Staffordshire University

Department of Engineering

MSc Robotics and Smart Technologies

Biologically Plausible Curiosity-Driven Multisensory Learning in BabyAI

Master's Final Project Thesis

Author: Syed Kumail Haider

Registration No: 24024612

Email: h024612o@student.staffs.ac.uk

Supervisor: Dr. Masum Billah

August 25, 2025

Abstract

This thesis proposes a biologically plausible, curiosity-driven framework for early multimodal learning centered on audition and extensible to vision. Incoming sound is encoded on a cochlea-inspired tonotopic sheet and converted to sparse spike patterns. Synaptic change combines local rules (Hebbian co-activation, pair-based STDP, and a predictive delta update) with a dopamine-like third factor derived from an intrinsic curiosity signal that fuses normalized prediction error, "Goldilocks" spectral entropy, and deviance from a running prototype; safety gates (loudness/roughness) down-regulate plasticity. Long-term memory consists of label-indexed spectral prototypes updated by exponential moving average and time-canonicalized via dynamic time warping, with stability maintained by L_2 decay and sparsification. Open-set recognition uses temperature-scaled softmax with an abstention threshold, and a perception–production loop resynthesizes waveforms from prototypes for self-evaluation. Experiments show heightened sensitivity to novelty, incremental consolidation without retraining, invariance to speaking rate, and principled rejection of out-of-distribution inputs. Cross-modal links to vision via normalized similarity yield transient co-firing consistent with a minimal global-workspace broadcast. The result is a mechanistic, online account of early concept formation that unifies intrinsic motivation, biologically grounded plasticity, and persistent memory.

Acknowledgements

I owe my deepest gratitude to my supervisor Dr. Masum Billah whose consistent supervision and in-depth comments molded this thesis from an idea to its final version. His ability to link biological theory with algorithms that we can implement motivated me to improve both the scientific questions and technical implementation. I am particularly grateful for his accessibility at key moments, and for the extremely high bar he established for the precision and quality of the prose.

I would like to thank Staffordshire University, the School of Engineering and the MSc Robotics and Smart Technologies course for the support they have provided and the use of various facilities without which this work would not have been possible. I would like to thank the administrative and technical staff for their responsiveness, and patience. Casual conversations with classmates and friends, frequently accompanied by whiteboards and coffee, inspired many aspects of the auditory pipeline, the plasticity rules, and the memory architecture.

Other studies also made contributions to the instant project. I would like to thank the authors whose seminal work in synaptic plasticity, intrinsic motivation, and auditory modelling formed the backbone of this thesis. I want to thank the maintainers and contributors to open-source scientific software, many of whose projects provided the numerical computing libraries to implement the things I describe, and the visualization frameworks and signal-processing packages for analysis and reproduction. Their dedication to open knowledge allowed to make progress faster and add transparency to this work.

Last, but not least, I would like to thank my family and friends for their support and patience when working hard on the thesis. Special thanks to Eman who with his patience, optimism, and faith in this work during some of the most challenging periods,

kept me going.

This work did not have any specific external funding. The few mistakes, omissions and limitations that remain are solely mine.

Plagiarism Statement

I, **Syed Kumail Haider** (Registration No. **24024612**), understand that plagiarism constitutes a serious breach of academic integrity. I hereby affirm that this thesis is my own work and that I have observed the University's regulations on academic honesty. In particular, I confirm that:

1. All ideas, data, figures, tables, equations, code and text that are not my own have been **clearly acknowledged and cited** using the Harvard referencing style. Direct quotations are marked as such and accompanied by full references; paraphrased material is properly attributed.

2. Any **self-plagiarism** (reuse of my previously submitted work) has been declared and cited where applicable, and has not been submitted elsewhere for academic credit without explicit permission.

3. Any **third-party material** (including images, datasets, and software) is used in accordance with licences or permissions; sources and licences are stated in the thesis where relevant.

4. I have not engaged in **contract cheating**, ghost-writing, or any form of undisclosed collaboration that would misrepresent authorship.

5. I consent to the **electronic storage** of this thesis and to its submission to **similarity-detection systems** for the purposes of academic integrity checking. I understand that detected irregularities may be investigated under University procedures.

6. I am aware of the **penalties** for plagiarism and academic misconduct and accept full responsibility for the integrity of the work submitted.

Optional Where language-editing or generative tools were used for *writing clarity or formatting*, their use is acknowledged in the Acknowledgements; all technical content, analyses, results and conclusions are my own and have been independently verified by me.

Signature: 

Date: 25 August, 2025

Declaration

I, **Syed Kumail Haider** (Registration No. **24024612**), hereby declare that this thesis, entitled *Biologically Plausible Curiosity-Driven Auditory Learning in BabyAI*, is the result of my own independent work carried out in the Department of Engineering, Staffordshire University, in partial fulfilment of the requirements for the degree of MSc Robotics and Smart Technologies.

I confirm that:

1. This thesis has **not** been submitted, in whole or in part, for any other degree or qualification at this or any other institution.

2. All sources of information, ideas, data, figures and code that are not my own have been **appropriately acknowledged and cited**. Quotations are clearly indicated and full references are provided in the bibliography.

3. Any **third–party material** (e.g., images, diagrams, datasets, or code) reproduced in this thesis has been used in accordance with the applicable licence or with the explicit permission of the rights holder.

4. The research was conducted in accordance with **Staffordshire University's research ethics and integrity policies**. Where applicable, ethical approval was obtained prior to data collection; all participants (if any) provided informed consent.

5. The **software and experimental code** used to produce the results herein were authored by me unless stated otherwise. External libraries are credited and used under their respective licences.

6. The **data and code** supporting the findings of this study are available as described in the *Data Availability* statement (see Appendix/Repository link), subject to any

legal or ethical restrictions.

7. **Assistance and supervision:** my supervisor, Dr. Masum Billah, provided academic guidance only; the analysis, implementation, and writing are my own. Any additional assistance is acknowledged in the *Acknowledgements*.

Optional I used generative AI tools for *editing and formatting assistance only* (e.g., language clarity, structure); all technical content, results, and conclusions are my own. Any AI-generated text has been critically reviewed and verified by me.

I understand that the University may submit this thesis to text–matching software and that any breach of academic integrity will be dealt with under the University's regulations.

Author: Syed Kumail Haider **Programme:** MSc Robotics and Smart Technologies

Supervisor: Dr. Masum Billah

Signature: 

Date: 25 August, 2025

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation and Vision

Creating artificial agents that learn *in the manner of* human infants is a key challenge for cognitive robotics and computational neuroscience. Infant learners use small quantities of montaged, noisy and unsupervised sensory data to build freighted percepts and concepts; they are selective for novelty, converge upon the most useful regularities with experience, and flexibly integrate multimodal information. Modern instantiations of pattern-recognition pipelines have had great success on large supervised datasets with fixed training sets, but are ill-equipped to reproduce this developmental, curiosity-driven and *continual* learning regimen. The challenge of the present work stems from the necessity to develop a biologically plausible system that (i) represents sensory stimuli grounded in neural activity, (ii) learns in an online manner through local plasticity rules modulated by global signals, and (iii) maintains and updates longlasting stimulus prototypes without suffering catastrophic forgetting.

Audition is a particularly interesting testbed. As described in great detail in other studies, the human auditory system shows clear frequency-place mapping (tonotopy) in both the cochlea and ascending pathways, which underpins both fine temporal coding and robust perceptual organisation (Moore, 2012; Stevens et al., 1937). Computational learner that respects this organisation—by treating spectral channels as order ed neural populations and preserving temporal structur e—may better model the dynamics of infant auditory learning compared with purely task-optimised feature extractor s. Likewise, vision is formed by early receptive fields and hierarchical transformations by retinal centre–surround and LGN relay to orientation-selective populations in V1, and progressively invariant object-selective representations in IT cortex (Hubel and Wiesel, 1959, 1962;

Marr and Hildreth, 1980; Olshausen and Field, 1996; DiCarlo et al., 2012). Such biological priors justify the use of sparse, Gabor-like filtering and final tier pooling methods for visual descriptors that are analogous to possible auditory patterns.

At the scale of synaptic change, classic Hebbian learning and spike-time-dependent plasticity (STDP) are local, time-dependent rules that determine the strengthening and weakening of connections (Hebb, 1949; Bi and Poo, 1998). But local rules are not sufficient to resolve distal credit assignment, because neuromodulatory "third-factor" signals (e.g., dopamine) have been shown to gate plasticity according to prediction, salience, or outcome (Schultz et al., 1997; Frémaux and Gerstner, 2016). This drives a learning model where eligibility traces from local pre/post events are multiplied with a global modulatory term, originating from an intrinsic computation of surprise or curiosity based on the current sensory flow. This, in effect, meaning that in the current state the system *prioritizes learning what is most informative now*, mirroring developmental principles and departing from offline optimisation.

Even more than with single modalities, our cognitive faculties depend on the coordination and integration of specialised systems. The GNW perspective holds that distributed processors enter into competition to broadcast processed, or partial information for flexible, reportable use (Baars, 1988; Dehaene and Changeux, 2011; Mashour et al., 2020). While a full GNW can be beyond our scope, this thesis takes a related design position: keep interpretable, modality-specific encodings (the auditory and the visual), allow cross-modal connections in a common representational interface, and develop a lasting memory that can be strengthened or pruned depending on evidence.

Vision for *BabyAI*. Motivated by these biological and computational observations, we aim to build a small learning agent that:

1. **Encodes** audio as a tonotopic, time-resolved population activity and vision as a hierarchy of orientation-/scale-selective responses that are invariant with increasing levels (Moore, 2012; Hubel and Wiesel, 1962; Olshausen and Field, 1996; DiCarlo et al., 2012).

2. **Learns online** by local Hebbian/STDP-like updates which are gated by a global curiosity signal that integrates prediction error and structural complexity, similarly

as implied by three-factor learning theories (Hebb, 1949; Bi and Poo, 1998; Frémaux and Gerstner, 2016; Schultz et al., 1997).

3. **Merges prototypes** using time-based alignment (e.g., DTW) and exponential moving averages, and relies on principled decay to prevent interference and facilitate life-long adaptation (Sakoe and Chiba, 1978).

4. **Operates in open set**, by making calibrated confidence and thresholded decisions to abstain when considering unknown inputs, akin to real-world uncertainty handling (Cover and Hart, 1967; Guo et al., 2017; Hendrycks and Gimpel, 2017; Scheirer et al., 2013; Bendale and Boult, 2016).

5. **Is compatible with cross-modal binding** of auditory and visual codes, toward a more extensive workspace-like integration without loss of mechanistic interpretability (Baars, 1988; Dehaene and Changeux, 2011; Mashour et al., 2020).

This thesis attempts to bridge this gap between theory and practice: we translate existing neurobiological motifs (tonotopy, receptive fields, hebbian/stdp, and neuromodulation) into an implementable, real-time learning loop, and show that such a loop is capable of obtaining, retaining, and synthesising auditory patterns while creating beneficial ties to visual inputs. Building here on interpretability and biological plausibility, *BabyAI* seeks to become a platform for future developmental agents trained on the wide-range of multisensory input available in the environment and receiving close to no supervision.

1.2 Background

This chapter's objective is to provide an overview of the biological and computational underpinnings of the motivations behind the design choices in *BabyAI*. We discuss visual and auditory processing from first principles, the underlying neural coding and plasticity mechanisms for life-long learning, and a modern rationale for recognition under uncertainty and conscious-level integration.

1.2.1 Human Vision: From the Eye to the Cortex

Optical apparatus and image formation. The light is refracted by the cornea (which accounts for the major optical power) and is further refracted by the crystalline

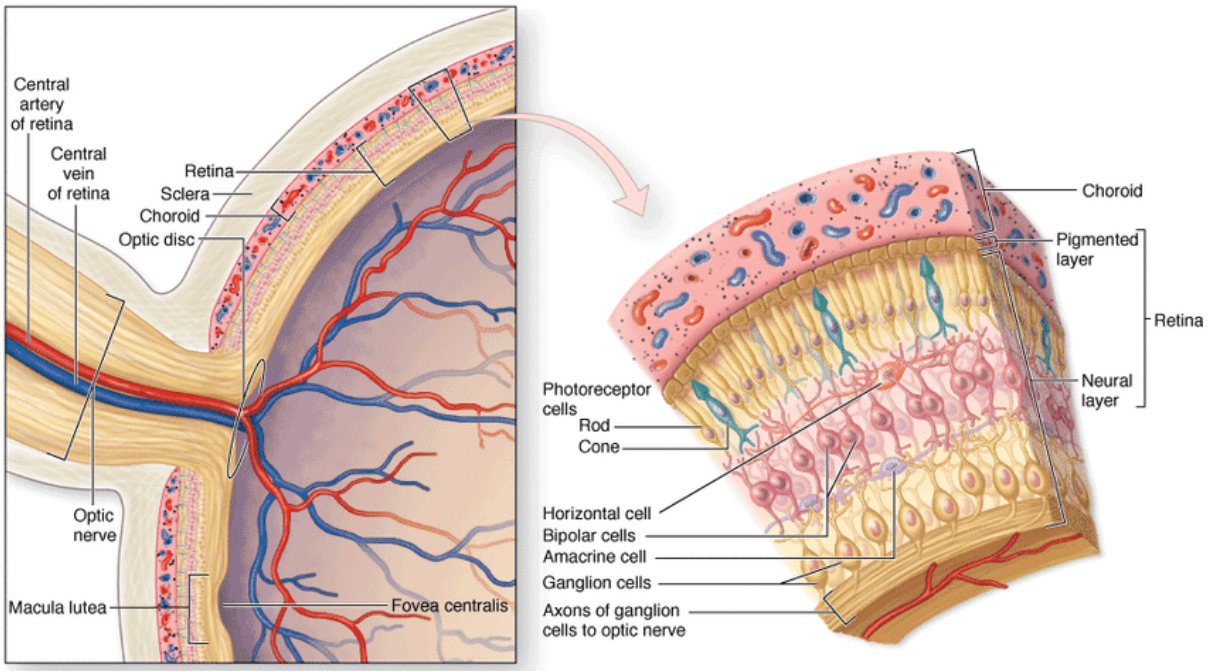

Figure 1.1: Layered anatomy of the eye and retina (adapted from Adiga (2019)).

lens, whose curvature varies during the process of accommodation to form images of objects at various distances on the retina. The pupil of the eye controls luminance, balancing photon flux with depth of field. The retinal image is inverted and include optical defects; ultimate spatial resolution is capped by photoreceptor sampling density and optics at the fovea.

