finger tapping alone, and we test whether tones are still detected more accurately in the expected troughs of the (now self-

generated) rhythm. A central feature of this design is that removing the auditory rhythm must be done cleanly; the steps taken to prevent the tapping itself from reintroducing an auditory cue are described in the Methods.

4.2 Methods

4.2.1 Participants

Eleven right-handed adults with normal hearing took part in the study. All participants provided written informed consent, and the study was approved by the Institutional Review Board of the University of California, Irvine. Two participants did not complete the full experiment and two performed at or below chance overall; these four were excluded, leaving seven participants (age mean 21.29 ± SD 4.31; all right-handed) in the final analysis. Participants identified their gender as female (n = 6) or male (n = 1). Each of the remaining participants completed a six-trial practice block, which allowed them to become familiar with the task, followed by ten sessions of 100 trials.

4.2.2 Stimuli and apparatus

Stimuli were generated in MATLAB with Psychtoolbox (PsychPortAudio) at 44.1 kHz and presented over a Focusrite USB audio interface (PsychPortAudio latency class 3, master–slave configuration with separate slave devices for masker and target). The masker was Gaussian white noise, unfiltered and therefore spanning 0–22.05 kHz, 12 s in duration, with 5 ms raised-cosine onset and offset ramps. Level was set by root-mean-square normalization rather than peak normalization, so that -65 dB FS denotes an RMS of 10ˆ(-65/20) relative to digital full scale; signal-to-noise ratios are correspondingly RMS-referenced. For the first

1.75 s of each trial the noise was sinusoidally amplitude modulated at f = 2 Hz with 80% modulation depth,

$$m(t) := \frac{1 + 0.8 \sin\left(2\pi f t + \frac{3\pi}{2}\right)}{1.8} = \frac{1 - 0.8 \cos(2\pi f t)}{1.8}, \tag{4.1}$$

so that the envelope began at a trough and reached maxima 250, 750, 1250, and 1750 ms after noise onset (3.5 cycles). After normalization the modulator spanned 0.111 to 1.00, a peak-to-trough range of 19.1 dB. Thereafter the noise continued unmodulated for the remainder of the trial, so that the target was always masked by unmodulated noise. Targets were 50 ms tones of 1000 or 2000 Hz with 5 ms raised-cosine ramps, presented at an RMS of -65 dB FS plus the trial's SNR.

In-ear headphone Etymotic ER4SR and over-the-head 3M™ PELTOR™ X5 Earmuffs X5A/37274 (AAD) were used together to provide good quality of sound as well as prevent subjects hearing the tapping sound. Sound interface Focusrite Scarlett 2i2 was also used. Apple Empirisoft DirectIN keyboard was used to provide relatively precise timing recording of tapping.

One implementation detail is recorded here for accuracy. The trial routine drew a random starting phase and passed it to the tone generator, but the generator's fourth positional argument is amplitude rather than phase, so the value was bound to amplitude and subsequently removed by the generator's internal peak normalization; every target therefore began at phase zero. Because the masker was freshly randomized on every trial and the analysis concerns target timing rather than target fine structure, this introduces no systematic bias into any result reported below, and is noted only so that the description matches what the code executed.

4.2.3 Thresholds

Reference levels were obtained from a two-down one-up adaptive staircase measuring tone-in-noise detection. On each trial a 1 s sample of steady noise at -65 dB FS was presented, containing a centered 50 ms tone on a random half of trials, and the listener reported whether a signal had been present. The track began at 0 dB SNR, stepped 4 dB until two reversals had occurred and 2 dB thereafter, was bounded to [-40, 0] dB SNR, and terminated after twelve reversals or eighty trials. The threshold estimate for a run was the mean of its final six reversal levels. Eight runs were completed in strict alternation between 1 kHz and 2 kHz, four per frequency; alternating in this way distributes any order or fatigue effect evenly across the two carriers. The mean of the final three runs at each frequency defined that participant's reference level, the first run being treated as practice on the staircase task itself. Trial-by-trial feedback was provided during threshold measurement but not during the main experiment.

In the main experiment the target was presented at one of five signal-to-noise ratios spaced $20 \cdot \log_{10}(w/w_1)$ dB above that reference, with w = [0.025, 0.0312, 0.0375, 0.0437, 0.05], i.e. [0, 1.93, 3.52, 4.85, 6.02] dB. Level was randomized trial by trial. Anchoring the five levels to each frequency's own threshold allows accuracy at the two carrier frequencies to be pooled onto a common relative scale.

Three features of this arrangement are noted. First, the reference is a detection threshold whereas the experimental task is discrimination, so the mapping between the two is an assumption rather than a measurement; the present data bear on it incidentally, since discrimination accuracy ran from .54–.75 at the detection threshold to .93–.99 at +6 dB, placing the discrimination 75% point roughly 0–3.5 dB above the detection threshold. Second, the staircase ceiling coincides with its starting level, so an observer whose threshold lies above 0 dB SNR cannot be distinguished from one performing normally; in the present data all

tracks converged normally, so no threshold was invalid. Third, the gating of the two tones differs: the threshold tone receives a 10 ms linear ramp followed by a 5 ms raised-cosine ramp with level set after both, whereas the experimental tone receives only the 5 ms raised-cosine ramp with level set before it, so experimental tones are approximately 0.6 dB quieter than nominal relative to the measured threshold. This offset is common to every condition and participant and translates the psychometric function without affecting any phase measure.

4.2.4 Experimental task

The hypothesis was tested in two steps, both adapted from the procedure of Hickok et al. (2015) and modified to incorporate finger tapping.

Step one:

Listeners were presented with a wideband Gaussian-noise stimulus lasting 4 s per trial. The noise was amplitude modulated at 3 Hz (80% modulation depth) for the first 3 s (the entrainment phase); the modulation then terminated on the cosine phase of the following cycle, leaving the final 1 s of the noise unmodulated (Figure 4.2). Participants tapped to the beat of the 3 Hz entraining stimulus and continued tapping through the steady-state noise after the modulation had stopped. In the original study, a tone was present or absent at one of nine temporal positions during the unmodulated portion. Here, a tone of one of two frequencies was presented on every trial, and tones were restricted to the final five temporal positions rather than nine, for two reasons: the original study showed that the effect is present in both the first and second post-offset cycles, and restricting the positions discouraged participants from counting beats during the entrainment phase. Because participants might continue tapping after the modulated noise ended, the design sought to remove any contribution of tapping and to ensure that the judgement was made

on the basis of listening alone.

On half of the trials a low (1 kHz) tone was presented and on the other half a high (2 kHz) tone; both were 50 ms in duration with 5 ms rise and decay times. The five temporal positions began one cycle (333 ms) after the offset of the modulation and were spaced a quarter-cycle (83.3 ms) apart, so that they covered one full cycle of the modulation waveform had it continued. Because one modulation cycle at 3 Hz equals 333 ms, the tone was presented between 3.333 and 3.666 s after noise onset (333 to 666 ms after termination of the modulation).

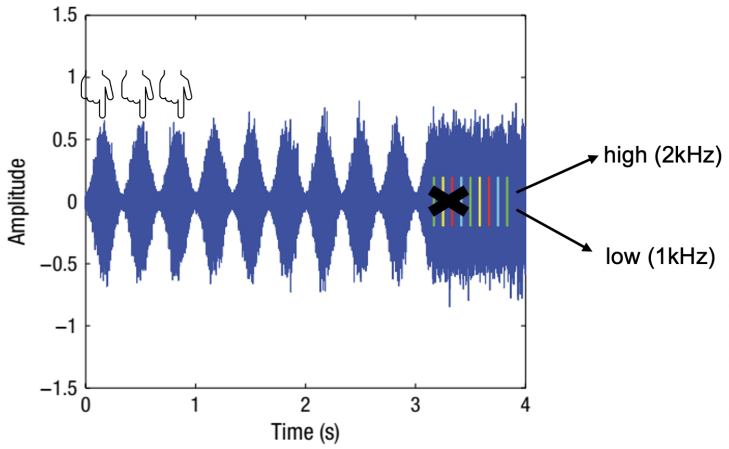


Figure 4.2: Stimulus of the current experiment, adapted from Hickok et al. (2015). The principal differences from the original study are that (1) participants tapped along with the modulated noise, and (2) the temporal positions of the tone were reduced from nine to the final five.

Before the main experiment, a single-interval tone-in-noise detection task estimated intensity thresholds for the 1 kHz and 2 kHz tones separately. On each trial the participant heard either a signal (a tone embedded in unmodulated Gaussian noise) or no signal (Gaussian noise alone) and reported whether a tone was present. Tone intensity was adjusted with a two-down one-up adaptive procedure, and the 71%-correct detection threshold (Levitt, 1971) was recorded for each frequency. The difference in detection thresholds between the two frequencies was preserved in the main experiment to account for the asymmetry in sensitivity across tones.

In the main experiment, each trial was a single-interval two-alternative forced-choice task in which the participant pressed one of two keys ("1" or "2") to indicate whether the tone presented during the unmodulated segment was high or low. The prior probability of a signal was 50%. When a tone was presented, its temporal position was drawn randomly from the five values and its level from five values spanning 0 to 6 dB, evenly spaced on a log scale, to allow estimation of psychometric functions (Figure 4.3). Each run comprised 100 trials, and each participant completed a minimum of 20 runs.

