Time-yoked integration throughout human auditory cortex

Sam V. Norman-Haignere (samuel_norman-haignere@urmc.rochester.edu)
Biostatistics & Computational Biology, Neuroscience, Brain & Cognitive Science, Biomedical Engineering
University of Rochester Medical Center, Rochester, NY 14604 United States

Menoua K. Keshishian (mk4011@columbia.edu)
Department of Electrical Engineering, Columbia University
New York, NY 10027 United States

Orrin Devinsky (orrin.devinsky@nyulangone.org)
Department of Neurology, NYU Langone Medical Center
New York, NY 10016 United States

Werner Doyle (werner.doyle@nyulangone.org)
Department of Neurosurgery, NYU Langone Medical Center
New York, NY 10016 United States

Guy M. McKhann (gm317@cumc.columbia.edu)
Department of Neurological Surgery, Columbia University Medical Center
New York, NY 10019 United States

Catherine A. Schevon (cas2044@cumc.columbia.edu)
Department of Neurology, Columbia University Medical Center
New York, NY 10019 United States

Adeen Flinker (adeen.flinker@nyulangone.org)
Department of Neurology, NYU Langone Medical Center
New York, NY 10016 United States

Nima Mesgarani (nima@ee.columbia.edu)
Department of Electrical Engineering, Columbia University
New York, NY 10027 United States

The sound structures that convey meaning in speech such as phonemes and words vary widely in their duration. As a consequence, integrating across absolute time (e.g., 100 ms) and sound structure (e.g., phonemes) reflect fundamentally distinct neural computations. Auditory and cognitive model have often cast neural integration in terms of time and structure, respectively, but whether neural computations in the auditory cortex reflect time or structure remains unknown. To answer this question, we rescaled the duration of all speech structures using time stretching/compression and measured integration windows using a new paradigm, effective in nonlinear systems. Our approach revealed a clear transition from time- to structure-yoked computation across the layers of a popular deep neural network model trained to recognize structure from natural speech. When applied to spatiotemporally precise intracranial recordings from the human auditory cortex, we observed significantly longer integration windows for stretched vs. compressed speech, but this lengthening was very small (~5%) relative to the change in structure durations, even in non-primary regions strongly implicated in speech-specific processing. These findings demonstrate that time-yoked computations dominate throughout the human auditory cortex, placing strong constraints on neurocomputational models of structure processing.

Keywords: auditory cortex, temporal integration, intracranial EEG, natural sounds, deep neural networks

Natural sounds are composed of hierarchically organized structures that span many temporal scales, such as phonemes, syllables, and words in speech (Hickok & Poeppel, 2007) and notes, contours, and melodies in music (Patel, 2007). Importantly, the duration of these structures is highly variable (House, 1961) (**Fig 1A**), and thus there is a fundamental distinction between integrating across absolute time (e.g., 100 milliseconds, **time-yoked integration**) vs. sound structure (e.g., a phoneme, **structure-yoked integration**). If a neural population were to integrate across a structure such as a phoneme or word – or sequences of phonemes and words – then the effective integration time would necessarily vary with the duration of those structures. Auditory models have typically assumed that neural integration is tied to

absolute cortical timescales (Chi et al., 2005; Dau et al., 1997; Khatami & Escabí, 2020; Nelson & Carney, 2004), whereas cognitive and psycholinguistic models have often assumed that information integration is tied to abstract structures such as phonemes or words (Brodbeck et al., 2018; Caucheteux et al., 2023; Norris & McQueen, 2008; Temperley, 2007). Distinguishing between time- and structure-yoked integration is therefore important for relating auditory and cognitive models, building more accurate neurocomputational models of auditory processing, and interpreting findings from the prior literature.

In this project, we tested whether neural integration in the human auditory cortex reflects time or structure. Integration windows are often defined as the time window within which stimuli alter a neural response and outside of which stimuli have little effect (Theunissen & Miller, 1995). We measured integration windows using a recently developed paradigm, in which sound segments are surrounded by different "context" segments (the temporal context invariance or TCI paradigm) (Norman-Haignere et al., 2022) (**Fig 1B**). If the integration window is less than the segment duration, there will be a moment when it is fully contained within each segment and thus unaffected by surrounding context. We can therefore estimate the integration window as the smallest segment yielding a context-invariant response. Our approach does not depend on any assumptions about the features that underlie the response or the nature of the stimulus-response mapping (e.g., linearity), and thus is broadly applicable to nonlinear systems.