Retinal architecture and transduction. The Retina is a Laminated Neural Circuit. *Rods* confer high sensitivity in scotopic (low) light; *cones* mediate photopic vision with high temporal resolution, and support trichromatic colour (L/M/S cone types). In photoreceptors, phototr Signal flow then passes to *bipolar cells* (divided according to the ON and OFF pathways), with lateral connections mediated by *horizontal cells* (spatial inhibition) and *amacrine cells* (temporal filtering, motion-related inhibition), leading to the first action potentials being generated at *retinal ganglion cell*s (RGCs). Classical centre–surround receptive fields of RGCs [5, 37] perform contrast enhancement through lateral inhibition (ON-centre/OFF-surround and vice versa), sharpening spatial edges by the level of retinal output (Marr and Hildreth, 1980; see also foundational physiology in Hubel and Wiesel, 1959, 1962). A: The fovea, consisting of closely spaced cones with the smallest receptive fields, provides maximum visual acuity; the optic disc has no photoreceptors (physiologic blind spot).

Parallel pathways and retinotopy. RGC axons form the optic nerve, partially decussate at the optic chiasm, and project topographically to the lateral geniculate nucleus (LGN) of the thalamus. Parallel channels (classically described as magnocellular and parvocellular) preserve complementary sensitivities to contrast, temporal frequency, and spatial detail. Retinotopy is largely preserved through LGN to primary visual cortex (V1), with *cortical magnification* over-representing the fovea relative to the periphery.

Primary visual cortex (V1): oriented filtering. Seminal experiments revealed that V1 neurons are tuned to oriented bars and edges (Hubel and Wiesel, 1959, 1962). *Simple cells* exhibit spatially segregated ON/OFF subregions and are phase-sensitive; *complex cells* pool over position/phase to yield tolerance to small translations. Many V1 receptive fields are well described by localised, oriented band-pass functions (*Gabor*-like filters; Gabor, 1946), capturing selectivity to orientation, spatial frequency, and phase. Multi-scale edge operators provide a complementary formulation of early vision computations (Marr and Hildreth, 1980).

Hierarchical processing and increasing invariance. Beyond V1, ventral-stream areas (V2, V4, inferotemporal cortex, IT) exhibit progressively larger receptive fields, increasing tolerance to position, scale, and clutter (Felleman and Van Essen, 1991; Riesenhuber and Poggio, 1999; DiCarlo et al., 2012). Behavioural and neurophysiological evidence indicates that selectivity and tolerance (*invariance*) jointly support robust object recognition in IT (DiCarlo and Cox, 2007; Rust and DiCarlo, 2010; DiCarlo et al., 2012). A useful computational picture is that early stages extract oriented energy across scale and phase, while later stages pool and integrate these signals into shape-selective, view-tolerant codes (Riesenhuber and Poggio, 1999; DiCarlo et al., 2012).

Statistical principles and coding efficiency. From an efficient-coding viewpoint, learning sparse, overcomplete bases on natural images yields filters resembling V1 simple-cell receptive fields (localised, oriented, band-pass) (Olshausen and Field, 1996). This links environmental statistics to cortical tuning and motivates the prevalence of orientation and spatial-frequency selectivity observed experimentally. At a broader level, computational theories emphasise the separation between the *computational* goal, *algorithmic* strategy, and *implementational* substrate in vision (Marr, 1982).

Temporal sampling and eye movements. Fixational eye movements (microsaccades, drifts) and saccades continually refresh retinal input, effectively providing temporal sampling that helps decorrelate adaptation and probe spatial structure. Neural responses thus reflect both the retinal mosaic and the dynamics of gaze; early visual neurons track temporal contrast and motion through their spatiotemporal receptive fields.

Summary. In sum, biological vision proceeds from optical formation of an image to contrast- and edge-enhanced spiking output at the retina, through retinotopic thalamic relay, to cortex where oriented, multi-scale features are extracted and progressively pooled into tolerant object representations. Key motifs include centre–surround antagonism, orientation and spatial-frequency tuning, hierarchical pooling, and efficient/sparse coding (Gabor, 1946; Hubel and Wiesel, 1959, 1962; Marr and Hildreth, 1980; Felleman and Van Essen, 1991; Olshausen and Field, 1996; Riesenhuber and Poggio, 1999; DiCarlo and Cox, 2007; Rust and DiCarlo, 2010; DiCarlo et al., 2012; Marr, 1982).

1.2.2 Human Audition: Sound to Auditory Cortex

Acoustic front end and impedance matching Airborne sound is filtered directionally and frequency dependently as an external sound wave by the *pinna* and the *external auditory canal* before it strikes the *tympanic membrane*. The middle ear *ossicular chain* (malleus–incus–stapes) imparts impedance matching from air into the fluid filled inner ear, focusing force at the stapes footplate on the oval window. The round window is a release point that allows for motion of the cochlear fluid (Moore, 2012).

Cochlear mechanics and tonotopy. Through the coiled cochlear-like structure, the basilar membrane (BM) has a graded stiffness, which is stiff and narrow at the base (high characteristic frequencies) and compliant and wide at the apex (low characteristic frequencies). A *traveling wave* is evoked along the BM by incoming sound which peaks at a location that is determined by the frequency of the stimulus and this constitutes a mechanical *place* code (tonotopy). On top of the BM is the organ of Corti, which contains a single row of *inner hair cells* (IHCs) and three rows of *outer hair cells* (OHCs). The cochlear amplifier (OHC electromotility) narrows mechanical tuning and increases sensitivity, particularly at low levels; IHCs convey the main synaptic drive to the afferent

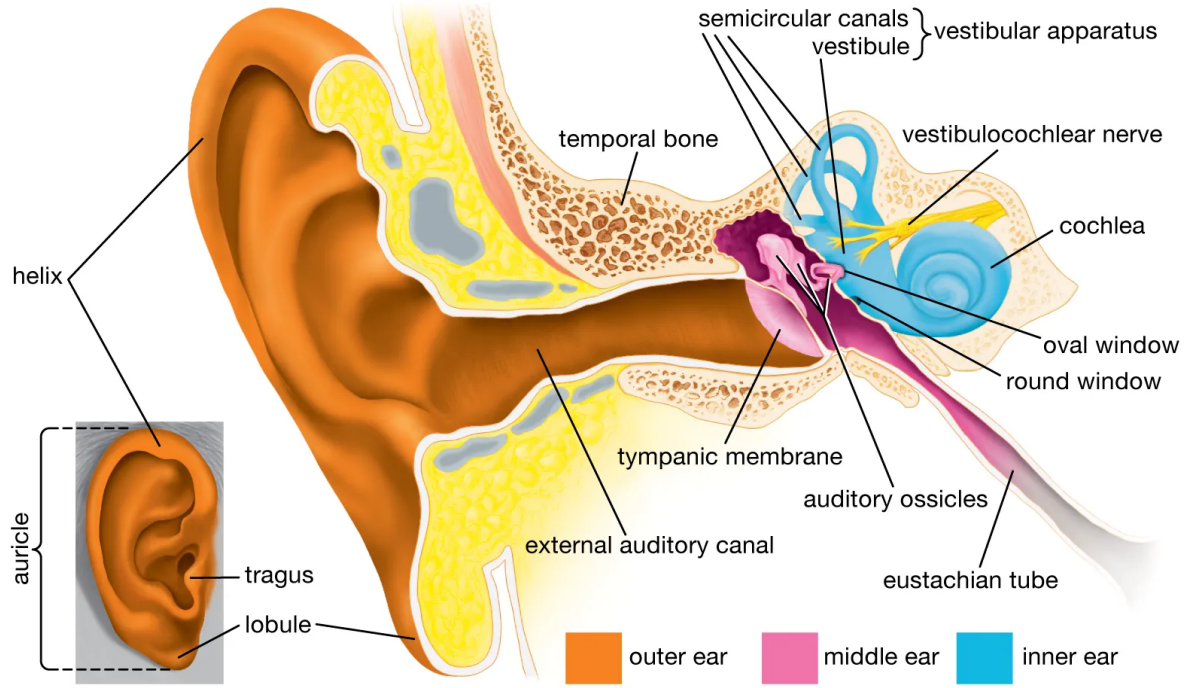

Figure 1.2: Anatomy of the human ear with the cochlea and vestibular organs

auditory nerve (Moore, 2012).

Mechanoelectrical transduction and synaptic release. Deflection of hair cell stereocilia opens MET channels through tip links, depolarising the cell in coherence with BM motion. Receptor potentials of the IHCs control neurotransmitter release at *ribbon synapses*, and result in phase-locked firing in type I AN fibres. OHCs are richly innervated by the efferent system and are also actively involved in BM mechanics via prestin based motility (Moore, 2012).

Auditory-nerve coding: rate and time. Auditory-nerve fibres are sorted in terms of CF and demonstrate both level-dependent *rate* coding and *temporal* coding, the latter based on phase-locking with the fine structure at low frequencies and to the envelope at higher frequencies. Phase locking provides accurate timing information for localization at low and middle frequencies, while spike-rate growth indicates level and spectral envelope (Moore, 2012). Physically, pitch perception squeezes frequency in something close to logarithmically; the traditional *mel* scale models this nonlinearity (Stevens et al., 1937).

Binaural processing in the brainstem. Auditory-nerve fibres relay on the *cochlear nucleus* with some of these diverging to the (SOC). Medial SOC neurons are sensitive to ITDs, while lateral SOC neurons code for ILDs—the major cues for localization in the horizontal plane. These pathways ascend through the nuclei of the lateral lemniscus to the *inferior colliculus* (IC), where monaural and binaural information is integrated and where reflexive orienting is mediated (Moore, 2012).

Thalamocortical projections and cortical tonotopy. From IC, signals are transmitted to the *medial geniculate body* (MGB) of the thalamus and further to primary auditory cortex (A1). A1 has smooth *tonotopic maps* in which neurons are typically tuned for frequency, level and the spectrotemporal pattern; neurons in the non-primary belt and parabelt areas become more selective for complex sounds and speech-like modulations. Cortical representations are capable of integrating place (tonotopy), time (phase/envelope locking) and levels cues into higher order features that form the basis for pitch, timbre and phonemic structure Moore (2012).

Computational models and analysis frameworks. So that engineered models that rely on BM frequency selecticity and bandwidth scaling so often approximate peripheral filtering with gammatone or gammachirp based filterbanks (Slaney, 1998). The analysis formalism of the short-time Fourier transform simply encapsulates the interplay between time and the frequency domain and windowed spectral estimation (see, e.g., Oppenheim and Schafer, 1989). Cepstral analyses such as MFCCs (Mel-Frequency Cepstral Coefficients) summarise spectral envelopes following a psychoacoustic frequency warping and a logarithmic compression (Davis and Mermelstein, 1980; Rabiner and Juang, 1993). Temporal representations of voiced and unvoiced signals as a sum of sine waves, i.e., the classic sinusoidal model (McAulay and Quatieri, 1986), have been contrasted with phase/vocoder techniques (Griffin and Lim, 1984). Although they are algorithmic, many of them have their origin in auditory psychophysics and peripheral physiology (e.g., filter shapes, loudness, and pitch scaling (Moore, 2012; Stevens et al., 1937; Slaney, 1998)).

Summary. Human hearing converts acoustic pressure patterns into spike trains by: (a) finely tuned mechanics of the cochlea; (b) transduction by hair cells; and (c) generation and longitudinal distribution of information along ototopic representations to the cortex,

where spectrotemporal features are integrated and abstracted. This cascade enables critical perception of pitch, loudness, timbre and spatial location in changing acoustics(l43) and (44).

1.2.3 Synaptic Plasticity: From Co-activity to Neuromodulated Change

Why plasticity? Synaptic plasticity refers to activity-dependent change in synaptic efficacy and is the biological substrate of learning and memory in nervous systems. Plasticity spans timescales from milliseconds to months and operates through multiple interacting mechanisms that tune networks to their sensory environments and behavioural outcomes.

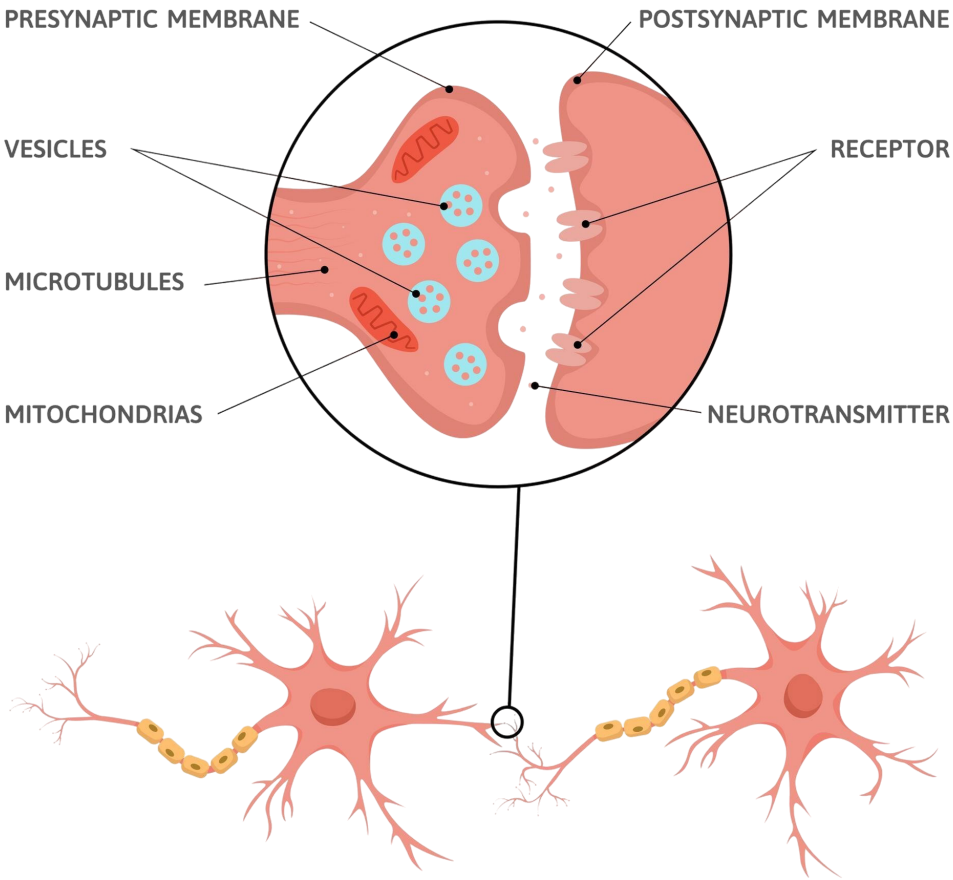

Figure 1.3: Formation of synaptic plasticity between two neurons

Hebbian co-activity and long-term change. Hebb's postulate provided a formal statement of the argument that if a presynaptic neuron *persistently and repeatedly* helps fire a post-synaptic neuron, the connection between them is (in some sense) strengthened (Hebb, 1949). In physiological terms, prolonged co-activity can elicit *long-term potentiation* (LTP) or *long-term depression* (LTD), which are lasting and reversible bidirectional changes in the synaptic gain. While it is timing independent, Hebb's rule is based on the abstract idea that correlated activity leaves associations in the synaptic matrix.