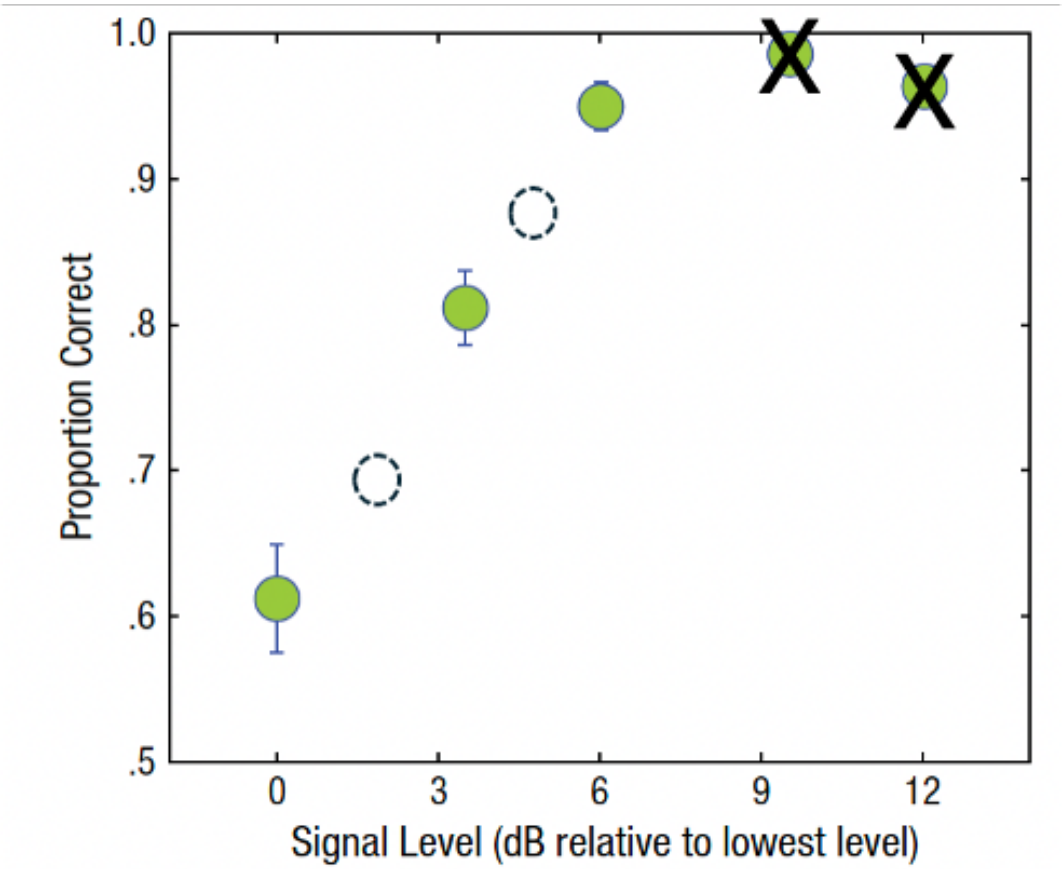


Figure 4.3: Mean proportion of correctly detected tones as a function of signal level in the main experiment. Adapted from Hickok et al. (2015).

Stimuli were generated in MATLAB (The MathWorks, Natick, MA) on a Dell computer [model — to be specified] and presented at 44.1 kHz through 16-bit digital-to-analog converters (Focusrite Scarlett 2i2) and Sennheiser HD 280 Pro headphones in a steel-walled,

acoustically isolated chamber (IAC Acoustics, Stafford, TX).

Step Two:

Before the first experimental session each participant completed a six-trial practice block using the identical trial routine. Step Two built on Step One with a few modifications. The amplitude-modulated stimulus was set to 2 Hz rather than 3 Hz: as noted above in the discussion of preferred rates for different motor effectors, approximately 2 Hz is the average rate at which participants freely tap rhythms with the finger (MacDougall & Moore, 2005), which makes 2 Hz a more suitable choice here than 3 Hz.

The design comprised three stages. In the synchronization stage, participants synchronized their finger tapping to 2 Hz for the first three taps, ensuring that all participants tapped at approximately the same rate. In the continuation stage, participants produced seven further taps at the learned 2 Hz rate, thereby generating a motor rhythm in place of the auditory rhythm used in Step One. In the detection stage, following the first ten taps, a tone was triggered on one of the next three taps if tapping continued, at one of four positions (quarter-cycle intervals) within a tapping cycle, as in Step One.

Participants tapped the space bar in time with the rhythm and continued until they heard a tone. Tap times were recorded with a keyboard queue that logs press events only, returning a two-row array of key codes and timestamps on the same clock as the noise-onset time t0. The trial comprised three taps during the modulated portion (synchronization), seven taps with the noise now unmodulated (continuation), and one, two, or three additional taps chosen at random. The last of these, timestamped tCue, triggered the target: the tone was scheduled at tCue + Δ, with $\Delta \in \{125, 250, 375, 500\}$ ms drawn at random. Participants then reported whether the tone was low (1) or high (2). No feedback was given. Randomizing which tap served as the trigger rendered the absolute time of the target unpredictable while

leaving its phase within the tapping cycle unchanged.

Three concerns were addressed by the Step Two design. The first is that the sound of tapping could itself reintroduce an auditory cue, which would undermine the aim of removing auditory input; participants therefore wore insert earphones together with noise-reducing earmuffs, allowing them to tap without hearing the tap. The second is the quality of the self-generated rhythm—that is, how accurately participants tapped—which was assessed from the recorded tap times after the experiment. The third is the timing delay associated with tapping. The second and third concerns were both addressed by using a MIDI keyboard to obtain accurate timing for each tap.

The task and remaining procedures, including the threshold measurement, were the same as in Step One. (Figure 4.4)

4.2.5 Data Analysis

Analyses were first conducted separately for each participant, in two steps. First, a psychometric function was constructed from the proportion correct at the five SNRs, and the SNR closest to 69% correct—corresponding approximately to the discrimination threshold ($d' = 1$)—was aimed to be identified for use in the subsequent analysis. However, to be practical, in the following analysis for Stage Two, either the lowest or optimal signal level was chosen for exploratory reason. Second, at that SNR, the proportion correct was computed at each temporal position and inspected for any modulation as a function of position within the rhythm cycle; d' could be used in place of proportion correct. This procedure was identical for Step One and Step Two. For Step Two, a simple average was additionally computed across participants and plotted as proportion correct across temporal positions, in the same form as the single-subject plots.

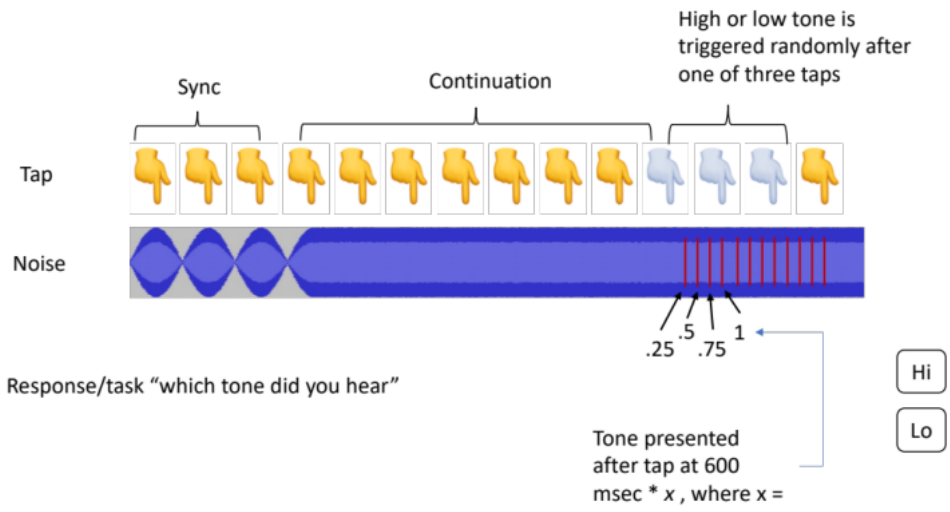


Figure 4.4: Schematic of the experimental design in Step Two.

4.3 Results

4.3.1 Stage One

Step One served to confirm that the anti-phasic forward-entrainment effect of Hickok et al. (2015) was reproducible under the present procedure, in which participants tapped along with the entraining stimulus. Data were obtained from two participants. For subject "a", the psychometric function (Figure 4.5, right) indicated that proportion correct was closest to 69% at the lowest SNR tested (Δ 0 dB, signal level 1); at that level, proportion correct as a function of temporal position showed a relatively anti-phasic pattern (Figure 4.5, left). For subject "x", proportion correct was closest to 71% at the second-lowest SNR (≈ Δ 2 dB, signal level 2), and the corresponding temporal-position curve likewise showed a relatively anti-phasic pattern (Figure 4.6).

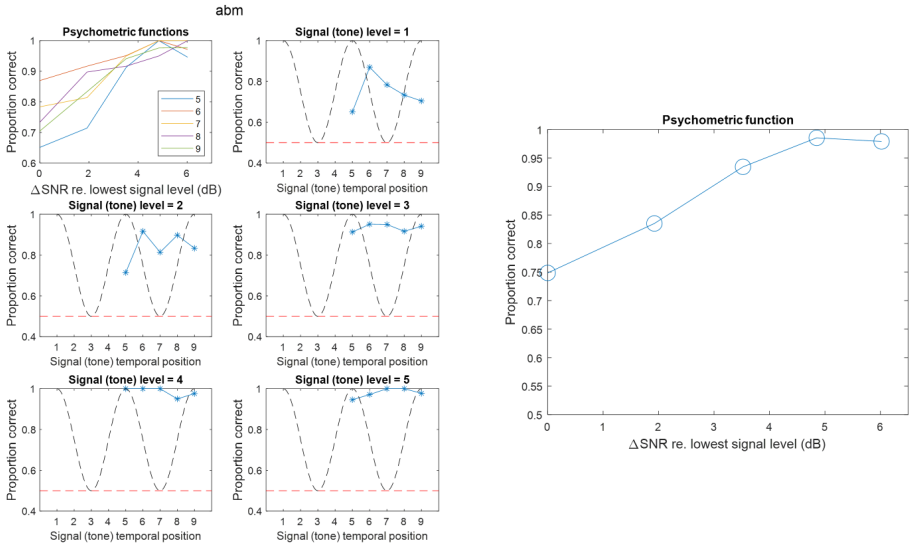


Figure 4.5: Individual data from "a". The psychometric function (right) shows proportion correct closest to 69% at SNR = Δ 0 dB (signal level 1). Proportion correct as a function of signal temporal position (left), at signal level 1, shows a relatively anti-phasic pattern. Chance level is indicated by the red dashed line. The five SNRs are evenly spaced on a log scale over a 0–6 dB range.