We tested whether neural integration windows varied with the duration of speech structures by stretching and compressing speech (preserving pitch) so as to rescale the duration of all structures (**Fig 1C**). A structure-yoked integration window should thus rescale with the magnitude of stretching/compression, irrespective of the particular structures that underlie the window (**Fig 1C**, right panel), while a time-yoked window should be invariant to stretching/compression (**Fig 1C**, left panel).

TCI paradigm effectively distinguishes time- vs. structure-yoked integration in DNN models. To test the efficacy of our approach, we first applied it to a popular deep neural network (DNN) model (DeepSpeech2), trained to recognize speech from a spectrogram representation of sound (Amodei et al., 2016; Keshishian et al., 2021) (**Fig 1D-E**). Task-trained DNNs have shown strong predictive power in sensory cortices, and have replicated important aspects of hierarchical functional organization (Kell et al., 2018; Kriegeskorte, 2015; Yamins et al., 2014), and thus provide a useful testbed for evaluating new methods and generating hypotheses for neural experiments (Kell & McDermott, 2019; Skrill & Norman-Haignere, 2023). The DNN model used here was only ever trained on natural speech (Keshishian et al., 2021).

We measured integration windows for each unit of the DNN model, after stretching and compressing speech by $\sqrt{3}$ (a factor that preserves the intelligibility in human listeners), producing a 3-fold difference in structure durations (**Fig 1D**). We found that integration windows in early DNN layers were narrow and similar for stretched and compressed speech, while integration windows in late layers were much longer and increased substantially with stretching. We computed a structure-yoking index by measuring the difference in integration windows on a logarithmic scale between stretched and compressed speech, divided by the difference in structure durations (**Fig 1E**). We observed a clear transition from time- to structure-yoked integration across DNN layers, that was completely absent from an

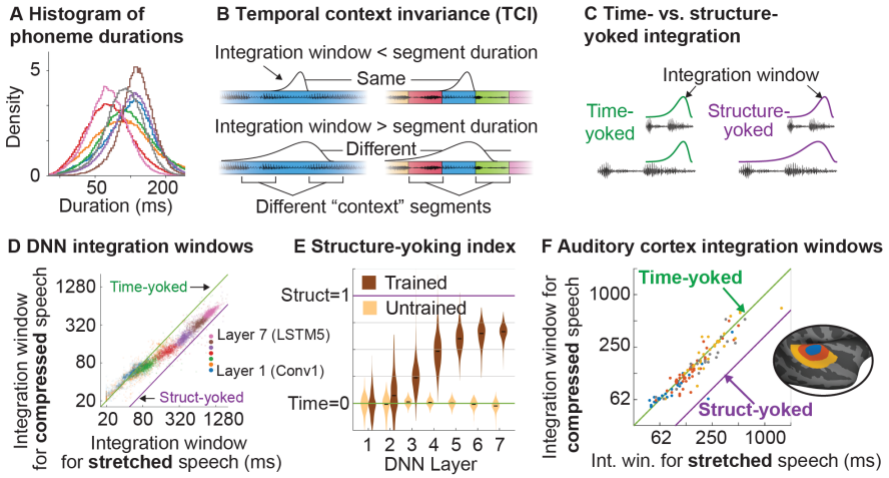

Figure 1. **A**, Duration histogram for several phonemes across a large corpus (LibriSpeech), illustrating their variability. **B**, Illustration of the temporal context invariance (TCI) paradigm. The same speech segment (in blue) is surrounded by its natural context (left) and randomly selected alternative segments (right). Context segments can only alter the response if the integration window is larger than the segment duration. **C,** Compression/stretching rescales the duration of all speech structures and will therefore compress/stretch the integration window if it reflects structure (purple, right) and not time (green, left). **D**, Integration window of DNN units for stretched and compressed speech. Purple line shows the difference in structure durations on a logarithmic scale; green is the line of unity. **E**, Distribution of structure-yoking indices across all units from each layer of a trained and untrained DNN model. **F,** Integration windows in the human auditory cortex for stretched and compressed speech. Electrodes are colored based on distance to primary auditory cortex (inset).

untrained model, demonstrating that it was learned from the durational variability of natural speech.