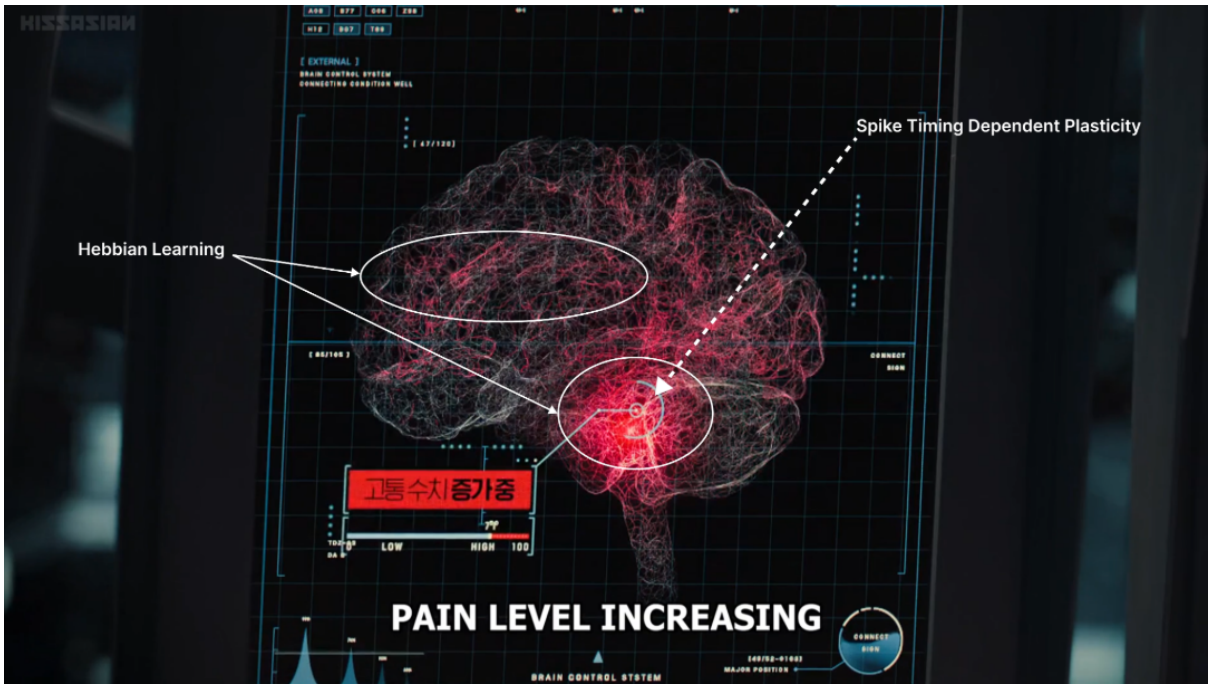

Figure 1.4: Patterns Mapping Leading to form Hebbian Learning

Spike-timing-dependent plasticity (STDP). Decades after Hebb, experiments demonstrated that the timing of spikes, at *millisecond scale*, becomes important: if *causal* pairings of presynaptic spikes/postsynaptic spikes (in which a presynaptic spike reliably precedes a postsynaptic spike) occurs, synapses potentiate, while *anti-causal pairings* (in which the two spike timing is reversed) cause depression of synaptic strength. Those effects are encoded into its purely pair-based STDP window, as was measured in hippocampal cultures, with exponentially decaying potentiation and depression lobes around $\Delta t = t_{post} - t_{pre} = 0$ (Bi and Poo, 1998). STDP establishes an association between temporal coding and synaptic change so that circuits can master predictive relations and shape temporal selectivity.

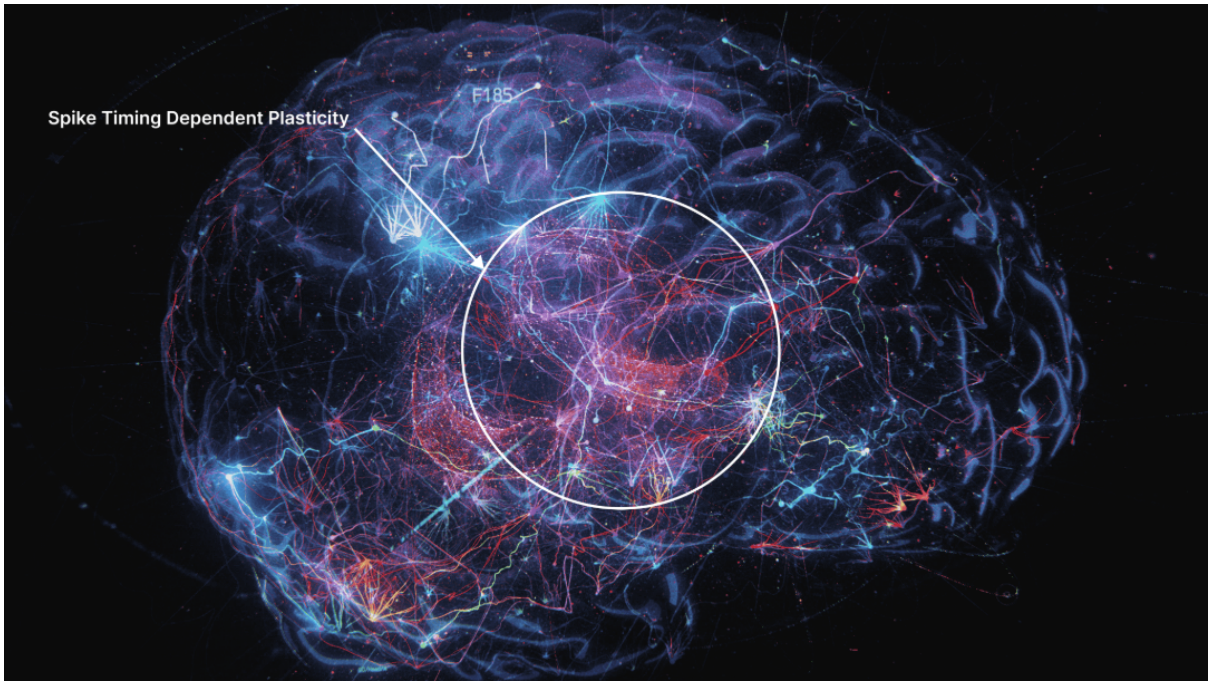

Figure 1.5: Spikes on different neurons at same time

Three-factor learning: eligibility and neuromodulation. Local pre/post (within an environment) coincidence is not enough to solve credit assignment in the real world, since (good or bad) consequences at the end can be delayed. An example of a plausible biological solution is *three-factor* plasticity: (1) pre–post interaction, which forms a transient *eligibility trace* at the synapse, (2) a strong, but global, *modulatory signal* (e.g., of dopaminergic origin) conveying information about the outcome or its saliency, and (3) the synaptic modifications occurs only when eligibility and modulation are coincident (Frémaux and Gerstner, 2016). The neurons respond with a reward-prediction error (RPE) signal—the difference between what happened and what they predicted—that can gate plasticity so only surprising or valuable events gets consolidated (Schultz et al., 1997). This transforms timing-dependent, local Hebbian/STDP mechanisms into learning of behaviourally relevant features.

Short-term dynamics and resource constraints. On time scales shorter than LTP/LTD synapses show *short-term plasticity*, however, which provides strong dynamic frequency-dependent filtering of information. On longer times, networks maintain stability via mechanisms such as synaptic normalisation and synaptic competition, which permits efficient use of limited dynamical range while retaining learned structure.

Circuit-level considerations. Plasticity rarely acts in isolation. In sensory pathways, finely-tuned excitatory-inhibitory balance determines receptive fields, whereas recurrent and feedforward connectivity interact to selectivity across space, time and frequency. Plasticity throughout these layers reshapes representations to more closely match the regularities of the sensory world, while neuromodulators control where and when such reshaping takes place.

Summary. Linking timing-dependent refinements, Hebbian association, and neuromodulated three-factor rules (Frémaux and Gerstner, 2016; Schultz et al., 1997), on a more abstract level, offers a biologically inspired set of tools. They explain how synapses store correlations, how spike timing determines circuits congenial to prediction, and how global evaluative signals make the judgment on which transient coincidences should be fated to long-term memory.

1.2.4 Recognition, Uncertainty, and Open-Set Conditions

Recognition as inference over distributed codes. In biological sensory systems, recognition emerges from populations of neurons that encode stimulus features in a distributed manner. In vision, ascending pathways—from retina to LGN to V1 and ventral stream—progressively build selectivity and tolerance (invariance) to changes in position, scale, pose and illumination (Felleman and Van Essen, 1991; Riesenhuber and Poggio, 1999; DiCarlo and Cox, 2007; DiCarlo et al., 2012; Rust and DiCarlo, 2010). Recognition can be viewed as inferring a latent category y from sensory evidence x under uncertainty, where downstream readouts pool and compare activity patterns across tuned units (e.g., via similarity to stored prototypes or decision surfaces). In audition, analogous hierarchies transform cochlear frequency–time patterns into phonetic and lexical categories, with time alignment often necessary to compare sequences of different durations (Sakoe and Chiba, 1978; Davis and Mermelstein, 1980; Rabiner and Juang, 1993).

Representations and invariances. A core challenge is to *untangle* object (or phoneme/word) manifolds so that categories become linearly separable or easily clusterable in neural space, while preserving sufficient detail for discrimination (DiCarlo and Cox, 2007; DiCarlo et al., 2012). Hierarchical feature composition and pooling explain physiological

observations of increasing receptive-field size and tolerance from early to late stages (e.g., simple/complex cells to IT in vision) (Riesenhuber and Poggio, 1999; Rust and DiCarlo, 2010). In audition, short-time spectral features (e.g., cepstra) summarise envelope and formant structure, while alignment (e.g., DTW) compensates for speaking-rate variability (Sakoe and Chiba, 1978; Davis and Mermelstein, 1980; Rabiner and Juang, 1993).

Decision confidence and uncertainty. Neural and computational decision systems must contend with noise, ambiguity, and limited evidence. A classifier producing scores $\{z_k\}$ over classes typically reports a *confidence* by mapping them through a normalised function such as softmax

$$p(y{=}k \,|\, x) \;=\; \frac{e^{z_k/T}}{\sum_j e^{z_j/T}},$$

where $T{>}0$ is a temperature that controls the sharpness of the distribution: lower T yields more peaky posteriors, higher T flattens them. In practice, many recognition systems are *miscalibrated*: their predicted probabilities do not match empirical accuracies, especially under distribution shift (Guo et al., 2017). Simple *post-hoc* calibration such as temperature scaling improves reliability of confidence without altering the decision boundary (Guo et al., 2017). From a signal-detection viewpoint, confidence supports setting operating points (e.g., ROC/PR trade-offs) appropriate to downstream costs.

Out-of-distribution (OOD) detection. Confidence also underpins detection of inputs that fall far from the training distribution. Even simple maximum softmax probability or energy scores can separate in-distribution from OOD samples surprisingly well (Hendrycks and Gimpel, 2017). Reliable OOD detection is essential in natural settings where sensors encounter novel objects, unheard words, or corrupted signals, and where deferring a decision is safer than forcing a (wrong) label.

Open-set recognition and selective rejection. The *closed-set* assumption—that every test input belongs to a known class—rarely holds outside the lab. *Open-set recognition* formalises the need to both classify known categories and *reject* unknowns (Scheirer et al., 2013). Operationally, a system sets a *rejection threshold* τ on a calibrated confidence or margin: if $\max_k p(y{=}k|x) < \tau$, the system abstains instead of over-committing. Modern treatments include open-max and EVT-based tails to better model the extreme regions of score distributions (Bendale and Boult, 2016). In nonparametric regimes (e.g., nearest

neighbours), confidence can be tied to neighbourhood density or distance ratios (Cover and Hart, 1967). Across modalities, the principle is the same: accept decisions only when evidence is sufficiently strong and familiar; otherwise, flag for deferral, acquisition of more data, or human oversight.

Sequence recognition under uncertainty (audition). Temporal variability compounds uncertainty in speech and acoustic recognition. Classical pipelines rely on robust short-time features (e.g., MFCCs) and explicit time warping or HMM-style alignment to compare variable-length sequences (Davis and Mermelstein, 1980; Rabiner and Juang, 1993; Sakoe and Chiba, 1978). Confidence can be derived from normalised alignment costs, posterior path probabilities, or calibrated posteriorgrams, again enabling abstention when sequences deviate strongly from known prototypes.

Summary. Biological and computational recognition systems balance *representational power* (selectivity and invariance) with *decision hygiene* (calibrated confidence, OOD detection, and principled rejection). Hierarchical coding (Felleman and Van Essen, 1991; Riesenhuber and Poggio, 1999; DiCarlo et al., 2012), sequence alignment (Sakoe and Chiba, 1978; Rabiner and Juang, 1993), and modern calibration/open-set methods (Guo et al., 2017; Hendrycks and Gimpel, 2017; Scheirer et al., 2013; Bendale and Boult, 2016) together provide a rigorous foundation for safe, reliable recognition under real-world uncertainty.

1.2.5 Curiosity, Intrinsic Motivation, and Neuromodulatory Gating

Conceptual role. Curiosity refers to the intrinsic drive to sample stimuli and actions that are expected to improve internal models of the world. In early development, organisms preferentially attend to signals that are neither trivial nor intractably random (a "Goldilocks" regime of complexity), allocating time and plasticity where learning progress is greatest. Functionally, curiosity shapes *what* is encoded first, *how* strongly it is consolidated, and *when* exploration should override exploitation.