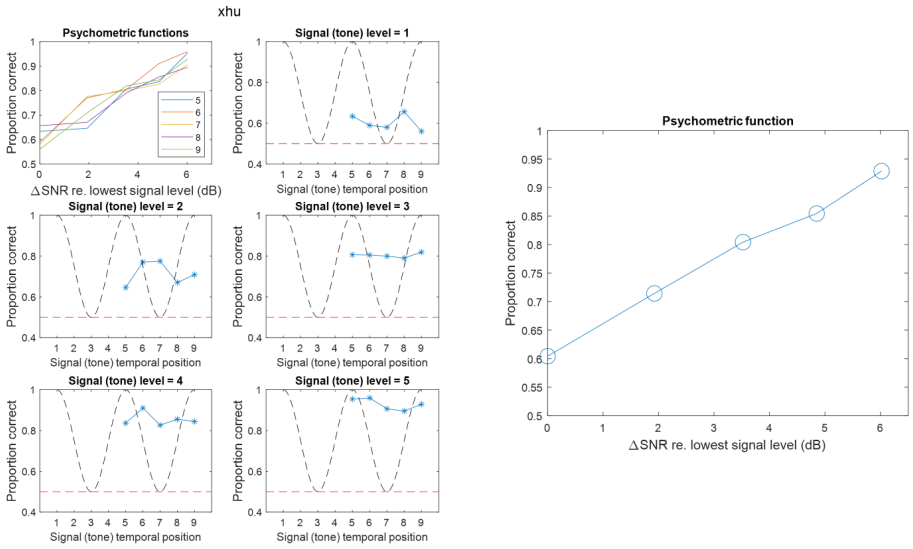


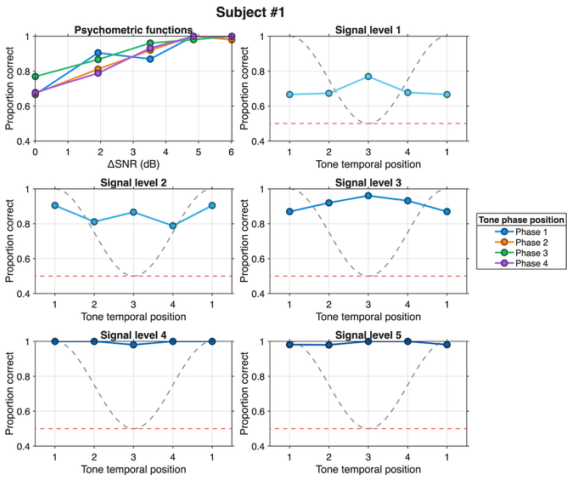
Figure 4.6: Individual data from "x". The psychometric function (right) shows proportion correct closest to 69% at SNR $\approx \Delta$ 2 dB (signal level 2). Proportion correct as a function of signal temporal position (left), at signal level 2, shows a relatively anti-phasic pattern. Chance level is indicated by the red dashed line. The five SNRs are evenly spaced on a log scale over a 0–6 dB range.

Both participants reproduced the anti-phasic pattern reported by Hickok et al. (2015) under the modified procedure, confirming that the paradigm behaves as expected in our hands. Having established this, we proceeded to Step Two.

4.3.2 Stage Two

Single-subject results are shown for the seven participants (Figures 4.7–4.13), comprising the psychometric function together with proportion correct as a function of tone temporal position, plotted separately for the five signal levels. A subset of participants showed an anti-phasic pattern at a given SNR—typically the SNR at which proportion correct was near 69% or somewhat lower—including subjects 1, 2, 4, and 6; the remaining participants did not show a clear pattern.

We then pooled the participants by selecting, for each, the SNR closest to 69% correct or



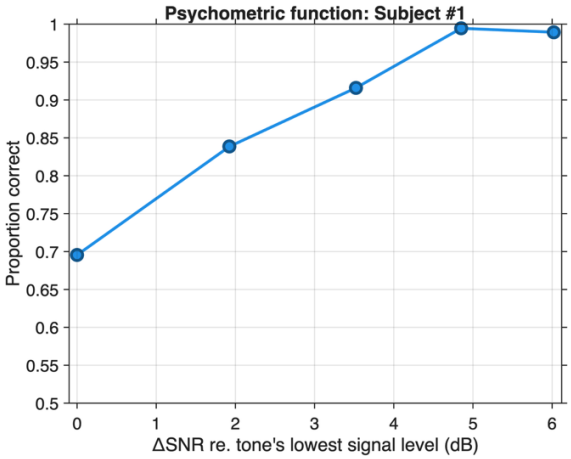
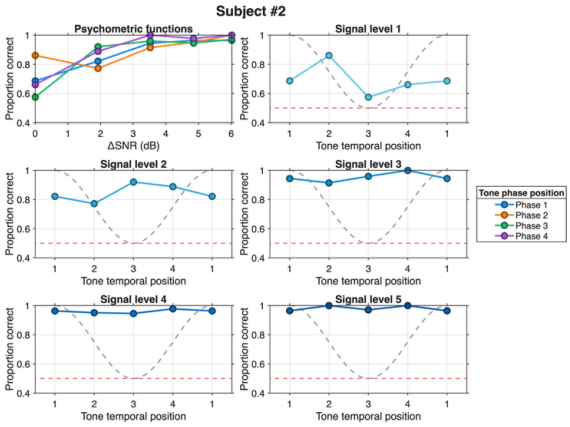


Figure 4.7: Individual data from Subject 1. The psychometric function (right) shows proportion correct closest to 69% at SNR = Δ 0 dB (signal level 1). Proportion correct as a function of signal temporal position (left), at signal level 1, shows a relatively anti-phasic pattern. Chance level is indicated by the red dashed line.



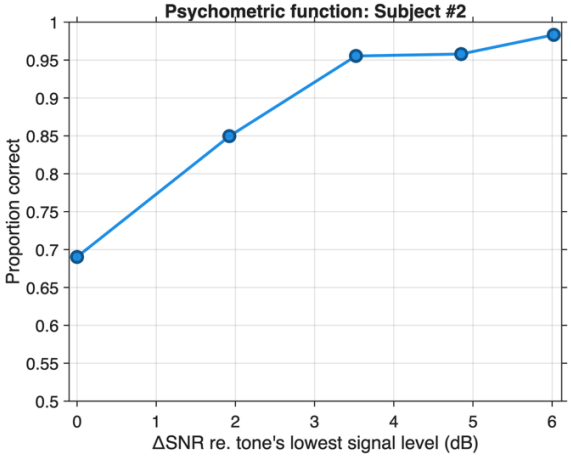


Figure 4.8: Individual data from Subject 2. The psychometric function (right) shows proportion correct closest to 69% at SNR = Δ 0 dB (signal level 1). Proportion correct as a function of signal temporal position (left), at signal level 1, shows a relatively anti-phasic pattern. Chance level is indicated by the red dashed line.

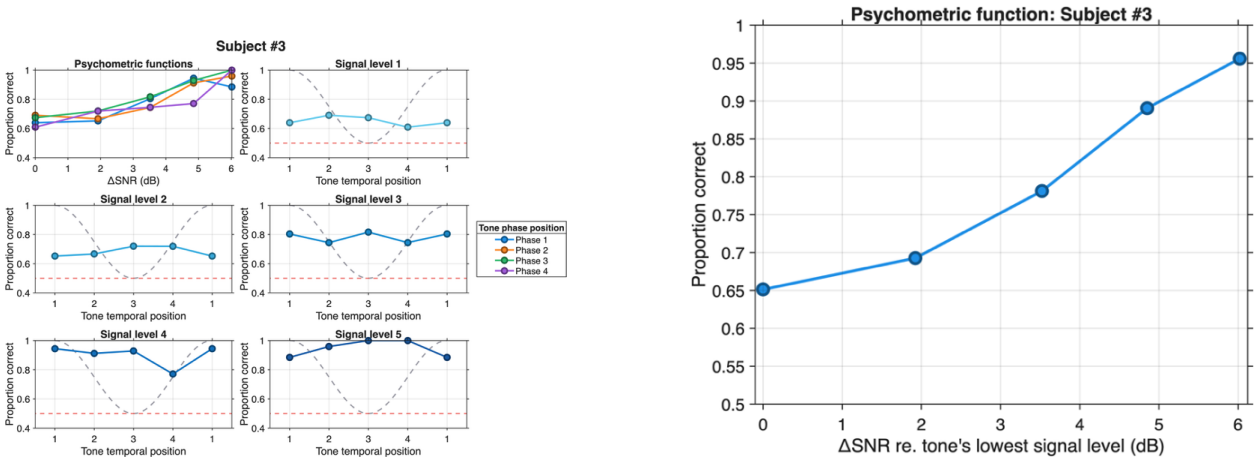


Figure 4.9: Individual data from Subject 3. The psychometric function (right) shows proportion correct around 65% at SNR = Δ 0 dB (signal level 1). Proportion correct as a function of signal temporal position (left), at signal level 1, shows a relatively anti-phasic pattern. Chance level is indicated by the red dashed line.

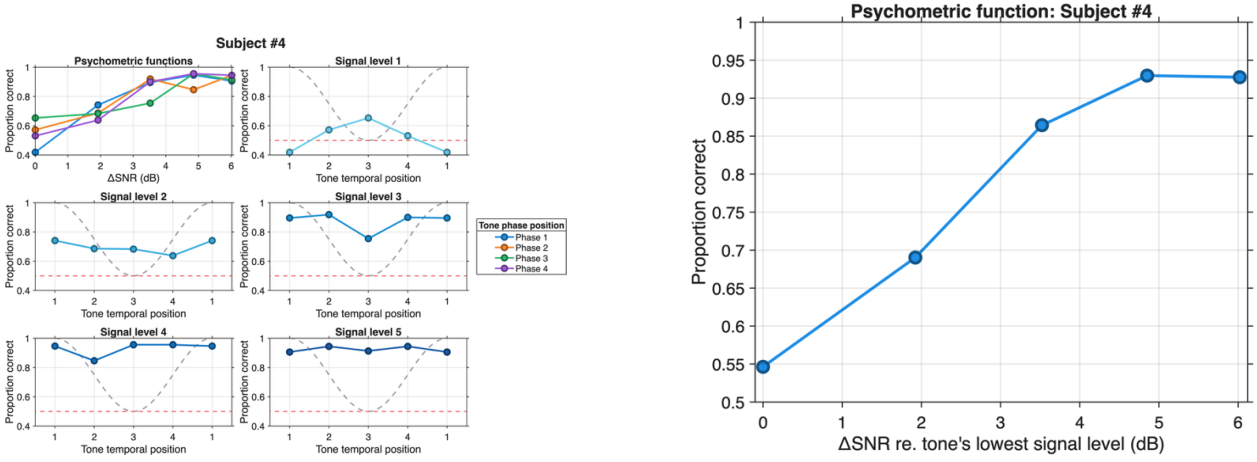


Figure 4.10: Individual data from Subject 4. The psychometric function (right) shows proportion correct around 55% at SNR = Δ 0 dB (signal level 1). Proportion correct as a function of signal temporal position (left), at signal level 1, shows a relatively anti-phasic pattern. Chance level is indicated by the red dashed line.