Neural integration throughout human auditory cortex is predominantly time-yoked. We next tested whether a similar transition to structure-yoked computation would be evident in the human auditory cortex (**Fig 1F**). We used our TCI paradigm to measure integration windows from throughout the human auditory cortex from patients undergoing intracranial monitoring for epilepsy (112 sound-responsive electrodes, 16 patients, broadband gamma power response: 70-140 Hz). We used the cortical surface distance of each electrode from primary auditory cortex (TE1.1) as a measure of its hierarchical position within the auditory cortex (Norman-Haignere et al., 2022) (inset of **Fig 1F**).

We found that the overall integration window, averaged across stretched and compressed speech, increased substantially across the cortical hierarchy replicating prior work (Norman-Haignere et al., 2022) (median integration for the annular ROIs: 75, 130, and 266 ms). We also observed a significant increase in integration windows for stretched compared with compressed speech ($F_{1,112} = 12.265, p < 0.001, \beta = 0.10$). The magnitude of this increase (0.061 octaves), however, was very small relative to the three-fold difference (1.58-octaves) in speech structure durations, yielding a structure-yoking index of only 0.04. Structure-yoking was similarly weak throughout the auditory cortex (**Fig 1F**).

These findings indicate that the primary unit of integration in the auditory cortex is absolute time and not structure duration, even in non-primary regions strongly implicated in speech-specific processing (Mesgarani et al., 2014; Norman-Haignere et al., 2015; Overath et al., 2015). How can people recognize speech structures using time-yoked integration windows, given their large durational variability (**Fig 1A**)? One possibility is that integration windows in non-primary auditory cortex are sufficiently long to achieve recognition of the relevant sound structures, even if yoked to absolute time, potentially analogous to higher-order regions of visual cortex that have large spatial receptive fields, sufficient to recognize objects across many spatial scales (Gross et al., 1969). Structure-yoked computations may also be instantiated in downstream regions (e.g., superior temporal sulcus) that integrate across longer, multi-second timescales (Lerner et al., 2011), either by enhancing weak structure-yoked computations already present in the auditory cortex or by explicitly aligning their computations to speech structures and structural boundaries (Graves et al., 2006; Norris & McQueen, 2008). The scientific findings from our study and the approach developed for distinguishing time- and structure-yoked integration will be useful in answering these questions in future research.

Acknowledgments

We thank Laura Long for help with data collection. This study was supported by the National Institutes of Health (NIDCD-K99-DC018051, NIDCD-R00-DC018051 to S.V.N.-H.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

References

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., & Chen, G. (2016). Deep speech 2: End-to-end speech recognition in english and mandarin. International conference on machine learning,

Brodbeck, C., Hong, L. E., & Simon, J. Z. (2018). Rapid transformation from auditory to linguistic representations of continuous speech. *Current Biology*, *28*(24), 3976-3983. https://www.sciencedirect.com/science/article/pii/S096098221831409X

Caucheteux, C., Gramfort, A., & King, J.-R. (2023). Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, *7*(3), 430-441.

Chi, T., Ru, P., & Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, *118*(2), 887-906. https://doi.org/10.1121/1.1945807

Dau, T., Kollmeier, B., & Kohlrausch, A. (1997). Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration. *The Journal of the Acoustical Society of America*, *102*(5), 2906-2919. https://doi.org/10.1121/1.420345

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. *Proceedings of the 23rd international conference on Machine learning*, 369-376. https://dl.acm.org/doi/abs/10.1145/1143844.1143891?casa_token=ItuJmaemKJUAAAAA:E00W4S1WSWiXxvvGT17SnY3Fq9BBV0I077U2wI49rgp2qBj1-ImktomjDhzDyp6PF9-i4mJjevMH