Computational signals. Several measurable quantities can stand in for curiosity: (i) *novelty*, often approximated by distance from familiar prototypes or low density in feature space; (ii) *surprise* (or prediction error), the discrepancy between expected and observed sensory evidence; (iii) *uncertainty/entropy*, which favours stimuli of intermediate complexity that promise discriminative structure. These signals are not exclusive and are frequently combined to balance salience with learnability.

Neurobiological substrates. Dopaminergic systems encode teaching signals related to reward prediction error, salience, and expectancy violation (Schultz et al., 1997). These neuromodulatory transients regulate synaptic plasticity in downstream circuits, effectively *gating* when coincident pre/post activity should be consolidated. Formal three-factor learning rules capture this by multiplying local eligibility traces (from Hebbian or spike-timing coincidence) with a global modulatory factor (e.g., dopamine-like) that carries task or salience context (Frémaux and Gerstner, 2016). Thus, surprising or meaningful events induce stronger long-term changes than predictable ones, aligning synaptic updates with behavioural relevance.

Attention and orienting. Curiosity interacts with attention and orienting reflexes: novel or surprising inputs capture gaze and listening, increase gain in early sensory pathways, and bias selection in competitive circuits. This selective amplification raises the probability that informative features drive downstream neurons to threshold, making learning both faster and more energy-efficient.

Homeostasis and safety. Curiosity is tempered by homeostatic control. Excessive loudness, spectral "roughness," or sustained stressors trigger avoidance and reduction of plasticity to protect tissue and maintain stable operating ranges. From a control perspective, curiosity optimises information gain *subject to* physiological constraints, trading off exploration with safety.

Developmental implications. Across vision and audition, curiosity accelerates the acquisition of robust, invariant representations by prioritising richly structured yet learnable inputs. Coupled with neuromodulatory gating (Schultz et al., 1997; Frémaux and Gerstner, 2016), it yields a principled mechanism for when experience should be writ-

ten into long-term memory versus ignored or only transiently buffered. The quantitative forms of these signals (e.g., prediction-error normalisation, entropy targets, prototype deviation) are specified later in the Methodology, where their mathematical properties and stability constraints are laid out.

1.2.6 Memory Decay, Consolidation, and the Stability–Plasticity Trade-off

Why forgetting is functional. Biological memory is not a static archive but a continuously reorganised substrate that balances *plasticity* (to incorporate new regularities) with *stability* (to preserve useful structure). Forgetting is therefore not merely loss, but an adaptive mechanism that frees capacity, reduces interference, and biases the system toward information that remains behaviourally relevant.

Cellular and synaptic mechanisms. At synapses, multiple processes reduce effective connection strength over time when activity is absent or uncorrelated: (i) passive weight relaxation (molecular turnover and loss of potentiated states); (ii) long-term depression (LTD) and depotentiation driven by specific spike-timing and calcium dynamics that counterbalance LTP; (iii) homeostatic plasticity (e.g., synaptic scaling) that renormalises overall excitability, preventing runaway potentiation and maintaining dynamic range. These mechanisms act on different time-scales and interact with neuromodulatory control so that consolidation competes with decay depending on salience and recent usage (Frémaux and Gerstner, 2016).

Working memory vs. long-term consolidation. Short-lived, activity-dependent traces ("working memory") sustain information through recurrent excitation and transient synaptic changes; without rehearsal or reinforcement they fade within seconds to minutes. Longer-term storage relies on synaptic consolidation (hours to days) and systems-level consolidation (days to months), progressively transferring dependence from fast-learning circuits toward more stable cortical representations (Rolls, 2012). Computational treatments of working memory emphasise capacity limits, interference, and attractor stability under noise and decay (Van Vugt and Broers, 2016).

Interference and attractor dynamics. Associative networks exhibit *interference*: new patterns can overlap with and erode existing basins of attraction. Classical analyses show that capacity and robustness depend on coding sparsity and weight normalisation; forgetting (via decay or pruning) can in fact *improve* robustness by removing weak, noisy associations and sharpening attractor basins (Hopfield, 1982). Thus, moderate decay mitigates catastrophic interference while preserving strongly rehearsed structures.

Resource reallocation and pruning. Developmental and adult brains both display synaptic and dendritic pruning, reallocating resources from rarely used pathways to frequently engaged ones. Functionally, this yields a sparse, energy-efficient code where only persistently predictive links survive. Pruning works in concert with decay: weights that repeatedly fall below efficacy thresholds are removed, while reinforced connections are stabilised structurally.

Role of neuromodulators and replay. Neuromodulatory systems (e.g., dopamine, acetylcholine, norepinephrine) gate consolidation by signalling salience, novelty, and uncertainty, biasing the competition between decay and strengthening (Frémaux and Gerstner, 2016). Offline replay during sleep and quiet wakefulness reactivates recent ensembles, counteracting decay for important experiences and supporting systems consolidation (Rolls, 2012).

Stability–plasticity in practice. From a control perspective, memory decay implements a regulariser: it bounds synaptic growth, encourages sparsity, and prioritises representations that are repeatedly supported by data. Together with selective consolidation and occasional replay, it resolves the stability–plasticity dilemma by ensuring that the system remains adaptable to novelty without continually overwriting its most useful knowledge (Van Vugt and Broers, 2016; Rolls, 2012).

1.2.7 Global Workspace Theory

Global Workspace Theory (GWT) proposes that the brain comprises many specialised, mostly unconscious processors whose contents can occasionally enter a *global workspace*—a capacity-limited hub that, once ignited, *broadcasts* information widely to other systems for flexible, goal-directed use (Baars, 1988). In its neurobiological formulation (Global

Neuronal Workspace, GNW), conscious access is associated with large-scale, recurrent fronto–parietal activation that amplifies, stabilises, and disseminates selected information across distant cortical territories (Dehaene and Changeux, 2011; Mashour et al., 2020).

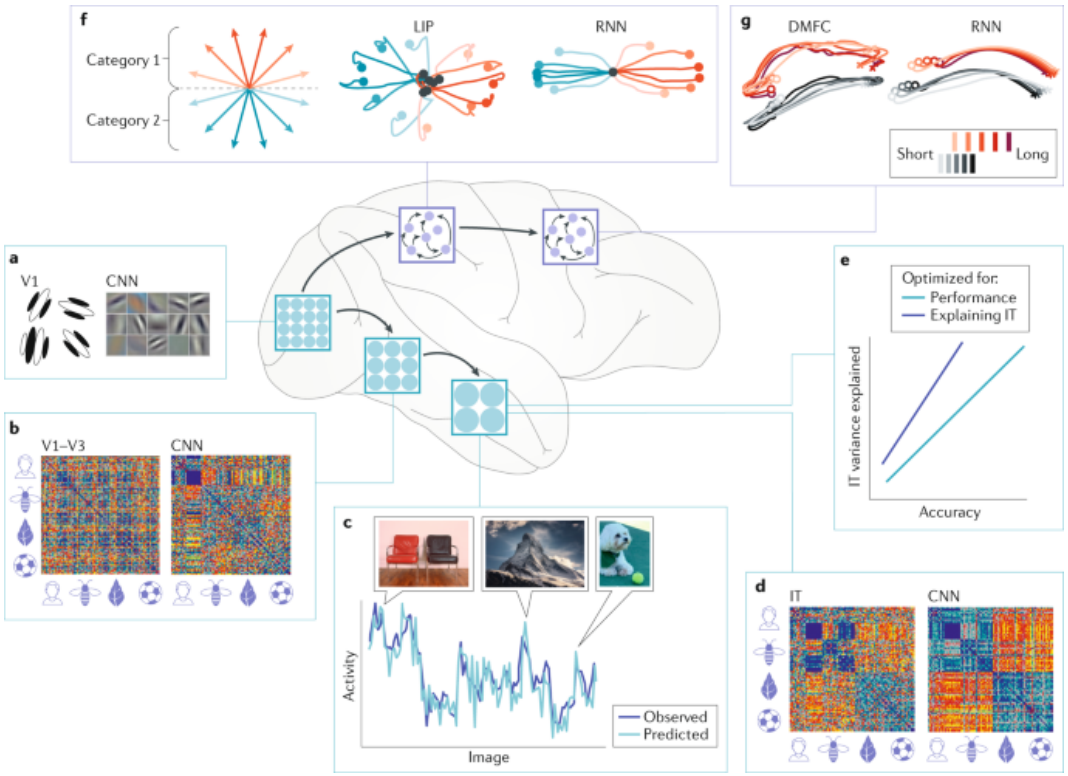

Figure 1.6: Computational architecture of Global Workspace Theory.

Core principles (conceptual).

- **Many unconscious specialists.** Sensory, mnemonic, motor, and valuation subsystems process in parallel; most activity remains local and preconscious.
- **Global workspace/broadcast.** A sparse set of high-level "workspace neurons" (long-range pyramidal cells) form a hub with widespread cortico–cortical connectivity; when a representation reaches sufficient strength and coherence, this hub *ignites* and broadcasts it globally (Dehaene and Changeux, 2011).
- **Access vs. phenomenology.** GWT/GNW focuses on *access consciousness*: the functional availability of content to reasoning, report, and voluntary control, without making strong claims about phenomenal qualities (Baars, 1988; Mashour et al., 2020).

- **Competition and threshold.** Multiple candidates vie for access; attention and top–down expectations bias the competition. Conscious access is characterised by a non-linear, threshold-like transition ("all-or-none" ignition) once recurrent amplification surpasses a critical level (Dehaene and Changeux, 2011).

Neurobiological motifs (GNW).

- **Long-range recurrent loops.** Dense reciprocal links between higher-order sensory cortices and fronto–parietal association areas support sustained, distributed activity; ignition correlates with widespread late components and large-scale synchrony (Dehaene and Changeux, 2011; Mashour et al., 2020).

- **Top–down amplification.** Prefrontal and parietal control signals (attention, task set) amplify and stabilise relevant representations, enabling report and flexible routing to memory, valuation, and action systems (Dehaene and Changeux, 2011).

- **Workspace neurons as hubs.** A minority of neurons with long corticocortical axons serve as broadcasting units, enabling cross-modal binding and coordination among otherwise modular processors (Mashour et al., 2020).

Dynamics and signatures. Empirically, conscious access is associated with (i) late, sustained activation beyond early sensory transients; (ii) widespread fronto–parietal recruitment; and (iii) enhanced effective connectivity indicating recurrent amplification and global availability (Dehaene and Changeux, 2011; Mashour et al., 2020). Failures of access (masking, attentional blink, deep anaesthesia) suppress ignition despite preserved early sensory responses, consistent with a gating role for the workspace.

Functional implications. By making selected content globally available, the workspace underwrites *flexible* cognition: cross-modal integration, sequence planning, rule application, and serial report. Attention, expectation, and goals shape which representations ignite; neuromodulatory context alters gain and stability, linking conscious access to arousal and behavioural relevance (Mashour et al., 2020).

1.3 Problem Statement

Current auditory and multitask recognition systems have reached high utility under supervision, but typically those are (i) based on large-labelled dataset, offline learning and (ii) do not incorporate online operational plasticity with biologically motivated scaling or forgetting; (iii) do not develop a long-term memory that is evolving, scaling down as less relevant information usage fades over time; and (iv) do not manage calibrated open-set decisions under uncertainty. (p. 200) Classicpipelines such as (MFCCs + HMM/DTW or deep endtoend ASR) strive for task performance over development plausibility, and no constant learning. At the level of synapses, decades of neuroscience suggest local Hebbian/STDP update rules and three-factor neuromodulated learning Hebb (1949); Bi and Poo (1998); Frémaux and Gerstner (2016); Schultz et al. (1997), but there is a large gap between these mechanisms and deployable real-time learning loops that can learn auditory patterns from sparse interaction and can be integrated with vision in a manner that is inspectable and stable across time.

Problem. Design and validate a *biologically plausible, online* learning architecture that:

1. Encodes raw audio into a cochlear/tonotopic, time–frequency representation suitable for spike-like processing (without requiring large-scale supervised training) (Oppenheim and Schafer, 1989; Slaney, 1998; Moore, 2012).

2. Updates synaptic associations *online* using Hebbian/STDP principles with a dopamine-like modulatory factor for credit assignment and salience (Hebb, 1949; Bi and Poo, 1998; Schultz et al., 1997; Frémaux and Gerstner, 2016).

3. Consolidates few-shot exemplars into durable prototypes while applying principled temporal alignment (e.g., DTW) and controlled forgetting to mitigate interference (Sakoe and Chiba, 1978).

4. Produces *calibrated*, open-set recognition decisions with explicit uncertainty control (temperature scaling and acceptance thresholds) (Guo et al., 2017; Scheirer et al., 2013; Bendale and Boult, 2016).

5. Integrates with vision in a manner consistent with hierarchical sensory processing

and large-scale broadcasting posited by Global Workspace accounts (Dehaene and Changeux, 2011; Mashour et al., 2020; Marr, 1982; Hubel and Wiesel, 1959, 1962).

The goal is a system that can *learn like an infant*: incrementally, from few exposures, prioritising salient/novel inputs, retaining what matters, and abstaining under uncertainty—while remaining transparent enough for scientific analysis.

1.4 Research Gap

Despite rich foundations in auditory signal processing and synaptic plasticity, we identify the following gaps that hinder a unified, developmentally plausible learning system:

G1 — From Plasticity Theory to an Online, End-to-End Loop. Hebbian and STDP rules (two-factor) and their neuromodulated extensions (three-factor) are well characterised experimentally and theoretically (Hebb, 1949; Bi and Poo, 1998; Frémaux and Gerstner, 2016; Schultz et al., 1997), yet practical frameworks that embed these updates into a *continuous* acquisition–consolidation–evaluation loop for real audio streams are scarce.

G2 — Curiosity/Salience as a Modulator Rather than External Reward. Most three-factor demonstrations couple modulation to extrinsic reward signals (Schultz et al., 1997; Frémaux and Gerstner, 2016). There is limited work operationalising an *intrinsic* salience/novelty signal (derived directly from the acoustics) to gate plasticity online in realistic, noisy conditions.