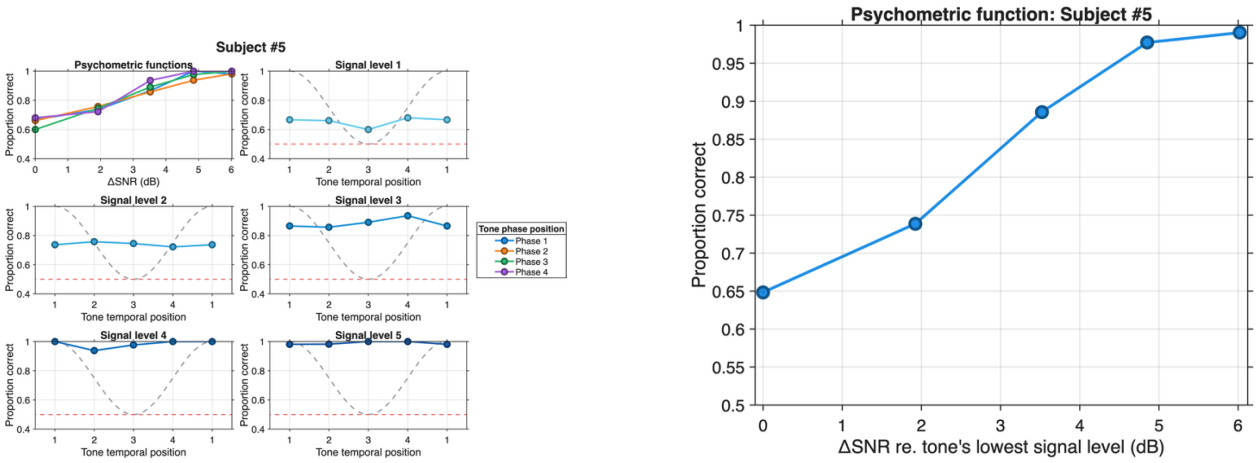


Figure 4.11: Individual data from Subject 5. The psychometric function (right) shows proportion correct around 74% at SNR = Δ 0 dB (signal level 2). Proportion correct as a function of signal temporal position (left), at signal level 2, shows a relatively anti-phasic pattern. Chance level is indicated by the red dashed line.

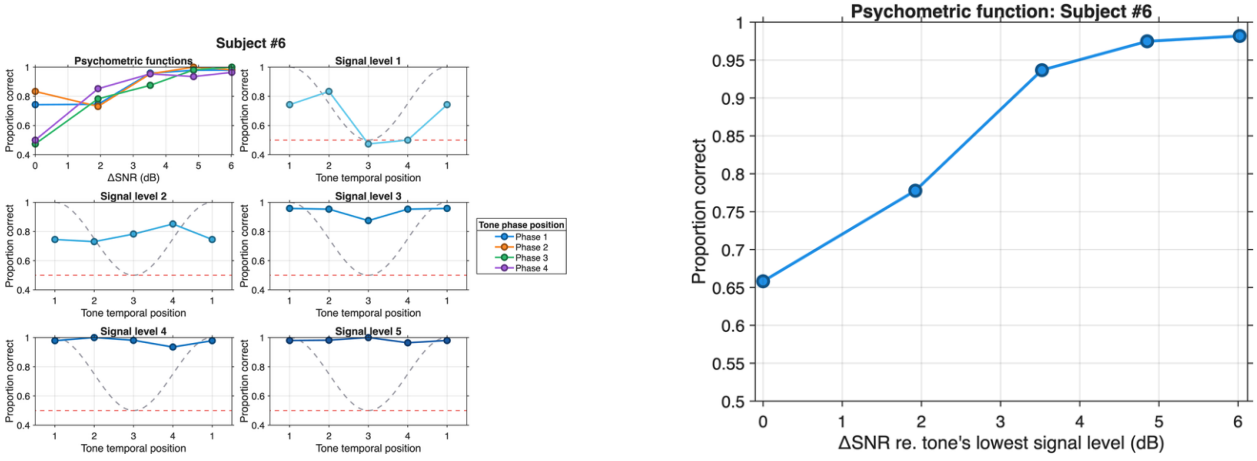


Figure 4.12: Individual data from Subject 6. The psychometric function (right) shows proportion correct closest to 69% at SNR = Δ 0 dB (signal level 1). Proportion correct as a function of signal temporal position (left), at signal level 1, shows a relatively anti-phasic pattern. Chance level is indicated by the red dashed line.

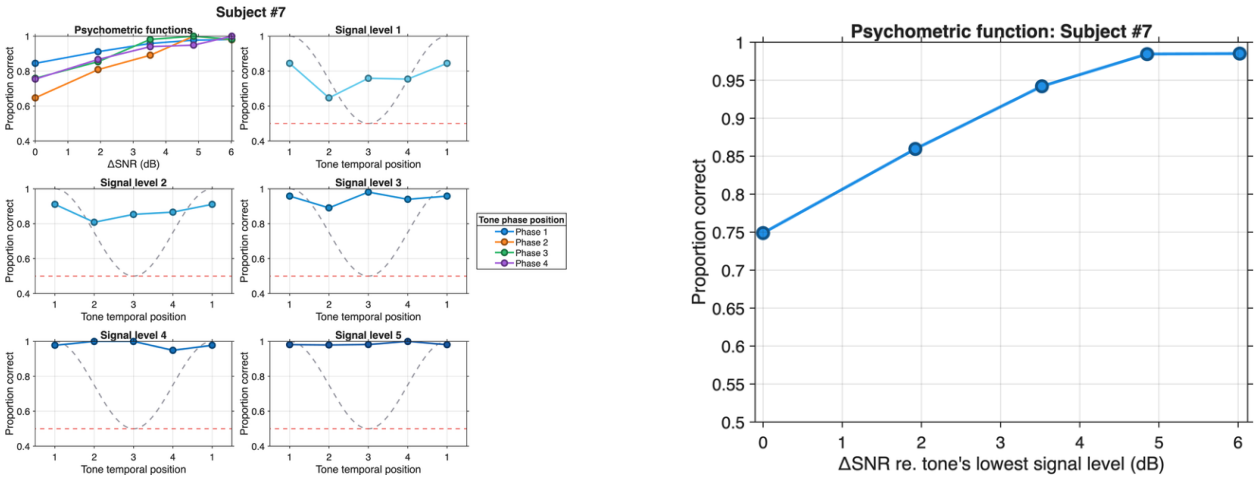


Figure 4.13: Individual data from Subject 7. The psychometric function (right) shows proportion correct closest to 69% at SNR = Δ 0 dB (signal level 1). Proportion correct as a function of signal temporal position (left), at signal level 1, shows a relatively anti-phasic pattern. Chance level is indicated by the red dashed line.

the optimal signal level (for exploratory purpose), and computed the group average across temporal positions; an analogous averaging approach has been used previously (Saberi & Hickok, 2022b; Figures 4.14–4.15).

Because prior work has documented that a proportion of participants may not show the anti-phasic effect under identical conditions, and on the basis of the pooled curves, we grouped the participants into two sets: Group A (subjects 2, 4, and 6), with a relatively clear peak-and-trough pattern, and Group B (subjects 1, 3, 5, and 7), with a less distinct pattern (Figures 4.16–4.18). These groupings are examined further in the Discussion.

4.4 Discussion

4.4.1 Overall pattern and magnitude

Across the seven Step Two participants, the picture was mixed rather than uniform. A subset of participants (Group A: subjects 2, 4, and 6) showed a relatively clear peak-and-trough

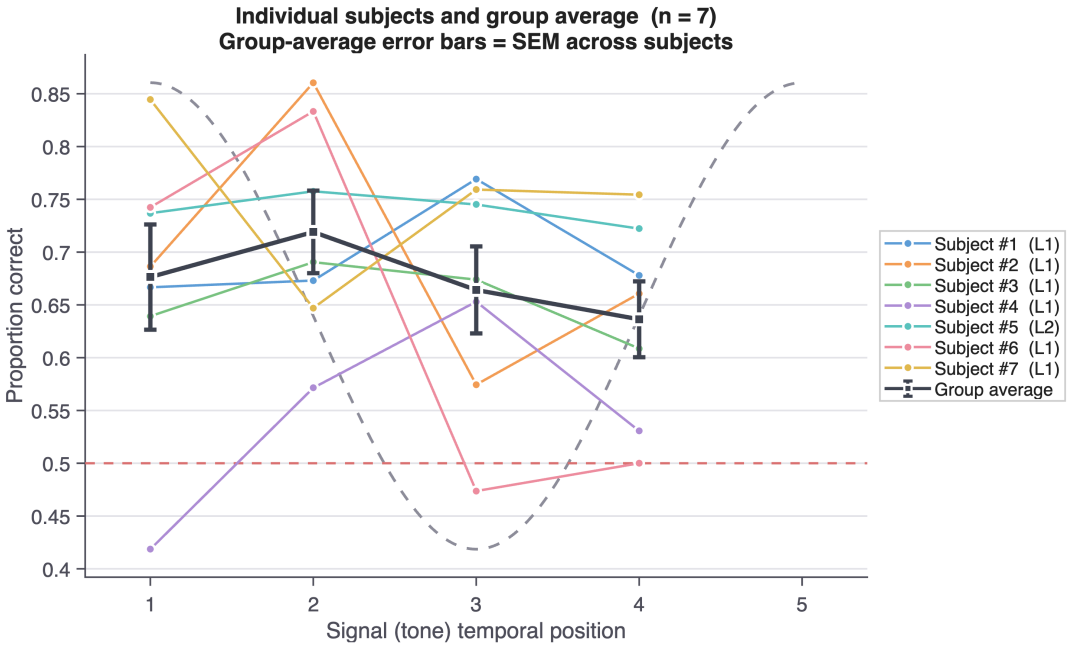


Figure 4.14: Group average across seven subjects as well as single subject data for proportion correct as a function of signal temporal position.

pattern in detection performance across the tapping cycle, while the remainder (Group B) did not show a clear pattern. For the Group A participants, the difference between the peak and trough of proportion correct reached approximately 15% or more, and the Group A average was of comparable magnitude. This is on the order of the modulation reported for the original auditory paradigm (Hickok et al., 2015), indicating that a sinusoidal modulation of performance can be present when the rhythm is self-generated rather than heard.

Two qualifications temper this observation. First, the effect was confined to a subset of participants; such intersubject variability is itself characteristic of forward-entrainment paradigms, where a proportion of subjects show the effect and a proportion do not under identical conditions (Henry et al., 2025; Saberi & Hickok, 2023). Second, and importantly, the peaks in the motor condition were not well aligned to the troughs of the tapping cycle in the way that the auditory effect is aligned to the modulator. This misalignment suggests that a phase correction may be required before the group pattern can be fairly evaluated (see below).