Gross, C. G., Bender, D. B., & Rocha-Miranda, C. d. (1969). Visual receptive fields of neurons in inferotemporal cortex of the monkey. *Science*, *166*(3910), 1303-1306.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393-402. https://www.nature.com/articles/nrn2113

House, A. S. (1961). On vowel duration in English. *The Journal of the Acoustical Society of America*, *33*(9), 1174-1178. https://asa.scitation.org/doi/abs/10.1121/1.1908941

https://asa.scitation.org/doi/pdf/10.1121/1.1908941

Kell, A. J. E., & McDermott, J. H. (2019). Deep neural network models of sensory systems: windows onto the role of task constraints. *Current Opinion in Neurobiology*, *55*, 121-132. https://doi.org/10.1016/j.conb.2019.02.003 (Machine Learning, Big Data, and Neuroscience)

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, *98*(3), 630-644. https://www.sciencedirect.com/science/article/pii/S0896627318302502

Keshishian, M., Norman-Haignere, S., & Mesgarani, N. (2021). Understanding adaptive, multiscale temporal integration in deep speech recognition systems. *NeurIPS*, *34*.

Khatami, F., & Escabí, M. A. (2020). Spiking network optimized for word recognition in noise predicts auditory system hierarchy. *PLOS Computational Biology*, *16*(6), e1007558. https://doi.org/10.1371/journal.pcbi.1007558

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, *1*, 417-446. https://doi.org/10.1146/annurev-vision-082114-035447

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, *31*(8), 2906-2915. http://www.jneurosci.org/content/31/8/2906.short

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, *343*(6174), 1006-1010. http://www.sciencemag.org/content/343/6174/1006.short

Nelson, P. C., & Carney, L. H. (2004). A phenomenological model of peripheral and central neural responses to amplitude-modulated tones. *The Journal of the Acoustical Society of America*, *116*(4), 2173-2186. https://doi.org/10.1121/1.1784442

Norman-Haignere, S. V., Kanwisher, N. G., & McDermott, J. H. (2015). Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron*, *88*(6), 1281-1296. https://doi.org/10.1016/j.neuron.2015.11.035

Norman-Haignere, S. V., Long, L. K., Devinsky, O., Doyle, W., Irobunda, I., Merricks, E. M., Feldstein, N. A., McKhann, G. M., Schevon, C. A., Flinker, A., & Mesgarani, N. (2022). Multiscale integration organizes hierarchical computation in human auditory cortex. *Nature Human Behaviour*, *6*, 455-469. https://doi.org/https://doi.org/10.1038/s41562-021-01261-y

Norris, D., & McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological review*, *115*(2), 357.

Overath, T., McDermott, J. H., Zarate, J. M., & Poeppel, D. (2015). The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nature Neuroscience*, *18*(6), 903-911. http://www.nature.com/neuro/journal/v18/n6/abs/nn.4021.html

Patel, A. D. (2007). *Music, language, and the brain*. Oxford university press. https://books.google.com/books?hl=en&lr=&id=BDS2OSt1G-MC&oi=fnd&pg=PR9&dq=patel+aniruddh&ots=S4nSQfD3eB&sig=4n7IE5Hfly9O8gQxOaypnHMez5U#v=onepage&q=patel%20aniruddh&f=false

Skrill, D., & Norman-Haignere, S. V. (2023). Large language models transition from integrating across position-yoked, exponential windows to structure-yoked, power-law windows. Thirty-seventh Conference on Neural Information Processing Systems,

Temperley, D. (2007). *Music and probability*. Mit Press. https://www.google.com/books/edition/Music_and_Probability/NS8TDgAAQBAJ?hl=en&gbpv=1&dq=temperley+model+tempo&pg=PR9&printsec=frontcover

Theunissen, F., & Miller, J. P. (1995). Temporal encoding in nervous systems: A rigorous definition. *Journal of Computational Neuroscience*, *2*(2), 149-162. https://doi.org/10.1007/BF00961885

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619-8624.

http://www.pnas.org/content/111/23/8619.short

http://www.pnas.org/content/111/23/8619.long