G3 — Few-Shot Consolidation with Temporal Variability. Classical speech pipelines acknowledge time-warping variability (DTW) (Sakoe and Chiba, 1978), but few-shot *prototype consolidation* that blends DTW-aligned traces over repeated exposures—together with principled decay to avoid catastrophic interference—remains underexplored in biologically grounded systems.

G4 — Calibrated Open-Set Decisions in Developmental Regimes. Open-set recognition and confidence calibration are well studied in discriminative settings (Guo et al., 2017; Scheirer et al., 2013; Bendale and Boult, 2016), yet there is a gap in adapting these ideas to biologically plausible, prototype-based learners that abstain when evidence is weak or unfamiliar.

G5 — Multimodal Integration with Inspectability. Vision and audition each have

mature literature—from early vision models to hierarchical object recognition (Marr, 1982; Hubel and Wiesel, 1959, 1962; DiCarlo et al., 2012) and auditory front-ends (Slaney, 1998; Moore, 2012)—but end-to-end *inspectable* pipelines that fuse both streams, expose intermediate neural-like states, and align with large-scale broadcasting principles (GNW) are limited (Dehaene and Changeux, 2011; Mashour et al., 2020).

This thesis addresses the gap by:

1. Presenting a real-time auditory front-end (tonotopic time–frequency representation) that feeds an online, biologically motivated learning loop (Hebbian/STDP with neuromodulatory gating).

2. Deriving a dopamine-like *intrinsic* modulator from signal-level salience (prediction error/structure) to prioritise surprising inputs during learning (within a three-factor framework) (Schultz et al., 1997; Frémaux and Gerstner, 2016).

3. Combining DTW-based temporal alignment with exponential moving prototypes and synaptic decay for stable few-shot consolidation under temporal variability (Sakoe and Chiba, 1978).

4. Implementing calibrated, open-set decisions using temperature scaling and acceptance thresholds to ensure reliable abstention when inputs are uncertain or out-of-distribution (Guo et al., 2017; Scheirer et al., 2013; Bendale and Boult, 2016).

5. Outlining a pathway for cross-modal integration consistent with hierarchical sensory processing and GNW-style global availability, enabling interpretability and hypothesis testing (Marr, 1982; Dehaene and Changeux, 2011; Mashour et al., 2020).

Together, these elements form a coherent, defensible contribution: a biologically informed, curiosity-modulated, few-shot learning architecture that learns online, remembers robustly, forgets gracefully, and knows when to abstain.

1.5 Research Aims, Objectives, and Hypotheses

1.5.1 Aims

1. **Design and validate a biologically plausible, online auditory learning architecture** that learns from few exposures, is modulated by an intrinsic (curiosity-like) signal, and maintains a persistent yet adaptable memory with principled forgetting and open-set, uncertainty-aware decisions.

2. **Demonstrate interpretability and multimodal readiness** by exposing inspectable intermediate states (spike/weight dynamics, prototypes) and supporting cross-modal alignment with vision for future Global Workspace–style integration.

1.5.2 Objectives

1. **Auditory front-end.** Construct a tonotopic time–frequency representation with VAD trimming and per-frame normalisation to yield stable, spike-like activity suitable for online plasticity.

2. **Spike-like encoding & synaptic learning.** Implement within-frame Hebbian co-activation and between-frame pairwise STDP; maintain local eligibility traces and a global neuromodulatory gain to form a three-factor update; include safety gates to prevent saturation on harsh/clipped inputs.

3. **Curiosity (intrinsic) modulation.** Derive a bounded modulator from prediction error, Goldilocks (entropy) structure, and deviation from a running prototype; couple this signal to plasticity strength in real time.

4. **Prototype consolidation with temporal alignment.** Canonicalise utterances via DTW to a fixed frame budget and update class prototypes with exponential moving averages; persist memory to disk with versioning.

5. **Uncertainty calibration & open-set decisions.** Apply temperature scaling to similarity scores for calibrated confidence and set an acceptance threshold τ for abstention under uncertainty/out-of-distribution inputs.

6. **Memory dynamics.** Introduce synaptic decay and sparsification to mitigate interference and catastrophic forgetting; quantify retention and stability–plasticity trade-offs.

7. **Synthesis for self-evaluation (optional).** Reconstruct waveforms from learned spectral prototypes via additive sinusoidal synthesis to close the loop and enable spectral error diagnostics.

8. **Multimodal readiness & interpretability.** Provide a minimal vision feature path and cross-modal (audio–vision) co-firing visualisation to inspect correspondences and prepare for broader integration.

9. **Evaluation protocol.** Define metrics and ablations: few-shot learning curves, novelty–plasticity coupling, retention/interference indices, open-set AUROC/FPR, calibration error (ECE/NLL), and spectral reconstruction error.

1.5.3 Hypotheses

1. **Curiosity improves sample efficiency.** Three-factor (curiosity-gated) updates achieve higher few-shot identification after the same number of exposures than ungated Hebbian/STDP baselines, and yield faster learning curves (greater area under the learning curve).

2. **Temporal alignment + EMA prototypes increase robustness.** DTW canonicalisation combined with EMA prototype updating improves matching under speaking-rate variability versus uniform resampling or no alignment, reflected in higher cosine similarity to prototypes and better F1 at fixed coverage.

3. **Decay reduces interference without erasing competence.** Synaptic decay and sparsification preserve earlier competencies during new-class learning (higher retention index, lower interference) compared with no-decay baselines.

4. **Calibration reduces misconfidence.** Temperature scaling decreases Expected Calibration Error (ECE) and negative log-likelihood (NLL), and improves risk–coverage trade-offs relative to uncalibrated scores.

5. **Open-set thresholding lowers false acceptances.** A tuned acceptance threshold τ improves open-/closed-set separation (higher AUROC/AUPR, lower FPR at target TPR) while maintaining acceptable coverage on in-distribution inputs.

6. **Curiosity–plasticity coupling is predictive.** Framewise curiosity scores positively correlate with the magnitude of weight updates and with subsequent recognition improvements for those exemplars.

7. **Safety gates enhance stability.** Loudness/roughness-based gating reduces runaway weight growth and spurious associations in noisy conditions (lower weight-variance, fewer false links) without degrading clean-signal learning.

8. **Multimodal co-firing reflects confidence.** During cross-modal inspection, stronger audio–vision co-firing patterns correlate with higher calibrated recognition confidence for matching concepts, supporting interpretability and integration readiness.

1.5.4 Primary Success Criteria (measurable)

- **Few-shot learning:** significant improvement in top-1 accuracy or F1 after ≤ 3 enrolment exposures versus ungated baseline; higher AULC.

- **Open-set:** AUROC/AUPR improvements over baseline; reduced FPR at 95% TPR (FPR_{95}).

- **Calibration:** reduced ECE/NLL with temperature scaling; improved risk–coverage curves.

- **Retention/interference:** higher retention index for earlier classes after learning new ones; lower performance drop versus no-decay.

- **Curiosity coupling:** positive correlation between curiosity score and update magnitude/learning gains.

- **Reconstruction (optional):** decreased spectral MSE between synthesised and prototype spectra across iterations.

1.6 Contributions

This thesis advances biologically plausible, curiosity-driven auditory learning through a set of conceptual, algorithmic, and evaluative contributions. The key contributions are:

1. **Curiosity–gated three–factor plasticity for online learning.** We formulate and implement a neurally inspired learning rule that combines (i) local co-activity/timing (Hebbian + pairwise STDP), (ii) an eligibility trace, and (iii) a *bounded intrinsic modulator* derived from curiosity. The modulator integrates normalised prediction error, a Goldilocks–complexity (entropy) term, and deviation from a running prototype, enabling the system to prioritise surprising but structured inputs during continuous learning.

2. **Temporal alignment and prototype consolidation.** We introduce a lightweight consolidation pipeline for variable-rate utterances: dynamic time warping (DTW) canonicalises recordings to a fixed temporal budget, and an exponential moving average (EMA) updates class prototypes online. This yields invariance to speaking rate while preserving incremental, few-shot learning behaviour.

3. **Uncertainty calibration and open–set recognition.** We provide a calibrated recognition layer using temperature scaling over similarity scores and an acceptance threshold τ to support abstention under uncertainty or out-of-distribution inputs. This separates recognition from decision, improving reliability without sacrificing biological plausibility.

4. **Persistent memory with principled forgetting.** We design a two-timescale memory mechanism: fast prototype updates (EMA) for stability under drift and slower synaptic dynamics with L_2 decay and sparsification for capacity control. This mitigates interference and catastrophic forgetting in a continual-learning regime.

5. **Interpretable neural activity visualisation.** We develop inspection utilities that expose internal states at multiple levels of abstraction: spike-like activity on a tonotopic sheet, evolving synaptic connectivity, co-firing (audio–vision) patterns, and prototype spectra over time. These views make the learning dynamics auditable and scientifically communicable.

6. **Self-evaluative spectral synthesis loop (optional).** We close the perception–action loop by reconstructing waveforms from learned spectral prototypes via additive sinusoidal synthesis, enabling quantitative spectral error diagnostics and qualitative human-in-the-loop appraisal of learned representations.

7. **Multimodal readiness and cross–modal inspection.** While the primary emphasis is audition, we provide a minimal visual pathway and a cross-modal co-firing analysis to examine correspondence between auditory and visual features. This prepares the architecture for future integration with broader cognitive frameworks (e.g., Global Workspace–style broadcasting).

8. **Evaluation protocol for developmental learning.** We specify and employ metrics that align with the developmental learning goal rather than supervised benchmark accuracy alone: few-shot learning curves and area-under-learning-curve, retention/interference indices under sequential class addition, open-set AUROC/FPR, calibration (ECE/NLL), and spectral reconstruction error.

Secondary contributions.

- A practical recipe to compute a bounded curiosity signal from prediction error, entropy, and prototype deviance that is stable in real time and compatible with three-factor plasticity.

- A compact, persistent on-disk memory representation (with versioning) supporting reproducible ablations and longitudinal experiments.

- Clear ablation studies isolating the effects of curiosity gating, temporal alignment, decay/sparsification, and calibration on few-shot performance, robustness, and reliability.

1.7 Scope, Assumptions, and Delimitations

1.7.1 Scope

This thesis investigates a *biologically plausible, curiosity–driven auditory learning* system with provisions for minimal cross–modal inspection. The project scope comprises:

- **Modality focus:** Primary emphasis on **audition**. A lightweight visual pathway is included only to support *analysis* (e.g., co–firing visualisations), not to pursue state-of-the-art visual recognition.

- **Signal path and features:** Single–channel, 16 kHz audio; short–time spectral analysis with fixed frame hop (order 100 ms); a **tonotopic sheet** of $\sim$100 frequency bands (low–to–mid kHz range) for rate–like activation and spike thresholding.

- **Learning rule:** Local Hebbian/STDP updates augmented by an **intrinsic curiosity** modulator (three–factor plasticity). Curiosity aggregates prediction error, entropy–based "Goldilocks" complexity, and deviation from a running prototype.

- **Representation and consolidation:** Per–class spectral **prototypes** updated online via exponential moving average (EMA); temporal **canonicalisation** by dynamic time warping (DTW) to handle rate variability. The canonical length T_c is configurable (typical values reported per experiment).

- **Decision layer:** Similarity–based recognition with **temperature calibration** (confidence smoothing) and an **open–set threshold** τ that permits abstention under uncertainty.

- **Memory dynamics:** Persistent on–disk store for prototypes and synaptic statistics; **weight decay** and sparsification for principled forgetting and capacity control.

- **Self–evaluation (optional):** Additive sinusoidal **synthesis** from prototypes to assess spectral fidelity; qualitative listening plus quantitative spectral error metrics.

- **Visualisations:** Interpretable, multi–level views (tonotopic activity, synaptic changes, co–firing heatmaps/3D graphs) to communicate learning dynamics.

- **Environment:** Real–time, CPU–only Python implementation suitable for laptop–class hardware; **local processing** only (no cloud).

1.7.2 Assumptions

The design and analyses rely on the following assumptions:

- **Framewise stationarity:** Spectral content is approximately stationary within analysis frames; cross–frame dynamics capture meaningful temporal structure.
- **VAD reliability:** Voice activity detection (or onset/offset heuristics) sufficiently trims silence and avoids gross truncation of relevant segments.
- **Prototype adequacy:** Class prototypes (DTW–aligned EMAs) are adequate *summaries* of categories; intra–class variability is representable as perturbations around the prototype.
- **Curiosity proxy:** The composite curiosity signal (prediction error $\times$ entropy window $\times$ prototype deviance) correlates with *learnability/salience*, justifying its use as a plasticity modulator.
- **Locality of updates:** Pairwise co–activity and short–lag timing (Hebb/STDP) approximate the relevant synaptic credit assignment at the timescales considered.
- **User–provided labels:** When a human assigns or confirms a label, it is treated as ground truth for prototype consolidation and reinforcement.
- **Acoustic channel:** Mono microphone recordings with typical indoor noise; no beamforming, multi–mic array processing, or room impulse response inversion is assumed.

1.7.3 Delimitations

Several boundaries are set deliberately to keep the study focused and tractable:

- **Not a full spiking simulation:** We do not simulate full biophysical neuron models or conductance dynamics; spike–like events are derived from rate features via thresholding for efficiency.
- **Not a benchmark ASR:** The system is not intended to compete with state–of–the–art supervised automatic speech recognition on large corpora; the target is *developmental, few–shot, online learning.*
- **Limited linguistic scope:** No phoneme dictionary, lexicon, or language model; no explicit speaker normalisation beyond DTW alignment and prototype averaging.

- **Noise robustness:** While basic robustness emerges from normalisation and curiosity gating, we do not implement advanced denoising, source separation, or adversarial defences.
- **Visual pathway minimalism:** Vision is used for feature extraction and cross–modal inspection only; no large–scale visual training or benchmarking is performed.
- **GWT as framing, not implementation:** Global Workspace Theory is used to motivate future multimodal broadcasting/selection; a full GWT architecture is not implemented here.
- **Hardware and runtime:** Experiments are constrained to CPU–only execution in Python; no GPU acceleration or embedded/robotic deployment is evaluated.
- **Ethical/data limits:** All recordings are voluntary, local, and small–scale; no personal data is transmitted off–device; no clinical claims are made.

1.8 Ethics, Data, and Reproducibility Considerations

1.8.1 Ethical Framework

This research is guided by three principles: **(i) respect for persons** (informed consent, autonomy), **(ii) beneficence** (risk minimisation, safety), and **(iii) justice** (fair access, avoidance of harmful bias). The system is designed for on–device learning, explicit user control, and *open–set abstention* to reduce overconfident, potentially harmful decisions on out–of–distribution (OOD) inputs.