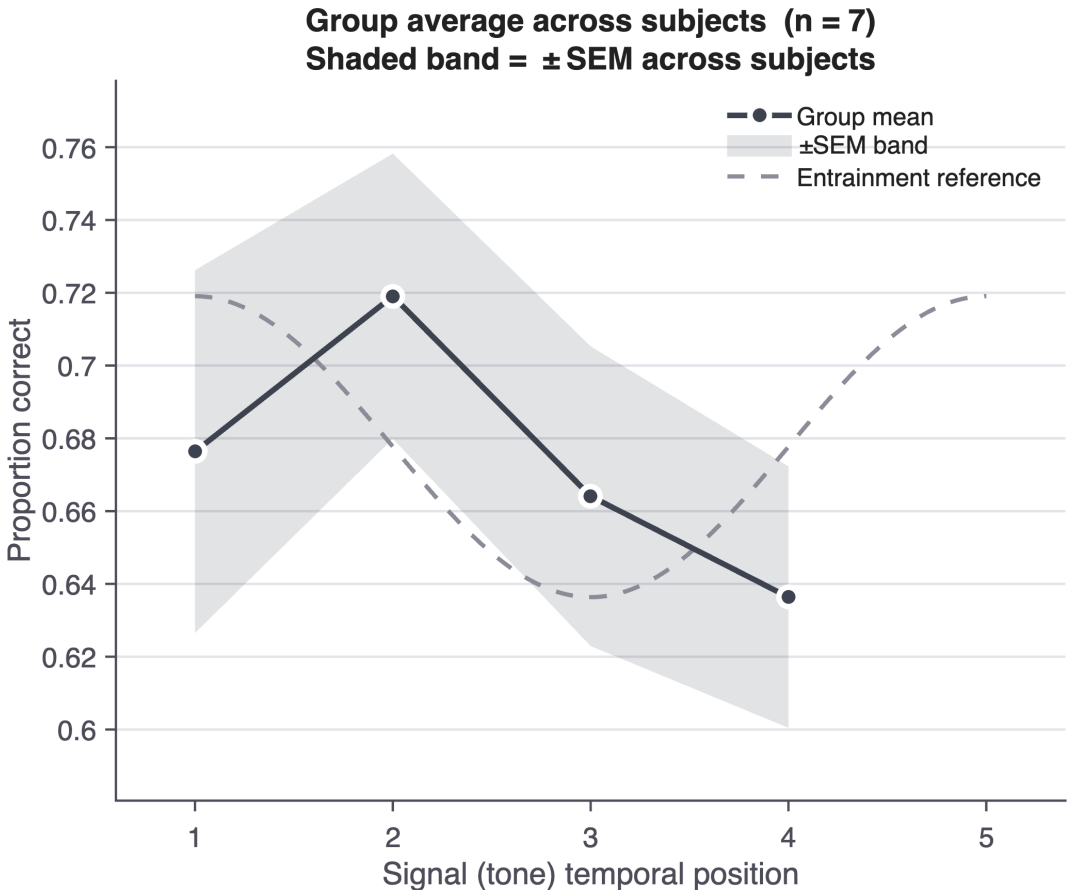


Figure 4.15: Group average across seven subjects for proportion correct as a function of signal temporal position.

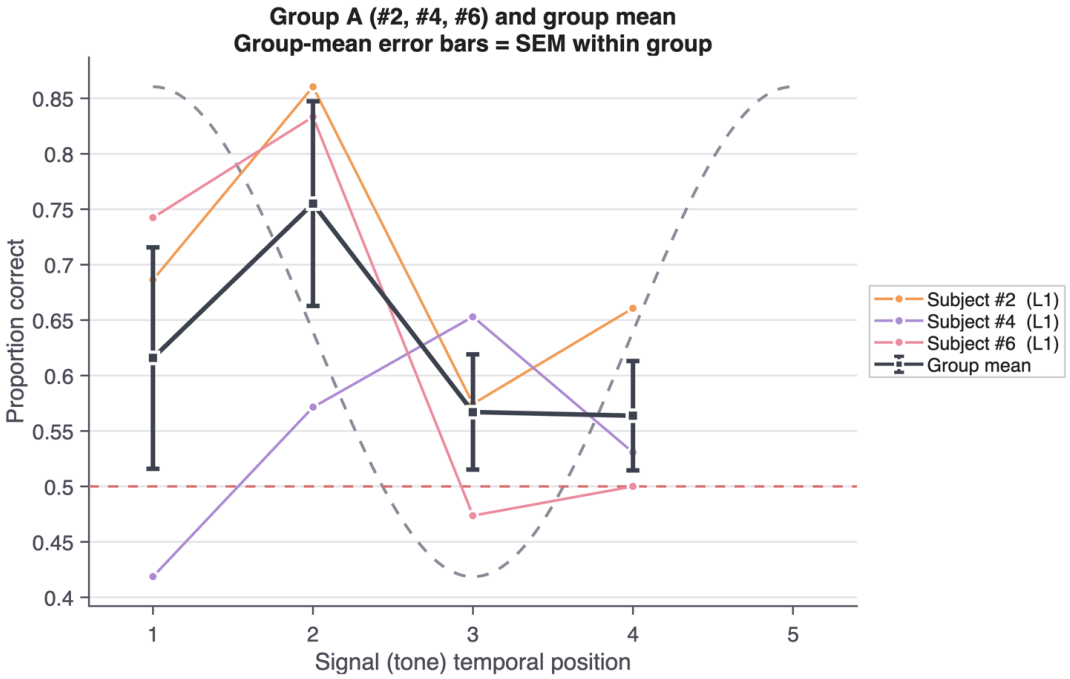


Figure 4.16: Group average across subjects 2, 4 & 6 as well as single subject data for proportion correct as a function of signal temporal position.

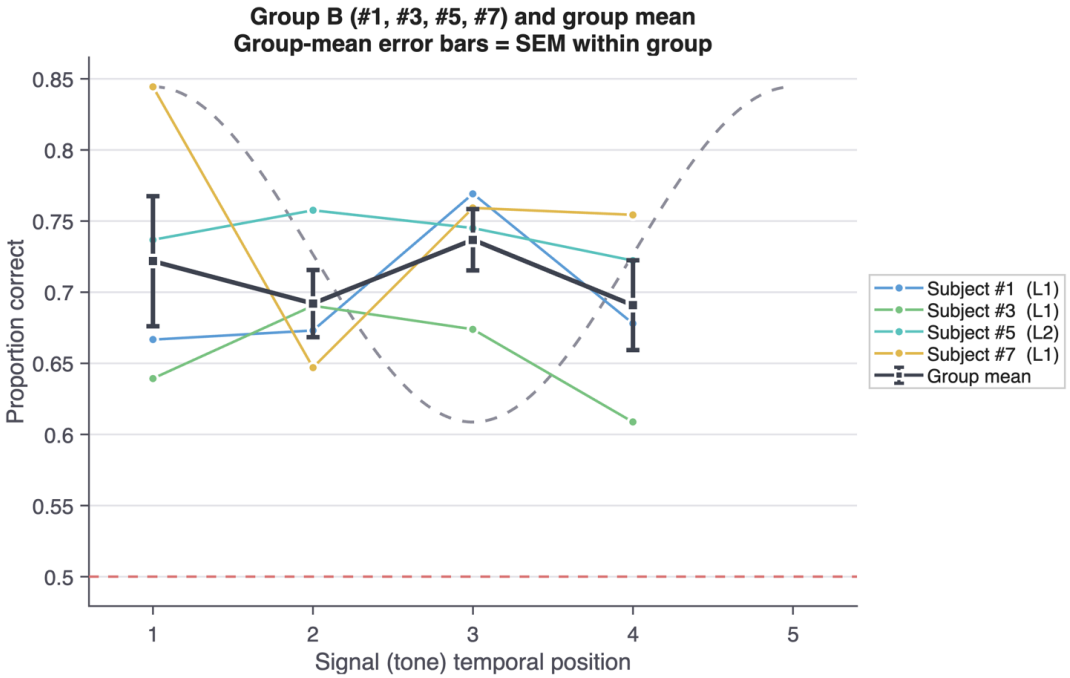


Figure 4.17: Group average across subjects 1, 3, 5 & 7 as well as single subject data for proportion correct as a function of signal temporal position.

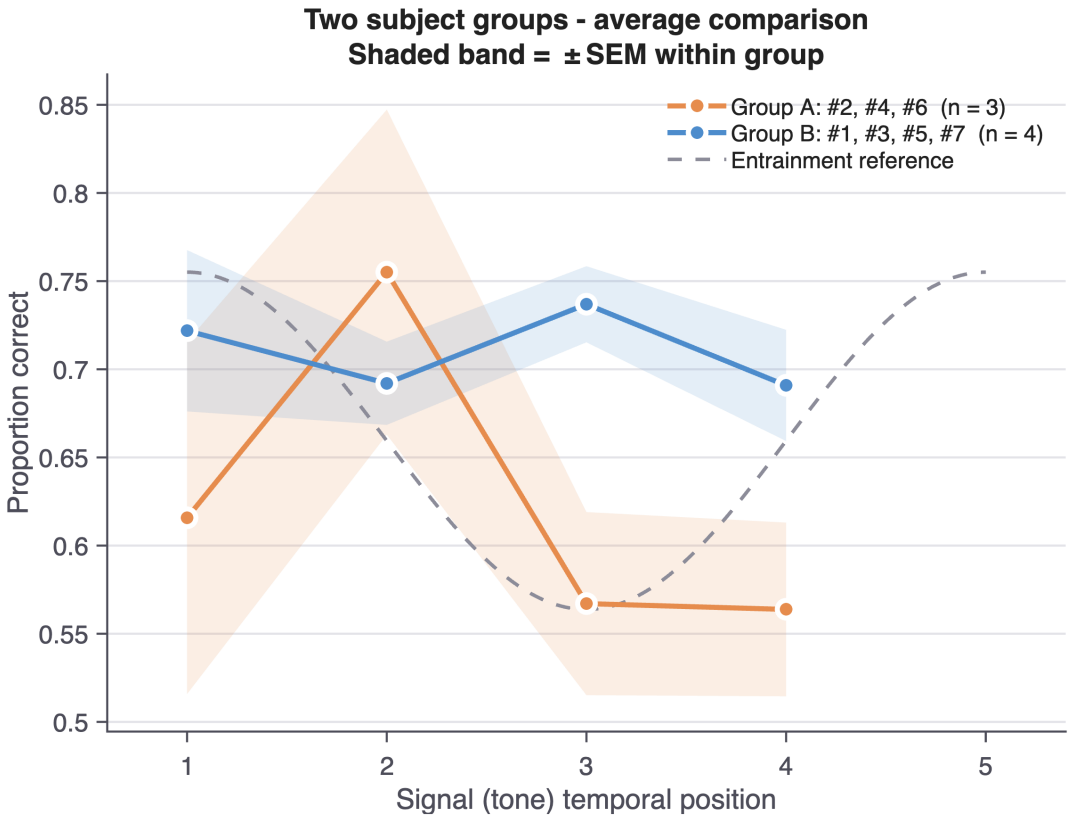


Figure 4.18: Group average across subjects for proportion correct as a function of signal temporal position. Group A includes subjects 2, 4 & 6; Group B includes subjects 1, 3, 5 & 7.

