AudioPrism: Oscillatory frequency decomposition enhances speech recognition through multi-scale temporal processing

B. Pietras[(1)], P. Carvalho[(1,2)], I. Dubinin[(1,2)], R. Ferrand[(1)], W. Singer [(1,2,3,4)] and F. Effenberger[(1,2)]

[(1)] Natural Intelligence (NISYS GmbH), [(2)] Ernst Strüngmann Institute (ESI) of the Max Planck Society,
[(3)] Max Planck Institute for Brain Research, [(4)] Frankfurt Institute for Advanced Studies, Frankfurt am Main, Germany.

Speech encodes information across vastly different timescales: phonetic features (∼10–100 ms), syllabic structure (∼150–300 ms), and prosodic contours (>500 ms). Current recognition systems address this multi-scale challenge with deep, parameter-heavy architectures poorly suited for real-time or low-resource deployment. Biological auditory systems, by contrast, exploit oscillatory dynamics at two processing stages: the cochlea decomposes sound into frequency-specific channels via a tonotopic map, while cortical δ–γ oscillations integrate information across timescales through resonance, synchronization, and fading memory [1]—a principle largely unexploited in neural network design.

We introduce AudioPrism, a minimal two-stage architecture that mirrors this biological processing hierarchy (Fig. 1A). The first stage (PRISM) uses a bank of heavily damped harmonic oscillators with log-spaced frequencies spanning the speech bandwidth, producing a tonotopic decomposition analogous to cochlear filtering (Fig. 1C). The second stage (HORN) employs recurrently coupled oscillators with heterogeneous frequencies in the δ–γ range, whose trainable coupling matrix enables multi-scale temporal integration [2], capturing both high-frequency phonetic transients and low-frequency prosodic modulations that are important for readout and classification (Fig. 1D).

We evaluated AudioPrism on a spoken-digit classification task using the Google Speech Command dataset. To quantify the PRISM's contribution, we compared it against baseline networks lacking the initial decomposition stage. The baseline failed to learn, performing near chance level. In contrast, AudioPrism achieved >70% classification accuracy in only a few training epochs (Fig. 1E). After training, recurrent weights efficiently streamlined acoustic information from high-frequency to low-frequency nodes, increasing readout performance and mimicking the cortical hierarchy's transition from fast signal transients to slow-frequency integration at higher processing levels [1].

Oscillatory dynamics enables the network to maintain multiple temporal representations simultaneously, with different frequency channels encoding information at distinct timescales. This multi-scale integration allows AudioPrism to capture the hierarchical structure of speech efficiently. Our results demonstrate that biologically grounded oscillatory dynamics yield parameter-efficient, multi-scale speech representations with strong generalization, opening new directions for resource-efficient neural architectures [3].

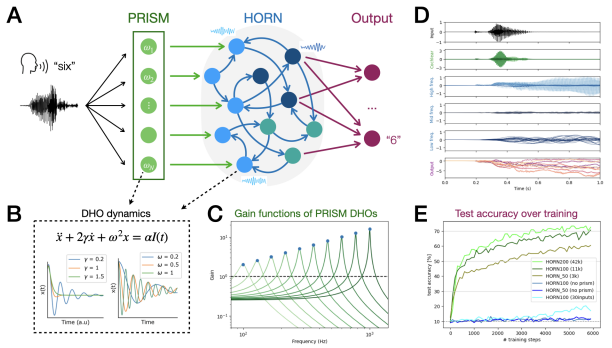

Figure 1: (A) AudioPrism = PRISM + HORN layer of damped harmonic oscillators with dynamics in (B). PRISM parameters are configured via analytic gain functions to mimic the mammalian cochlea (C). After training, the stimulus (D, top) is filtered through the PRISM and processed by the HORN, whose low-frequency components drive the output dynamics (D, bottom). Performance is measured as test accuracy for digit classification (E).

[1] A. E. Giraud and D. Poeppel, Nat. Neurosci., 15(4), 511–517, 2012.

[2] F. Effenberger et al., Proc. Natl. Acad. Sci., 122(4), e2412830122, 2025.

[3] P. Carvalho et al., Phys. Rev. Applied, 24 (6), 064055, 2025.