1.8.2 Human Participants, Consent, and Privacy

- **Participants:** Non–identifiable convenience samples recorded by the author and a small number of adult volunteers solely for methodological evaluation (no clinical use).
- **Consent:** Written informed consent was obtained for all recorded samples. Participants could withdraw at any time before data deletion. A plain–language information sheet and consent template are provided in **Appendix A**.

- **Privacy:** No personal identifiers beyond voluntary labels (e.g., class names chosen by the user) are stored. All data remain local to the machine used for experiments; no cloud services or external transfers were employed.

1.8.3 Data Handling, Security, and Retention

- **Local storage:** Audio waveforms, derived spectral features, and learned prototypes are stored in a project folder on the local device; files are human–readable (JSON/CSV/WAV) to support auditability.
- **Minimality:** Only data necessary for experiments are retained (prototype spectra and short reference clips); raw multi–take recordings were trimmed by VAD and either *not* persisted or deleted after processing.
- **Retention:** Experimental artefacts are retained for the duration of assessment and then deleted; users can delete all data via a single folder removal operation (documented in Appendix B).
- **Security:** The project assumes a trusted single–user machine; no network endpoints are exposed. For deployments, encrypt–at–rest and access controls are recommended.

1.8.4 Bias, Fairness, and Inclusivity

- **Data scale:** Small, convenience samples cannot represent speaker diversity (age, gender, accent, language, recording conditions). Results are *not* generalised to population–level speech recognition.
- **Mitigations:** (i) open–set threshold τ to abstain under high uncertainty, (ii) temperature calibration to avoid overconfidence, (iii) reporting per–class performance rather than aggregate accuracy.
- **Disclosure:** All figures report dataset sizes and class balance to avoid misleading impressions of robustness.

1.8.5 Safety and Misuse Considerations

- **Scope limits:** The system is a *developmental learning demonstrator*, not a surveillance, authentication, or medical device. It must not be used for identification or high–stakes decisions.
- **Synthesis:** The simple additive synthesiser is intended for diagnostic listening and pedagogy, not for deceptive voice cloning. It lacks timbral fidelity by design.
- **Abstention by default:** When confidence $< \tau$, the system returns "unknown" rather than forcing a label.

1.8.6 Transparency and Interpretability

- **Model introspection:** The code provides visualisations of tonotopic activations, prototype spectra, synaptic changes, and co–firing graphs, enabling inspection of *why* a decision was made.
- **Human–readable memory:** Prototypes and parameters are saved in transparent formats to facilitate auditing and replication.

1.8.7 Environmental Considerations

- **Compute budget:** All experiments ran on CPU–only, laptop–class hardware in real time. No large–scale training or GPU cycles were consumed, keeping the carbon footprint negligible relative to deep model training.

1.8.8 Reproducibility: Data, Code, and Experiments

- **Open artefacts:** The repository structure, configuration files, and seeds are documented in **Appendix B**. Scripts recreate all figures with fixed random seeds and pinned library versions.
- **Determinism:** Preprocessing (STFT, VAD thresholds), DTW settings, and EMA rates are fixed in a single `config.py`. Any stochastic elements (e.g., synthetic noise) are seeded.

- **Versioning:** Semantic version tags are used; JSON memory files include a `meta.version` field for backward compatibility.
- **Provenance:** Each experimental run logs (i) configuration snapshot, (ii) data hashes of inputs, (iii) software versions, and (iv) figure checksums to enable byte–level reproduction.

Reproducibility checklist (artefacts to release).

1. Source code with commit hash and license.
2. `requirements.txt/environment.yml` specifying exact package versions.
3. Example audio clips (consented), or synthetic stand–ins, with hashes and licenses.
4. Configuration files used to produce each figure/table.
5. Saved prototype JSONs for each experiment.
6. Scripts/notebooks to regenerate results and plots end–to–end.

1.8.9 Ethics Review and Compliance Statement

This work involves non–sensitive, non–clinical audio collected from adult volunteers exclusively for engineering evaluation, stored locally, and analysed without personal identifiers. In consultation with institutional guidelines, it qualifies as *low–risk* research. Consent materials and a data–management statement are provided (Appendix A–B). No data were shared with third parties. No attempt was made to infer protected attributes.

1.8.10 Limitations and Residual Risks

- Limited speaker diversity may bias qualitative impressions of performance.
- Misconfiguration of τ (too low) could yield over–confident labels; recommended defaults and calibration plots are provided.
- If deployed beyond the lab, additional safeguards (rate–limit, encrypted storage, user controls, local "panic–delete") are advised.

1.9 Rationale and Justification

This thesis adopts design choices that are explicitly motivated by—and traceable to—established theoretical and empirical research in neuroscience, psychophysics, and signal processing. Below we justify each major component with primary sources and explain why the chosen path is appropriate for a biologically–plausible, developmental auditory–visual learner.

Sensory front–ends grounded in biology and signal theory

Vision. The use of local, oriented filters is motivated by classical physiology: simple and complex cells in V1 exhibit orientation and spatial–frequency tuning (Hubel and Wiesel, 1959, 1962). Gabor functions provide a near–optimal joint localisation in space and spatial frequency, long proposed as a parsimonious model of cortical receptive fields (Gabor, 1946). Early computational vision formalised edge/zero–crossing detection and multi–scale analysis (Marr and Hildreth, 1980; Marr, 1982); hierarchical feedforward models later showed how such simple units can support robust object recognition (Riesenhuber and Poggio, 1999) and invariance (DiCarlo and Cox, 2007; DiCarlo et al., 2012), in line with distributed cortical hierarchies (Felleman and Van Essen, 1991). Sparse coding accounts further explain why Gabor–like bases emerge from natural images (Olshausen and Field, 1996). These works jointly justify a V1–like, filter–bank front–end as a biologically reasonable and computationally efficient starting point.

Audition. Time–frequency analysis via windowed Fourier methods remains the backbone of auditory feature extraction (Oppenheim and Schafer, 1989). Psychophysics and auditory physiology motivate a tonotopic, approximately logarithmic frequency organisation and pitch perception (Moore, 2012; Stevens et al., 1937), while engineering toolkits have long operationalised cochlear–inspired filterbanks for analysis (Slaney, 1998). Classical speech representations (e.g., spectrograms, MFCCs) and their variants provide stable features for downstream alignment and recognition (Davis and Mermelstein, 1980; Rabiner and Juang, 1993). Our reliance on explicit, interpretable spectra (rather than opaque latent embeddings) is thus grounded in both psychoacoustics and mature signal processing literature.

Learning rules with biological plausibility

At the synaptic level, Hebb (1949) established correlation–based strengthening; later, precise spike–timing dependence was demonstrated experimentally (Bi and Poo, 1998). Neuromodulated or "three–factor" formulations unify local coincidence with a delayed, global modulatory signal (Frémaux and Gerstner, 2016), consistent with dopamine–encoded reward prediction error (Schultz et al., 1997). These lines of work jointly justify a plasticity core that is local (pre–/post–activity), temporally sensitive (STDP), and globally gated (modulator), which is essential for credit assignment and prioritisation during developmental learning.

Prototype formation and temporal invariance

Natural speech exhibits tempo variability; aligning variable–length utterances to a canonical template is a long–standing practice, with dynamic time warping (DTW) providing an effective solution (Sakoe and Chiba, 1978). Canonical–prototype representations, coupled with hierarchical feature extraction (Riesenhuber and Poggio, 1999; DiCarlo and Cox, 2007), justify our use of template consolidation with time–normalisation to obtain robust, rate–invariant exemplars.

Decision rules, calibration, and open–set safety

Nearest–neighbour style similarity is a strong baseline under small–data, nonparametric regimes (Cover and Hart, 1967). However, uncalibrated scores can be misleading; temperature scaling offers a simple and effective post–hoc calibration method (Guo et al., 2017). For safety under distribution shift, open–set recognition explicitly separates known from unknown classes (Scheirer et al., 2013; Bendale and Boult, 2016), and energy/softmax–based uncertainty signals provide practical OOD heuristics (Hendrycks and Gimpel, 2017). Together, these works justify adopting (i) similarity–based decisions, (ii) confidence calibration, and (iii) an abstention threshold to avoid over–confident errors.

Synthesis choices for interpretability

Additive sinusoidal models of speech have a long history (McAulay and Quatieri, 1986); when phase information is unavailable or unreliable, iterative phase reconstruction (e.g.,

Griffin–Lim) can be used (Griffin and Lim, 1984). We favour simple, controllable sinusoidal synthesis because it is transparent and diagnostically useful, aligning with the scientific aim of understanding learned spectral patterns rather than achieving studio–grade speech quality.

System–level integration and cognition

Finally, the architectural motivation for a selective, broadcast–capable integration layer draws from Global Workspace/Global Neuronal Workspace theory (Baars, 1988; Dehaene and Changeux, 2011; Mashour et al., 2020), which frames how specialised processors can compete for access to a global, capacity–limited workspace that coordinates learning and report. This perspective supports our longer–term integration strategy across modalities and memory systems.

Summary. Each methodological pillar—sensory front–ends (Gabor, 1946; Hubel and Wiesel, 1959; Moore, 2012), plasticity (Hebb, 1949; Bi and Poo, 1998; Schultz et al., 1997; Frémaux and Gerstner, 2016), prototype alignment (Sakoe and Chiba, 1978), calibrated open–set decisions (Guo et al., 2017; Scheirer et al., 2013; Bendale and Boult, 2016; Hendrycks and Gimpel, 2017), interpretable synthesis (McAulay and Quatieri, 1986; Griffin and Lim, 1984), and a cognition–inspired integration lens (Baars, 1988; Dehaene and Changeux, 2011; Mashour et al., 2020)—is anchored in prior research. Their combination is justified by the goal of a transparent, biologically plausible system that learns incrementally from sparse experience while remaining safe under uncertainty.

2 Literature Review

2.1 Scope and Structure

This chapter surveys the principal research strands that underpin the thesis: (i) biological and computational accounts of **vision** from retina to cortex (Hubel and Wiesel, 1959, 1962; Gabor, 1946; Marr and Hildreth, 1980; Marr, 1982; Olshausen and Field, 1996; Felleman and Van Essen, 1991; Riesenhuber and Poggio, 1999; DiCarlo and Cox, 2007; DiCarlo et al., 2012), (ii) **audition** from cochlear mechanics to engineered features and temporal alignment (Moore, 2012; Stevens et al., 1937; Oppenheim and Schafer, 1989; Slaney, 1998; Davis and Mermelstein, 1980; Rabiner and Juang, 1993; Sakoe and Chiba, 1978; McAulay and Quatieri, 1986; Griffin and Lim, 1984), (iii) biologically **plausible plasticity** and neuromodulation (Hebb, 1949; Bi and Poo, 1998; Schultz et al., 1997; Frémaux and Gerstner, 2016), (iv) **recognition, calibration, and open-set** safety (Cover and Hart, 1967; Guo et al., 2017; Hendrycks and Gimpel, 2017; Scheirer et al., 2013; Bendale and Boult, 2016), and (v) **system-level integration** inspired by Global Workspace Theory (GWT/GNW) (Baars, 1988; Dehaene and Changeux, 2011; Mashour et al., 2020). For each theme, we present foundational results, engineering formalizations, and critical observations relevant to a developmental, biologically plausible learner.

2.2 Vision: Biological and Computational Foundations

2.2.1 From Retina to Primary Visual Cortex (V1)

Light is focused by the cornea and lens onto the retina, where phototransduction in rods and cones converts photons into graded potentials. Through bipolar, horizontal, and amacrine circuits, these signals are shaped by *center–surround* antagonism and transmitted as action potentials by retinal ganglion cells (RGCs) to the lateral geniculate nucleus (LGN) and onward to primary visual cortex (V1). Classical single–unit studies in cat and monkey revealed that many V1 neurons behave as *oriented edge* detectors, with tuning for orientation, spatial frequency, and position, and that V1 is organized into columns and hypercolumns spanning these stimulus dimensions (Hubel and Wiesel, 1959, 1962). These findings grounded a view of early vision as a bank of localized, oriented filters whose responses tile visual space, orientation, and scale.

2.2.2 Early Vision as Signal Processing: Gabor and Marr–Hildreth

A principled way to model localized, orientation–selective receptive fields is through 2D Gabor functions, which achieve near–minimal joint uncertainty in space and spatial frequency (Gabor, 1946). A real, even–symmetric Gabor filter at position (x, y) and preferred orientation θ may be written:

$$g_{\theta,\lambda,\sigma,\gamma}(x,y) \;=\; \exp\Big(-\frac{x_\theta^2+\gamma^2 y_\theta^2}{2\sigma^2}\Big)\cos\Big(2\pi\frac{x_\theta}{\lambda}\Big), \qquad x_\theta = x\cos\theta + y\sin\theta,\, y_\theta = -x\sin\theta + y\cos\theta, \tag{2.1}$$

where λ is the wavelength, σ the envelope width, and γ the aspect ratio. Applying a quadrature pair (even/odd phases) and computing local energy yields phase–insensitive edge/line responses consistent with simple/complex cell phenomenology.

In parallel, Marr's framework posits that early vision computes a *primal sketch*: a multi–scale description of intensity changes, with edges detected by zero–crossings of the Laplacian–of–Gaussian (LoG) or related operators (Marr and Hildreth, 1980; Marr, 1982). Both lines of work—Gabor wavelets and LoG/derivative–of–Gaussian filter banks—provide complementary, mathematically grounded models of V1–like feature

extraction across scales.

2.2.3 Efficient and Sparse Coding Perspectives

From an information–theoretic standpoint, V1 receptive fields may arise from the pressure to represent natural images efficiently. Training a linear generative model with sparsity constraints on natural scenes produces basis functions that closely resemble oriented, bandpass, localized filters (Olshausen and Field, 1996). This result supports the efficient–coding hypothesis and unifies physiological and computational observations: localized Gabor–like atoms emerge when seeking sparse, overcomplete representations.