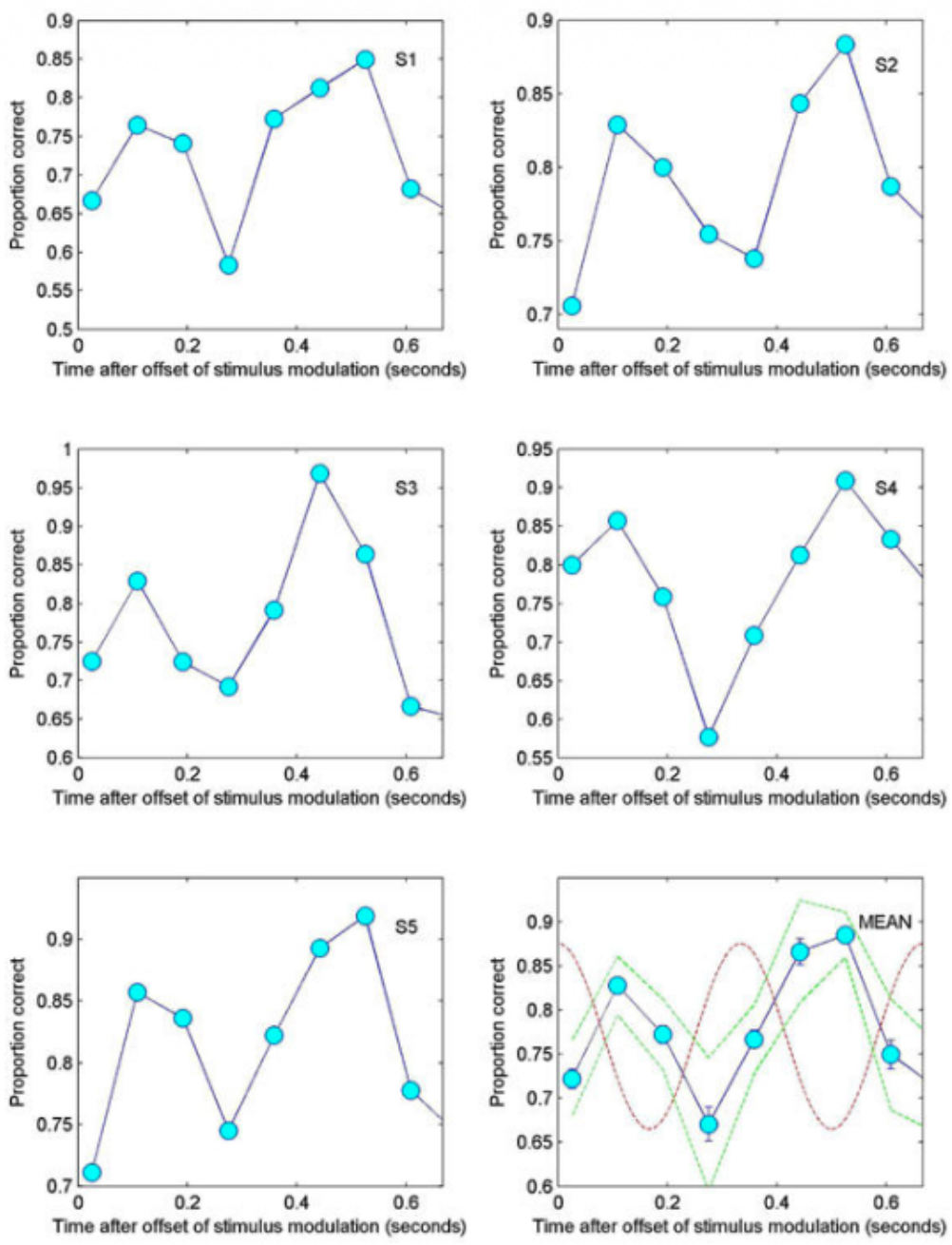


Figure 4.19: Tone detectability as a function of time after offset of noise modulation. The abscissa represents time after offset of masker modulation (at its cosine phase). Each panel displays data from one subject, with the bottom-right panel showing mean performance. The red dashed curve in the bottom-right panel shows expected masker modulation, had it continued. Green lines represent 95% confidence limits. Error bars represent 1 standard error. Reprinted from Hickok, Farahbod & Saberi, 2015.

4.4.2 Relation to the hypothesis

The motivating prediction was specific: if the motor system is the source of the anti-phasic forward-entrainment effect, then a purely motor version of the task should produce an effect *at least as strong* as the auditory version, and plausibly stronger (Hickok's prediction). Measured against this benchmark, the present results do not provide clear support for the motor account. In the motor condition the effect was, at most, comparable in magnitude to the auditory effect rather than stronger, was restricted to a subset of participants, and was not phase-aligned.

At the same time, the results do not license the opposite conclusion either. Some participants did show a modulation of the expected magnitude, which is not what one would expect if motor production had no bearing on the effect at all. The data therefore neither confirm that motor production is the source of forward entrainment nor establish that it is irrelevant; they are, at present, genuinely inconclusive. Two features of the design—phase alignment across subjects, and whether the most sensitive signal level was sampled—could each be masking or inflating the pattern, and both must be addressed before the motor account can be evaluated. We take these up in turn, and treat the present result as a step that sets up the necessary analyses rather than one that settles the question.

4.4.3 Phase alignment across subjects

The most likely confounding factor is phase misalignment. Even among the Group A participants, whose individual curves show clear peaks and troughs, the peaks do not occur at the same phase across subjects; for example, subject 4 peaks at a different position than subjects 2 and 6, which flattens the group-average curve relative to the single-subject curves. This is a recognized issue in this literature: minor phase misalignments across subjects can diminish or flatten modulation patterns when performance is averaged, so that a real effect is washed

out in the mean (Henry et al., 2025; Saberi & Hickok, 2023). Two sources of misalignment are worth separating. First, although participants were asked to hold a common rate, they may gradually drift from it; because tap times were recorded, this can be checked directly and, if present, corrected on the basis of each participant's own tapping cycle. Second, even without drift, the preferred phase at which an individual shows the effect may differ across participants, as has been reported both by us and by others (Henry et al., 2025; Saberi & Hickok, 2023). Aligning peaks on a subject-by-subject basis before averaging, or at minimum reporting the cross-subject spread of preferred phase, would yield a more informative group estimate. A complementary strategy would be to average across subjects separately at each SNR rather than pooling.

4.4.4 Locating the sweet spot

A second factor concerns whether the most sensitive signal level was even sampled. In the original auditory paradigm, the anti-phasic modulation is clearest at a particular near-threshold SNR and is small or absent at higher, near-ceiling SNRs (Hickok et al., 2015; Saberi & Hickok, 2023). It is therefore possible that, for the Group B participants, we did not sample the SNR at which the effect would be clearest—the relevant level may fall below the range spanned by our five SNRs, or between the values we tested. Collecting more trials per participant, or widening and refining the range of SNRs, could reveal an effect that the present sampling missed. This is not a limitation in the sense of a design flaw—small-sample psychophysical designs of this kind are standard—but an indication of where additional data could be most informative.

4.4.5 Taps following the target

One further methodological point deserves checking. On some trials a participant may have produced an additional tap *after* the target tone was presented, despite instructions to stop; because the target was triggered after a randomly selected tap, the inertia of ongoing tapping could carry a participant one tap past the cue. Whether and how this affects performance is not known a priori, but because tapping was recorded throughout, trials with and without a post-target tap can be separated and compared directly.

4.4.6 Summary

Returning to the motivating question: if the motor system were the source of the anti-phasic forward-entrainment effect, a purely motor version of the task should have produced an effect at least as strong as the auditory version. We did not observe this—the effect, where present, was at most comparable in magnitude, was limited to a subset of participants, and was not phase-aligned. On the current evidence we therefore do not have strong support for the motor theory of rhythm. We are equally clear, however, that the present data do not settle the matter in either direction: a modulation of the expected magnitude was present in some participants, and the phase-alignment and sweet-spot issues identified above could each be obscuring the underlying pattern. A firm conclusion awaits the further analyses these issues motivate—subject-wise phase alignment, per-SNR averaging, separation of post-target-tap trials, and additional data at a refined range of SNRs. This experiment is best understood, then, as establishing a clean paradigm for isolating the motor contribution to forward entrainment and identifying the analyses required to adjudicate it, with the question of whether rhythm perception depends on motor simulation left open for that work to resolve.

Chapter 5

General Discussion

This dissertation examined auditory and motor rhythm entrainment across three experiments, asking both where the capacity for beat-based synchronization comes from and whether perceiving a rhythm itself engages the motor system. The three studies approached the problem from complementary angles—behavioral, neuroimaging, and psychophysical—and each has been discussed in detail within its own chapter. Here I draw the threads together, consider what the experiments establish when taken as a whole, and address an apparent tension between them that is, I will argue, better understood as a distinction between two separable questions than as a contradiction.

5.1 Summary of findings

The behavioral experiment compared synchronization to isochronous rhythms with a non-vocal effector (finger tapping) and a supralaryngeal vocal effector (tongue clicking). Tongue clicking was as consistent as finger tapping and, on average, somewhat more closely aligned to the beat. The vocal system is thus at least as competent a synchronizer as a non-vocal

effector, leaving intact the premise that rhythmic synchronization is grounded in the speech system. As emphasized in that chapter, this comparison is a sanity check on a premise the two theoretical accounts share, not an adjudication between them.

The neuroimaging experiment mapped the neural substrate of the same comparison. Beyond the effector-specific regions expected from somatotopy, finger and tongue synchronization converged on common cortex, most notably the dorsal precentral speech area (dPCSA). Because the task involved no pitch or laryngeal control, the engagement of the dPCSA suggests that it is not modality-specific but multifunctional—an auditory-weighted region within the motor system that may serve a more general auditory-motor synchronization function and act as a common node through which auditory information reaches multiple motor effectors.