2.2.4 Hierarchy and Invariance in the Ventral Stream

Anatomical and physiological evidence indicates a distributed, hierarchical organization from V1 through V2/V4 to inferotemporal cortex (IT) (Felleman and Van Essen, 1991). Computationally, feedforward models such as HMAX instantiate alternating stages of *selectivity* (template/feature matching) and *tolerance* (pooling over position/scale), building progressively invariant representations (Riesenhuber and Poggio, 1999). Systems neuroscience links this hierarchy to behavior: object recognition requires *untangling*—transforming image manifolds so that object identity can be linearly read out despite nuisance variation (position, scale, pose, illumination) (DiCarlo and Cox, 2007). Empirical work in monkeys shows that IT neurons exhibit high selectivity with substantial tolerance to transformations, consistent with a culminating stage of the ventral stream where invariance is explicit (Rust and DiCarlo, 2010; DiCarlo et al., 2012).

2.2.5 Computational Ingredients Commonly Adopted

Drawing together these strands, biologically grounded vision pipelines typically comprise:

1. **Multi–scale, oriented filtering** (Gabor/DoG/LoG banks) to approximate V1 simple–cell responses (Gabor, 1946; Marr and Hildreth, 1980; Olshausen and Field, 1996).

2. **Nonlinear rectification and energy pooling** to model complex–cell invariance (e.g., quadrature energy, max/average pooling).

3. **Spatial pooling and normalization** to gain moderate translation/scale tolerance and contrast management.

4. **Hierarchical composition** (feature → part → object), consistent with ventral stream organization (Felleman and Van Essen, 1991; Riesenhuber and Poggio, 1999; DiCarlo et al., 2012).

These components provide interpretable, data–efficient feature extractors that align with classic physiology while remaining tractable for engineering.

2.2.6 Critical Appraisal and Relevance

Biologically inspired early–vision models offer transparency, controllability, and strong inductive biases, particularly valuable in low–data or developmental settings. However, relative to modern deep architectures, handcrafted banks may underperform on large, unconstrained recognition tasks and require careful parameterization (e.g., scale pyramids, normalization schemes). Hybrid approaches—combining biologically grounded front–ends with learned downstream pooling or classifiers—are a pragmatic compromise that preserve interpretability while improving performance. Crucially, the literature reviewed here supplies the theoretical basis for oriented multi–scale analysis, energy–based pooling, and hierarchical tolerances that underpin many modern, yet biologically aligned, visual processing pipelines (Hubel and Wiesel, 1959, 1962; Gabor, 1946; Marr, 1982; Olshausen and Field, 1996; Riesenhuber and Poggio, 1999; DiCarlo et al., 2012).

2.3 Audition: Biological and Computational Foundations

2.3.1 From Outer Ear to Auditory Nerve: Mechanics and Transduction

Acoustic pressure waves are funneled by the pinna and ear canal (outer ear) to the tympanic membrane, whose vibrations are impedance–matched to cochlear fluids by the ossicles (malleus–incus–stapes) in the middle ear. The stapes drives the oval window, launching a traveling wave along the basilar membrane (BM) within the fluid–filled cochlea.

Owing to its graded stiffness and width, the BM implements a *tonotopic* map: high frequencies peak near the base, low frequencies near the apex. Inner hair cells (IHCs) transduce BM motion into receptor potentials via mechano–electrical transduction in stereocilia; neurotransmitter release at ribbon synapses drives spiking in auditory nerve fibers (ANFs). ANF firing shows *phase locking* to low– and mid–frequency components (supporting temporal coding), while *place* (tonotopic) coding is expressed by the characteristic frequency of each fiber. Central pathways relay these spike trains through cochlear nuclei, superior olivary complex (binaural cues), inferior colliculus, and medial geniculate body to primary auditory cortex (A1), where frequency maps and spectro–temporal selectivity emerge (overview in Moore, 2012).

2.3.2 Temporal, Place, and Pitch Perception

Psychophysical work established relations between frequency and perceived pitch (e.g., Mel scale), reflecting combined temporal/place mechanisms and the ear's nonuniform frequency sensitivity (Stevens et al., 1937). At low frequencies, temporal fine structure and envelope cues support pitch via phase locking; at higher frequencies, place coding dominates as phase locking declines. This neurophysiology underpins filterbank designs and psychoacoustic frequency scales used in computational analysis.

2.3.3 Signal Processing View: Time–Frequency Analysis

A foundational representation is the short–time Fourier transform (STFT) of a windowed signal $y[n]$:

$$X(t,\omega) \;=\; \sum_{n} y[n]\, w[n-t]\, e^{-j\omega n}, \tag{2.2}$$

whose magnitude $|X(t,\omega)|$ yields a spectrogram (Oppenheim and Schafer, 1989). To approximate cochlear filtering, one applies a *filterbank* $H_k(\omega)$ across frequency and computes band energies

$$E_k(t) \;=\; \sum_{\omega} |H_k(\omega)|^2\, |X(t,\omega)|^2, \tag{2.3}$$

with $k = 1, \ldots, K$ bands laid out on a perceptual (e.g., Mel) or auditory (e.g., gammatone) scale. The *Auditory Toolbox* popularised practical designs for gammatone filterbanks and auditory spectrograms suitable for modeling early auditory processing (Slaney, 1998).

Mel–frequency cepstral coefficients (MFCCs). MFCCs summarise the spectral envelope by (i) mapping the linear frequency axis to Mel–spaced triangular filters, (ii) taking $\log E_k(t)$ to approximate loudness/compression, and (iii) applying a discrete cosine transform (DCT) to decorrelate:

$$c_m(t) \;=\; \sum_{k=1}^{K} \log E_k(t)\; \cos\!\Big[\frac{\pi m}{K}\,(k-12)\Big], \qquad m = 0, \ldots, M-1, \tag{2.4}$$

often augmented with temporal derivatives (Δ, Δ^2) to capture dynamics (Davis and Mermelstein, 1980; Rabiner and Juang, 1993). MFCCs and related cepstral features remain strong baselines for speech and general audition tasks.

Cochleagram and gammatone features. An alternative to MFCCs preserves a more biologically faithful time–frequency tiling using gammatone (or gammachirp) filters spaced on an auditory (ERB) scale, with half–wave rectification and compression to approximate IHC transduction and auditory nerve firing rates. The resulting *cochleagram* offers better temporal fidelity and explicit tonotopy than low–dimensional cepstra (Slaney, 1998).

2.3.4 Temporal Dynamics and Alignment

Speech and environmental sounds exhibit rate variability and local timing fluctuations. Classic *dynamic time warping* (DTW) aligns two sequences by minimizing accumulated frame–wise distances subject to monotone path constraints (Sakoe and Chiba, 1978). DTW enables template–based recognition robust to speaking rate and local elastic deformations, and remains a useful tool for alignment and canonicalization of time–frequency trajectories.

2.3.5 Analysis–by–Synthesis and Source–Filter Models

Speech signals are well described by a *source–filter* model: a quasi–periodic glottal source (pitch f_0) filtered by vocal–tract resonances (formants). Sinusoidal modeling represents frames as sums of time–varying sinusoids with amplitude and frequency trajectories, enabling high–fidelity analysis–synthesis (McAulay and Quatieri, 1986). When only magnitude spectra are available, consistent phase can be iteratively estimated (Griffin–Lim)

to reconstruct time–domain signals (Griffin and Lim, 1984). These frameworks link statistical features (e.g., envelopes, formants) to generative synthesis and are widely used in speech coding and TTS (Rabiner and Juang, 1993).

2.3.6 Pattern Recognition Perspectives

Template and nearest–neighbor methods on time–frequency features (spectrograms, cochleagrams, cepstra) provide interpretable baselines and are theoretically grounded in nonparametric classification (Cover and Hart, 1967). Modern systems often combine perceptually motivated front–ends (e.g., Mel/gammatone) with statistical learning back–ends (HMMs, DNNs), but the classical pipeline remains instructive: perceptual filtering $\rightarrow$ logarithmic compression $\rightarrow$ decorrelation/dimensionality reduction $\rightarrow$ sequence modeling (Rabiner and Juang, 1993).

2.3.7 Critical Appraisal and Relevance

Biologically aligned front–ends (gammatone/cochleagram, explicit tonotopy, compression, and rectification) capture peripheral auditory mechanics and yield features with good inductive bias for real–world sounds (Slaney, 1998; Moore, 2012). STFT–based cepstra (MFCCs) offer compactness and strong performance across tasks but abstract away fine temporal structure and detailed cochlear dynamics (Davis and Mermelstein, 1980; Rabiner and Juang, 1993). DTW provides rate–invariant alignment essential for variable–tempo utterances (Sakoe and Chiba, 1978). Analysis–by–synthesis models (sinusoidal, Griffin–Lim) bridge descriptive features and generative reconstruction (McAulay and Quatieri, 1986; Griffin and Lim, 1984). Together, these lines of work supply a mature toolkit that balances biological plausibility with computational tractability in auditory representation and recognition.

2.4 Synaptic Plasticity: Hebbian Learning, STDP, and Neuromodulation

2.4.1 Hebbian Co-activity

The modern study of learning in neural circuits begins with Hebb's postulate that synapses strengthen when presynaptic activity repeatedly or persistently contributes to postsynaptic firing (Hebb, 1949). Hebbian learning captures a core intuition about association—"cells that fire together, wire together"—and explains how co-activated features can become bound in cortex. Although Hebb's proposal was qualitative, it paved the way for quantitative plasticity rules that relate synaptic change to measurable spike events.

2.4.2 Timing Matters: Spike–Timing–Dependent Plasticity (STDP)

Physiological experiments established that the *timing* between presynaptic and postsynaptic spikes is critical for the *sign* and *magnitude* of synaptic change (Bi and Poo, 1998). In canonical pair–based STDP, presynaptic spikes that arrive slightly *before* postsynaptic spikes (causal order) induce long–term potentiation (LTP), whereas the reverse order induces long–term depression (LTD), with sensitivity that decays over tens of milliseconds. This asymmetric timing "window" supplies a biologically grounded mechanism for extracting temporal causality from spike trains and for shaping feedforward and recurrent microcircuits. Variants (e.g., nearest–neighbor, all–pairs, and triplet STDP) preserve the same qualitative dependence on relative timing while adapting to different firing regimes and synapse types; across these formulations, the core principle is that precise pre/post spike relations drive incremental weight changes that accumulate over experience.

2.4.3 From Two–Factor to Three–Factor Rules

Hebbian and pairwise STDP are *local* two–factor rules: synaptic updates depend on pre- and postsynaptic activity alone. Such locality makes them biologically plausible but limits their ability to assign credit for delayed outcomes. Converging theoretical and experimental work shows that neuromodulators (notably dopamine) provide a *third factor*

that gates plasticity as a function of behavioural salience or prediction error (Schultz et al., 1997; Frémaux and Gerstner, 2016). In three–factor learning, brief spike–based *eligibility traces* tag synapses when pre/post coincidences occur; if a modulatory signal arrives within a suitable time window, the tagged synapses are consolidated (potentiated or depressed) proportionally to the modulator. This mechanism solves distal credit assignment in a biologically realistic manner by decoupling fast local coincidence from slower global evaluation.

2.4.4 Eligibility Traces and Temporal Credit Assignment

Eligibility traces implement a synaptic memory of recent co-activity: they rise with spike coincidences and decay on a short timescale. When a phasic neuromodulatory burst (e.g., dopamine release signalling reward prediction error) follows, the trace is converted into a durable weight update (Schultz et al., 1997; Frémaux and Gerstner, 2016). This separation of timescales allows circuits to remain sensitive to precise spike timing while deferring the decision to consolidate until behavioural relevance is known, thereby aligning plasticity with learning objectives without abandoning locality.

2.4.5 Homeostasis, Stability, and Competition

Left unchecked, purely potentiating rules can cause runaway weights. Biological circuits therefore express homeostatic mechanisms (e.g., synaptic scaling, LTD components, and activity–dependent decay) and competitive interactions (e.g., lateral inhibition) that stabilise firing rates and promote sparse, selective codes. Although the exact mechanisms vary by system, the general theme is a balance between plasticity and stability: potentiation is countered by use–dependent depression and/or slow normalization, maintaining network operating points within functional ranges while preserving learned structure.

2.4.6 Implications for Auditory and Visual Learning

In sensory pathways, Hebbian/ STDP dynamics can bind co–occurring features within a frame (Hebbian co-activity) and shape temporal predictions across frames (causal STDP). When coupled to a modulatory signal reflecting behavioural significance or surprise, these local updates preferentially consolidate informative or salient experiences (Frémaux and

Gerstner, 2016). This triadic interaction—local coincidence, precise timing, and global modulation—offers a principled route to acquiring stable yet adaptable feature representations in early sensory hierarchies, consistent with the broader view that dopamine–like signals encode deviations from expectation to guide learning (Schultz et al., 1997).

2.4.7 Critical Appraisal

Hebbian and STDP rules are strongly supported at synaptic timescales and provide implementable learning dynamics with minimal assumptions (Hebb, 1949; Bi and Poo, 1998). Three–factor formulations integrate these local rules with behaviourally meaningful feedback while preserving biological plausibility (Frémaux and Gerstner, 2016). Open challenges include characterising how heterogeneous synapses and interneuron classes coordinate plasticity across layers, and how multiple neuromodulators (dopamine, acetylcholine, noradrenaline) interact to regulate attention, uncertainty, and consolidation in complex tasks. Nevertheless, the core ingredients—timing–sensitive coincidence, eligibility traces, and modulatory gating—form a robust conceptual foundation for modelling learning in sensory systems and beyond.

2.5 Recognition Confidence, Calibration, and Open–Set Conditions

2.5.1 Why Calibration Matters

Modern recognition systems typically output a *score* or a putative *probability* for each class. In practice, these probabilities are often *miscalibrated*—the reported confidence does not match empirical correctness (e.g., predictions at 90% confidence may be correct only 70% of the time). Guo et al. (2017) showed that state-of-the-art deep networks tend to be *overconfident*, and proposed *temperature scaling* (TS) as a simple, effective post-hoc fix applied to the pre-softmax logits $\mathbf{z}$:

$$p_i \;=\; \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}, \qquad T > 0, \tag{2.5}$$

where T is selected on a held-out set to minimise calibration error. Calibration quality is commonly summarised by the *Expected Calibration Error* (ECE), the bin-wise weighted gap between predicted confidence and empirical accuracy:

$$\mathrm{ECE} \;=\; \sum_{m=1}^{M} \frac{|B_m|}{n} \,|\,\mathrm{acc}(B_m) - \mathrm{conf}(B_m)|, \tag{2.6}$$

for bins B_m partitioning predictions by confidence (Guo et al., 2017). Reliability diagrams visualise these gaps by plotting accuracy vs. confidence per bin.