The forward-entrainment experiment turned from the effectors of synchronization to the mechanism of perception, testing the motor theory of rhythm. With the auditory rhythm removed and the rhythm self-generated by tapping, a subset of participants showed a modulation of the expected magnitude, but the enhancement predicted by a motor account was not observed and the pattern was not consistently phase-aligned across participants. The results therefore do not provide strong support for the motor theory of rhythm; equally, they do not exclude it. The question of whether rhythm perception depends on motor simulation is left open, and the experiment is best understood as establishing a clean paradigm for isolating the motor contribution and identifying the analyses required to resolve it.

5.2 Evolutionary source versus online mechanism

Read side by side, the three experiments might appear to pull in different directions. The behavioral and neuroimaging experiments point toward a motor system that is deeply in-

volved in rhythmic synchronization: a vocal effector synchronizes as well as a non-vocal one, and both engage a shared, auditory-weighted region of the motor system. The forward-entrainment experiment, by contrast, fails to show that motor production alone is sufficient to generate the perceptual signature of entrainment. Taken naively, the first two results seem to elevate the role of the motor system while the third seems to diminish it.

This tension dissolves once the two questions being asked are separated. The first two experiments concern the *evolutionary source* of the capacity for beat perception and synchronization—why the capacity exists at all, and how it comes to be available across the motor system. The third concerns the *online mechanism* of rhythm perception—whether, in the moment of perceiving a rhythm, the motor system is actively carrying out the perception. These are distinct questions, and an answer to one does not settle the other. A capacity can have evolved from, and remain neurally entangled with, the motor system while nevertheless not depending on motor simulation for every one of its online perceptual components. In short, vocal learning may be the evolutionary source of the capacity for beat-based synchronization without motor simulation being the mechanism of rhythm perception. What looks like a contradiction across chapters is, on this reading, a demonstration that the origin of a capacity and its moment-to-moment operation are separable levels of explanation—and the dissertation is stronger for keeping them apart.

5.3 Relation to the two hypotheses

The framing that motivated the first two experiments distinguished two accounts of the link between vocal learning and beat-based synchronization: Patel's fortuitous-enhancement account, on which strengthening of a laryngeal pitch-control pathway incidentally spread auditory-motor connectivity to adjacent non-vocal premotor regions, and Hickok's coordination conjecture, on which beat synchronization is a simple form of what is needed to

coordinate the vocal streams and, once built, spreads to the motor system as a whole.

It is worth being explicit about what the present work does and does not decide here. The behavioral and neuroimaging experiments test the premise these accounts *share*—that rhythmic synchronization is grounded in the speech system and has a substrate common to vocal and non-vocal effectors—rather than the point on which they *differ*. The behavioral result, in which a vocal effector synchronizes at least as well as a non-vocal one, is consistent with both accounts and, in particular, does not follow from any effector-level prediction that Patel makes. The neuroimaging result, in which the dPCSA is engaged by two non-laryngeal effectors, is consistent with Patel's proposal that connectivity spread from a laryngeal pathway, but it is not uniquely so: a general auditory-weighted hub of the kind Hickok's account would predict yields the same observation. This dissertation therefore does not adjudicate between the two hypotheses; rather, it supplies behavioral and neural evidence for the vocal-learning framework they hold in common.

The experiment that comes closest to bearing on a specific theoretical claim is the forward-entrainment study, because the Action Simulation for Auditory Prediction hypothesis occupies a dual position: it is one strand of the auditory–premotor coupling on which Patel's evolutionary argument draws, and it is at the same time the clearest statement of the motor theory of rhythm. The forward-entrainment result, being inconclusive, does not settle this claim either, but it establishes the means by which it could be settled.

5.4 The dorsal precentral speech area as a point of integration

The most substantive empirical contribution of this dissertation is the identification of the dPCSA as a region engaged by rhythmic synchronization regardless of the effector used. Be-

cause the region is auditory-weighted—it carries spectrotemporal receptive fields resembling those of auditory cortex and responds to acoustic input in the absence of an overt task—its engagement by effectors as different as the finger and the tongue is naturally read as reflecting a shared auditory input routed into the motor system, rather than the control of any particular effector. On this view the dPCSA behaves as an auditory region situated within the motor system, well placed to make auditory timing information available to multiple effectors for synchronization.

This observation connects two literatures that have largely developed separately. The dorsal premotor region implicated in the beat-processing literature and the dPCSA of the speech-coordination model may correspond to overlapping cortex; if so, a region characterized primarily in terms of pitch-related speech coordination and a region characterized in terms of beat-based timing may be, in part, the same region serving a common auditory-motor function. The present results do not establish this correspondence directly, and they leave open why such a region should be accessible to non-vocal effectors at all. But they point toward an integrated account in which speech-coordination and beat-synchronization networks share an auditory-weighted substrate, and they identify the dPCSA as a natural focus for that account.

5.5 Limitations and future directions

Several limitations cut across the dissertation as a whole. First, all three experiments used isochronous metronomic rhythms rather than music. This was a deliberate choice, made to isolate the basic auditory-motor synchronization mechanism from the additional processing required to extract a beat from complex, hierarchically structured sound. The conclusions therefore apply, in the first instance, to synchronization with simple rhythms; extending them to full beat perception and synchronization with musically structured rhythms is the

clearest next step, and would test whether the dPCSA hub, and the effector comparisons that revealed it, generalize once beat extraction is required.

Second, the behavioral and neuroimaging experiments compared vocal and non-vocal effectors but did not include a laryngeal, pitch-based effector, and none of the experiments manipulated vocal learning itself; the vocal-learning framework functions throughout as the motivating account rather than as a variable under direct experimental control. Third, the forward-entrainment experiment remains inconclusive, and its interpretation awaits the analyses identified in that chapter—subject-wise phase alignment, per-SNR averaging, separation of post-target-tap trials, and additional data at a refined range of signal-to-noise ratios—as well as a comparison against a pitch-based control task that could reveal whether the dPCSA's contribution to synchronization is specific to timing or extends to pitch.

Beyond addressing these limitations, the results motivate several further directions: characterizing the connectivity of the dPCSA with auditory and effector-specific motor regions using directed or effective-connectivity methods, so as to test the proposed auditory-to-dPCSA-to-effector pathway more directly than activation alone allows; examining whether individual differences in rhythmic ability relate to the structure or function of this circuit; and determining whether the effector-specific preferred rates suggested by the behavioral design have neural correlates.

5.6 Conclusion

Across behavioral, neuroimaging, and psychophysical experiments, this dissertation supports the view that the capacity for rhythmic synchronization is grounded in advanced vocal learning, and it identifies a shared, auditory-weighted cortical substrate—the dorsal precentral speech area—engaged by synchronization across both vocal and non-vocal effectors. At the

same time, it leaves open the question of whether rhythm perception itself depends on motor simulation, and it draws a distinction that unifies its otherwise disparate results: the evolutionary origin of a capacity and the mechanism by which it operates online are separable, and a capacity rooted in the motor system need not remain, in every respect, a motor phenomenon. Clarifying that relationship—between where beat-based synchronization came from and how rhythm is perceived in the moment—is the direction this work points toward.

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., ... Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, *8*, 14. doi: 10.3389/fninf.2014.00014

Andersson, J. L., Skare, S., & Ashburner, J. (2003). How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *NeuroImage*, *20*(2), 870–888. doi: 10.1016/S1053-8119(03)00336-7

Arnal, L. H. (2012). Predicting "when" using the motor system's beta-band oscillations. *Frontiers in human neuroscience*, *6*.

Avants, B., Epstein, C., Grossman, M., & Gee, J. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, *12*(1), 26–41. doi: 10.1016/j.media.2007.06.004

Barchet, A. V., Henry, M. J., Pelofi, C., & Rimmele, J. M. (2024). Auditory-motor synchronization and perception suggest partially distinct time scales in speech and music. *Communications Psychology*, *2*(1).

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fmri. *NeuroImage*, *37*(1), 90–101. doi: 10.1016/j.neuroimage.2007.04.042

Cannon, J. J., & Patel, A. D. (2021). How beat perception co-opts motor neurophysiology. *Trends in Cognitive Sciences*, *25*(2), 137–150.

Chakraborty, M., & Jarvis, E. D. (2015). Brain evolution by brain pathway duplication. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1684), 20150056.

Chen, J. L., Penhune, V. B., & Zatorre, R. J. (2008). Listening to musical rhythms recruits motor regions of the brain. *Cerebral cortex*, *18*(12), 2844–2854.

Chen, J. L., Penhune, V. B., & Zatorre, R. J. (2009). The role of auditory and premotor cortex in sensorimotor transformations. *Annals of the New York Academy of Sciences*, *1169*.

Chen, J. L., Zatorre, R. J., & Penhune, V. B. (2006). Interactions between auditory and dorsal premotor cortex during synchronization to musical rhythms. *Neuroimage*, *32*(4), 1771–1781.

Ciric, R., Thompson, W. H., Lorenz, R., Goncalves, M., MacNicol, E., Markiewicz, C. J., ... Esteban, O. (2022). TemplateFlow: FAIR-sharing of multi-scale, multi-species brain models. *Nature Methods*, *19*, 1568–1571. doi: 10.1038/s41592-022-01681-2

Coupé, C., Oh, Y. M., Dediu, D., & Pellegrino, F. (2019). Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science advances*, *5*(9), 2594.

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, *29*(3), 162–173. doi: 10.1006/cbmr.1996.0014

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I. segmentation and surface reconstruction. *NeuroImage*, *9*(2), 179–194. doi: 10.1006/nimg.1998.0395

Dalla Bella, S., Białuńska, A., & Sowiński, J. (2013). Why movement is captured by music, but less by speech: Role of temporal regularity. *PLoS ONE*, *8*(8), e71945.

Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., & Poeppel, D. (2017). Temporal modulations in speech and music. *Neuroscience & Biobehavioral Reviews*, *81*, 181–187.

Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A., & Gorgolewski, K. J. (2017). MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PLOS ONE*, *12*(9), e0184661. doi: 10.1371/journal.pone.0184661

Esteban, O., Blair, R., Markiewicz, C. J., et al. (2018). *fMRIPrep 25.0.0.* Zenodo. doi: 10.5281/zenodo.852659

Esteban, O., Markiewicz, C., Blair, R. W., Moodie, C., Isik, A. I., Erramuzpe Aliaga, A., ... Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*, 111–116. doi: 10.1038/s41592-018-0235-4

Evans, A., Janke, A., Collins, D., & Baillet, S. (2012). Brain templates and atlases. *NeuroImage*, *62*(2), 911–922. doi: 10.1016/j.neuroimage.2012.01.024

Farahbod, H., Saberi, K., & Hickok, G. (2020). The rhythm of attention: Perceptual modulation via rhythmic entrainment is lowpass and attention mediated. *Attention, Perception, & Psychophysics*, *82*, 3558–3570.

Fraisse, P. (1982). Rhythm and tempo. In D. Deutsch (Ed.), *The psychology of music* (pp. 149–180). Academic Press.

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in Neuroinformatics*, *5*, 13. doi: 10.3389/fninf.2011.00013

Gorgolewski, K. J., Esteban, O., Markiewicz, C. J., et al. (2018). *Nipype.* Zenodo. doi: 10.5281/zenodo.596855

Grahn, J. A., & Brett, M. (2007). Rhythm and beat perception in motor areas of the brain. *Journal of cognitive neuroscience*, *19*(5), 893–906.

Grahn, J. A., & Rowe, J. B. (2009). Feeling the beat: premotor and striatal interactions in musicians and nonmusicians during beat perception. *Journal of Neuroscience*, *29*(23), 7540–7548.

Grahn, J. A., & Rowe, J. B. (2013). Finding and feeling the musical beat: striatal dissociations between detection and prediction of regularity. *Cerebral cortex*, *23*(4), 913–921.

Greenberg, S., Carvey, H., Hitchcock, L., & Chang, S. (2003). Temporal properties of spontaneous speech—a syllable-centric perspective. *Journal of Phonetics*, *31*(3-4), 465–485.

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, *48*(1), 63–72. doi: 10.1016/j.neuroimage .2009.06.060

Henry, M. J., Obleser, J., Crusey, M. R., Fuller, E. R., Lee, Y. S., Meyer, M., & Peelle, J. E. (2025). How strong is the rhythm of perception? a registered replication of hickok et al. (2015). *Royal Society Open Science*, *12*(6), 220497.

Hickok, G. (2014). *Ww norton&company*.

Hickok, G. (2024). The "coordination conjecture" as an alternative to patel's fortuitous enhancement hypothesis for the relation between vocal learning and beat-based dancing. bmc neuroscience, 25(1), 59.

Hickok, G., Farahbod, H., & Saberi, K. (2015). The rhythm of perception: Entrainment to acoustic rhythms induces subsequent perceptual oscillation. *Psychological science*, *26*(7), 1006–1013.

Hickok, G., Venezia, J., & Teghipco, A. (2023). Beyond broca: neural architecture and evolution of a dual motor speech coordination system. *Brain*, *146*(5), 1775–1790.

Holm, T., & Isaksson, T. (2010). *Syncopes, duration and timing properties used to affect groove in monophonic melodies* (Unpublished master's thesis). Umeå University.

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, *17*(2), 825–841. doi: 10.1006/nimg.2002.1132

Jones, M. R., & Boltz, M. (1989). Dynamic attending and responses to time. *Psychological Review*, *96*(3), 459.

Katevas, K., Haddadi, H., Tokarchuk, L., & Clegg, R. G. (2015). Walking in sync: Two is company, threeś a crowd. *In Proceedings of the 2nd workshop on Workshop on Physical Analytics*, *25*.

Klein, A., Ghosh, S. S., Bao, F. S., Giard, J., Häme, Y., Stavsky, E., ... Keshavan, A. (2017). Mindboggling morphometry of human brains. *PLOS Computational Biology*, *13*(2), e1005350. doi: 10.1371/journal.pcbi.1005350

Kung, S. J., Chen, J. L., Zatorre, R. J., & Penhune, V. B. (2013). Interacting cortical and basal ganglia networks underlying finding and tapping to the musical beat. *Journal of cognitive neuroscience*, *25*(3), 401–420.

Levitt, H. C. C. H. (1971). Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical society of America*, *49*(2), 467–477.

MacDougall, H. G., & Moore, S. T. (2005). Marching to the beat of the same drummer: the spontaneous tempo of human locomotion. *Journal of applied physiology*, *99*(3), 1164–1173.

Matin, E. (1974). *Psychological bulletin*.

Matthews, T. E., Thibodeau, J. N., Gunther, B. P., & Penhune, V. B. (2016). The impact of instrument-specific musical training on rhythm perception and production. *Frontiers in Psychology*, *7*.

Matthews, T. E., Witek, M. A., Lund, T., Vuust, P., & Penhune, V. B. (2020). The sensation of groove engages motor and reward networks. *NeuroImage*, *214*, 116768.

Moelants, D. (2002). Preferred tempo reconsidered. *In Proceedings of the 7th international conference on music perception and cognition*, *2002*.

Morillon, B., & Baillet, S. (2017). Motor origin of temporal predictions in auditory attention. *Proceedings of the National Academy of Sciences*, *114*(42), E8921.

Murray, M. P., Drought, A. B., & Kory, R. C. (1964). Walking patterns of normal men. *JBJS*, *46*(2), 335–360.

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

Patel, A. D. (2021). Vocal learning as a preadaptation for the evolution of human beat perception and synchronization. *Philosophical Transactions of the Royal Society B*, *376*(1835), 20200326.

Patel, A. D. (2024). *Bmc neuroscience.*

Patel, A. D., & Iversen, J. R. (2014). The evolutionary neuroscience of musical beat perception: the action simulation for auditory prediction (asap) hypothesis. *Frontiers in systems neuroscience*, *8*.

Patel, A. D., Iversen, J. R., Chen, Y., & Repp, B. H. (2005). The influence of metricality and modality on synchronization with a beat. *Experimental brain research*, *163*, 226–238.

Patriat, R., Reynolds, R. C., & Birn, R. M. (2017). An improved model of motion-related signal changes in fMRI. *NeuroImage*, *144*, 74–82. doi: 10.1016/j.neuroimage.2016.08.051

Pellegrino, F., Coupé, C., & Marsico, E. (2011). A cross-language perspective on speech information rate. *Language*, *539*, 539–558.

Poeppel, D., & Assaneo, M. F. (2020). Speech rhythms and their neural foundations. *Nature reviews neuroscience*, *21*(6), 322–334.

Povel, D. J., & Essens, P. (1985). Perception of temporal patterns. *Music perception*, *2*(4), 411–440.

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fmri. *NeuroImage*, *84*, 320–341. doi: 10.1016/j.neuroimage.2013.08.048

Pressing, J. (2002). Black atlantic rhythm: Its computational and transcultural foundations. *Music Perception*, *19*.

Proksch, S., Comstock, D. C., Médé, B., Pabst, A., & Balasubramaniam, R. (2020). Motor and predictive processes in auditory beat and rhythm perception. *Frontiers in Human Neuroscience*, *14*, 578546.

Repp, B. H., & Su, Y. H. (2013). Sensorimotor synchronization: a review of recent research (2006–2012). *Psychonomic bulletin & review*, *20*, 403–452.

Ross, J. M., Warlaumont, A. S., Abney, D. H., Rigoli, L. M., & Balasubramaniam, R. (2016). Influence of musical groove on postural sway. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(3).

Saberi, K., & Hickok, G. (2022a). Confirming an antiphasic bicyclic pattern of forward entrainment in signal detection: A reanalysis of sun et al. (2021). *European Journal of Neuroscience*, *56*(8), 5274–5286.

Saberi, K., & Hickok, G. (2022b). A critical analysis of lin et al.'s (2021) failure to observe forward entrainment in pitch discrimination. *European Journal of Neuroscience*, *56*(8), 5191–5200.

Saberi, K., & Hickok, G. (2023). Forward entrainment: Psychophysics, neural correlates, and function. *Psychonomic Bulletin & Review*, *30*(3), 803–821.

Satterthwaite, T. D., Elliott, M. A., Gerraty, R. T., et al. (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *NeuroImage*, *64*(1), 240–256. doi: 10.1016/j.neuroimage.2012.08.052

Savage, P. E., Brown, S., Sakai, E., & Currie, T. E. (2015). Statistical universals reveal the structures and functions of human music. *Proceedings of the National Academy of Sciences*, *112*(29), 8987–8992.

Spratling, M. W. (2017). A review of predictive coding algorithms. *Brain and cognition*, *112*, 92–97.

Stupacher, J., Hove, M. J., Novembre, G., Schütz-Bosbach, S., & Keller, P. E. (2013). Musical groove modulates motor cortex excitability: A tms investigation. *Brain and cognition*, *82*(2), 127–136.

Thompson, E. C., White-Schwoch, T., Tierney, A., & Kraus, N. (2015). Beat synchronization across the lifespan: Intersection of development and musical experience. *PloS one*, *10*(6), e0128839.

Tsai, C. G., Chen, C. C., Chou, T. L., & Chen, J. H. (2010). Neural mechanisms involved in the oral representation of percussion music: An fmri study. *Brain and cognition*, *74*(2), 123–131.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4itk: Improved n3 bias correction. *IEEE Transactions on Medical Imaging*, *29*(6), 1310–1320. doi: 10.1109/TMI.2010.2046908

Van Noorden, L., & Moelants, D. (1999). Resonance in the perception of musical pulse. *Journal of New Music Research*, *28*(1), 43–66.

Varnet, L., Ortiz-Barajas, M. C., Erra, R. G., Gervain, J., & Lorenzi, C. (2017). A cross-linguistic study of speech modulation spectra. *The Journal of the Acoustical Society of America*, *142*(4), 1976–1989.

Venezia, J. H., Richards, V. M., & Hickok, G. (2021). Speech-driven spectrotemporal receptive fields beyond the auditory cortex. *Hearing Research*, *408*, 108307.

Vuust, P., & Witek, M. A. (2014). Rhythmic complexity and predictive coding: a novel approach to modeling rhythm and meter perception in music. *Frontiers in psychology*, *5*, 1111.

Witek, M. A. G., Clarke, E. F., Wallentin, M., Kringelbach, M. L., & Vuust, P. (2011). '...and i feel good!' the relationship between body movement, pleasure and rhythmic tension in groove. In *The neurosciences and music iv.*

Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, *20*(1), 45–57. doi: 10.1109/42.906424

Zhang, Y., Zou, J., & Ding, N. (2023). Acoustic correlates of the syllabic rhythm of speech: Modulation spectrum or local features of the temporal envelope. *Neuroscience & Biobehavioral Reviews*, *147*, 105111.