2.5.2 Uncertainty, Abstention, and Risk

A calibrated confidence enables *selective classification*: the system *abstains* when confidence falls below a threshold τ, trading coverage for accuracy. This is crucial in safety-critical or open-world settings where the true class may be absent from the training label set. While many models equate confidence with the maximum class probability, calibration ensures this quantity better reflects actual correctness risk, making threshold selection more principled.

2.5.3 Out-of-Distribution (OOD) Detection

When test inputs differ from the training distribution (*out-of-distribution*), standard classifiers can be confidently wrong. Hendrycks and Gimpel (2017) established a strong baseline: the *maximum softmax probability* (MSP) tends to be lower for OOD inputs than for in-distribution, enabling a simple threshold test for OOD detection. Despite its simplicity, MSP remains competitive and serves as a reference for evaluating more advanced OOD methods. In practice, MSP can be combined with temperature scaling to further improve separability between in- and out-of-distribution confidence profiles.

2.5.4 Open–Set Recognition (OSR)

Open–set recognition formalises the requirement that a classifier must *both* label known classes and *reject* unknowns without incurring excessive "open space risk" (Scheirer et al., 2013). Unlike generic OOD detection, OSR emphasises decision functions that decay away from known support and explicitly allocate probability mass to an `unknown` class

in ambiguous regions.

A seminal practical approach is **OpenMax** (Bendale and Boult, 2016). OpenMax replaces the conventional SoftMax layer with a meta-recognition step: for each known class, fit an Extreme Value Theory (EVT) model (e.g., Weibull) to the tail of the activation distances. At test time, large activation distances are partially reallocated from the predicted class to an `unknown` bucket via EVT-based probability correction, yielding improved rejection of unfamiliar inputs while maintaining accuracy on knowns.

2.5.5 Nonparametric Recognition and Score Calibration

Nonparametric methods (e.g., k-nearest neighbours; Cover and Hart 1967) classify by similarity in feature space rather than by parametric logits. Their raw *similarity scores* (cosine, Euclidean, DTW alignment, etc.) are not probabilities, but can be *monotonically calibrated*—for example, by passing scores through temperature-scaled SoftMax over class prototypes, or by fitting a monotone mapping on validation data (e.g., Platt-style logistic calibration, isotonic regression). This yields a single *confidence* that can be thresholded for abstention and used to compute ECE and reliability diagrams, just as with parametric classifiers (Guo et al., 2017).

2.5.6 Putting It Together: A Practical Recipe

Drawing on the literature above, a robust recognition stack in open environments typically:

1. Produces a *similarity vector* $\mathbf{s}$ to class prototypes (parametric logits or nonparametric distances converted to scores).

2. Applies *temperature scaling* to obtain calibrated class probabilities $\mathbf{p}$ (Guo et al., 2017).

3. Uses the *maximum probability* $\max_i p_i$ as a calibrated confidence for *selective prediction*; abstain if $\max_i p_i < \tau$.

4. Implements an *OOD/OSR guard*: MSP thresholding as a baseline (Hendrycks and Gimpel, 2017); optionally, EVT–based OpenMax to redistribute mass to `unknown` and reduce open space risk (Bendale and Boult, 2016; Scheirer et al., 2013).

5. Monitors *calibration* with ECE and reliability diagrams; tunes T and τ on a validation set comprising both in-distribution and proxy-unknown samples (Guo et al., 2017).

2.5.7 Critical Perspective

Temperature scaling is simple and effective but *post hoc*; it cannot fix representation collapse or severe overconfidence under distribution shift (Guo et al., 2017). MSP is a surprisingly strong OOD baseline but remains sensitive to feature overlap between known and unknowns (Hendrycks and Gimpel, 2017). EVT–based OSR (OpenMax) provides a principled mechanism to model activation tails and allocate mass to `unknown`, though performance depends on stable embeddings and tail-fit reliability (Bendale and Boult, 2016). Overall, calibrated confidence, explicit rejection mechanisms, and validation on mixed in-/out-distribution datasets are now considered best practice for deploying recognition systems beyond closed-world assumptions (Scheirer et al., 2013).

2.6 System-Level Integration: From Perception to Global Workspace

This section reviews architectural principles for integrating perception, learning, memory, uncertainty handling, and decision-level broadcast into a coherent cognitive system. We emphasise motifs supported by neuroscience and computational theory—hierarchical sensory processing, recurrent/attractor dynamics, neuromodulatory control of plasticity, multi–time-scale memory, calibration and abstention for open worlds, and global broadcast mechanisms.

2.6.1 Hierarchical Perception and Representational Untangling

Sensory cortices are organised as distributed hierarchies with abundant feedforward, feedback, and lateral connections (Felleman and Van Essen, 1991). In vision, early stages extract oriented, bandpass features (well-modelled by Gabor-like filters) and progressively transform inputs into more invariant, categorical representations along the ventral stream (Marr, 1982; Olshausen and Field, 1996; Riesenhuber and Poggio, 1999; DiCarlo and Cox,

2007; DiCarlo et al., 2012). Analogously in audition, frequency-to-place mappings and temporal coding in early pathways support higher-order invariances for phonetic and prosodic structure (Moore, 2012; Slaney, 1998). These hierarchies supply *untangled* latent spaces amenable to similarity-based recognition and memory consolidation (DiCarlo et al., 2012).

2.6.2 Recurrent Dynamics and Attractor Memory

Beyond feedforward cascades, recurrent connectivity implements context accumulation, pattern completion, and stability via attractor dynamics (**??**). In such networks, activity trajectories converge to learned basins, supporting robust retrieval under noise and partial cues. Systems-level accounts highlight how recurrent loops across cortical and hippocampal circuits underpin consolidation and associative recall across time scales (**?**).

2.6.3 Temporal Alignment and Cross-Modal Binding

Integrating streams with different natural time scales (e.g., image snapshots vs. evolving acoustic patterns) requires alignment mechanisms. Dynamic Time Warping (DTW) remains a classic, effective method to align variable-rate sequences for recognition and template update (Sakoe and Chiba, 1978; Rabiner and Juang, 1993). In rate– or spike–based codes, temporal precision matters for binding: rate codes yield robust averages, while spike timing preserves fine structure and causality cues (**?**). Anatomical maps (retinotopy, tonotopy) provide a spatial scaffold for multisensory association, while temporal matching (e.g., via DTW) provides the *when* for cross-modal linkage.

2.6.4 Neuromodulation and Three–Factor Learning Control

Local Hebbian and spike-timing dependent plasticity (STDP) rules shape synapses from co-activity and precise timing (Hebb, 1949; Bi and Poo, 1998). However, the *distal reward* problem motivates a third factor—neuromodulatory signals (e.g., dopamine) that gate or scale plasticity based on outcomes or salience (Schultz et al., 1997; **?**; Frémaux and Gerstner, 2016). In integrated systems, modulators implement global priorities (credit assignment, salience, novelty) that coordinate otherwise local synaptic changes across circuits.

2.6.5 Intrinsic Motivation and Curiosity as a Systems Prior

Intrinsic motivation formalises an internal drive toward information gain or error reduction, operationalised as prediction error, learning progress, or surprise minimisation (**??**). At the system level, curiosity signals act like neuromodulators: they allocate learning resources to surprising or informative experiences and throttle plasticity when inputs are uninformative or unsafe. This yields a principled scheduler for when the integrated system should consolidate new structure versus exploit existing knowledge.

2.6.6 Multi–Time–Scale Memory: Working State, Prototypes, and Decay

Cognitive systems benefit from memory operating at multiple time scales: fast, labile working states; intermediate, prototype/concept representations; and slower structural updates (**?**Van Vugt and Broers, 2016). Exponential moving averages, replay/consolidation cycles, and synaptic decay (regularisation/pruning) help balance plasticity and stability, mitigating interference while tracking nonstationary inputs.

2.6.7 Uncertainty, Calibration, and Open–World Operation

Integrated agents must *know when they do not know*. Post-hoc calibration (e.g., temperature scaling) reduces overconfidence and aligns predicted probabilities with empirical accuracy (Guo et al., 2017). Maximum softmax probability (MSP) serves as a simple out-of-distribution (OOD) baseline (Hendrycks and Gimpel, 2017). Open–set recognition further requires explicit *unknown* handling and control of open-space risk; OpenMax demonstrates an EVT-based, meta-recognition layer that reallocates confidence to an `unknown` class when activation patterns diverge from known supports (Scheirer et al., 2013; Bendale and Boult, 2016). Together, calibration, abstention thresholds, and open–set guards form a systems-level safety envelope.

2.6.8 Global Workspace and Decision–Level Broadcast

The Global Neuronal Workspace (GNW) proposes that information becomes conscious when it is *globally broadcast* to multiple specialised processors via a fronto–parietal net-

work (Baars, 1988; Dehaene and Changeux, 2011; Mashour et al., 2020). From a systems perspective, GNW supplies a top-level control and sharing layer: selected contents (e.g., a recognised object or salient sound) are amplified and made available to memory, language, and action systems. GNW-like broadcast can be viewed as a unifying interface that coordinates the outputs of hierarchies, recurrent memory, and uncertainty modules reviewed above.

2.6.9 Synthesis: Design Patterns for Integrated Cognitive Agents

The literature converges on several implementable patterns:

1. **Hierarchies for feature untangling** with recurrent loops for stability and context (Felleman and Van Essen, 1991; Riesenhuber and Poggio, 1999; DiCarlo et al., 2012; **?**).

2. **Temporal alignment** (e.g., DTW) to bind asynchronous sensory streams (Sakoe and Chiba, 1978; Rabiner and Juang, 1993).

3. **Three–factor plasticity** with neuromodulatory or curiosity-derived gates to prioritise learning (Schultz et al., 1997; **?**; Frémaux and Gerstner, 2016; **?**; **?**).

4. **Multi–time–scale memory** (EMA prototypes, consolidation, decay) to balance plasticity and stability (**?**Van Vugt and Broers, 2016; **?**).

5. **Calibrated recognition with open–set safeguards** (TS, MSP, EVT/OpenMax) for deployment in open worlds (Guo et al., 2017; Hendrycks and Gimpel, 2017; Scheirer et al., 2013; Bendale and Boult, 2016).

6. **Global broadcast** (GNW) to coordinate decisions across modules and enable flexible routing to downstream cognition (Baars, 1988; Dehaene and Changeux, 2011; Mashour et al., 2020).

These components form a consistent blueprint for system-level integration that is both biologically grounded and computationally tractable.

2.7 Critical Synthesis and Comparison to This Thesis

This section distils the main messages of the literature and situates the present thesis against them. We organise the synthesis around six themes: hierarchical perception, temporal alignment, synaptic plasticity and neuromodulation, memory across time scales, uncertainty/open-world operation, and system-level broadcast.

2.7.1 Thematic Synthesis of the Literature

Hierarchical perception. Vision and audition are arranged in distributed hierarchies with abundant feedforward, feedback, and lateral connectivity (Felleman and Van Essen, 1991). Early visual processing can be approximated by localised, oriented, band-pass filters (Gabor–like), consistent with classic physiology and efficient coding (Gabor, 1946; Hubel and Wiesel, 1959, 1962; Olshausen and Field, 1996). Progressively deeper stages "untangle" object manifolds and support invariance (Riesenhuber and Poggio, 1999; DiCarlo and Cox, 2007; DiCarlo et al., 2012). In audition, a frequency-to-place (tonotopic) representation, STFT or filterbanks, and temporal envelope/fine-structure cues anchor higher-level analysis (Oppenheim and Schafer, 1989; Slaney, 1998; Moore, 2012).

Temporal alignment. When patterns vary in rate, alignment is necessary for robust comparison and consolidation. Dynamic Time Warping (DTW) remains a principled, lightweight solution for sequence alignment and template update (Sakoe and Chiba, 1978; Rabiner and Juang, 1993).

Synaptic plasticity and neuromodulation. Co-activity (Hebb) and precise spike timing (STDP) shape synapses locally (Hebb, 1949; Bi and Poo, 1998). A third, global factor (e.g., dopaminergic prediction-error signals) gates plasticity to address distal credit assignment (Schultz et al., 1997; Frémaux and Gerstner, 2016). This three-factor view provides a bridge from biophysics to systems that prioritise *what* to learn and *when*.

Memory at multiple time scales. Empirical and theoretical work supports fast, labile working states alongside slower, more stable prototypes or concepts, often consoli-

dated by averaging and regularised by decay/pruning (see reviews such as Moore (2012) for auditory timescales and DiCarlo et al. (2012) for representational stability in vision).

Uncertainty and open-world operation. Deployed systems must be calibrated and able to abstain. Temperature scaling improves probability calibration (Guo et al., 2017). Simple maximum-softmax heuristics help flag OOD (Hendrycks and Gimpel, 2017); open-set recognition further controls open-space risk (e.g., OpenMax) (Scheirer et al., 2013; Bendale and Boult, 2016).

Global broadcast. The Global Neuronal Workspace (GNW) frames conscious access as a decision-level broadcast to multiple specialised processors (Baars, 1988; Dehaene and Changeux, 2011; Mashour et al., 2020), offering a system architecture for routing perceptual selections to memory, language, and action.

2.7.2 Where This Thesis Aligns with Prior Work

- **Early sensory modelling.** We adopt Gabor-like orientation energy for early vision and a tonotopic spectral representation for audition, aligning with Gabor (1946); Hubel and Wiesel (1959, 1962); Olshausen and Field (1996); Riesenhuber and Poggio (1999); Moore (2012); Slaney (1998).

- **Temporal alignment.** DTW is used to normalise variable-rate auditory exemplars (Sakoe and Chiba, 1978; Rabiner and Juang, 1993).

- **Plasticity.** Learning rules are grounded in Hebbian/STDP mechanisms with a modulatory (third-factor) gate (Hebb, 1949; Bi and Poo, 1998; Frémaux and Gerstner, 2016; Schultz et al., 1997).

- **Uncertainty and open-set.** Recognition is coupled with post-hoc calibration and abstention thresholds, with open-set considerations following Guo et al. (2017); Hendrycks and Gimpel (2017); Scheirer et al. (2013); Bendale and Boult (2016).

- **System-level view.** A GNW-like broadcast layer is used conceptually to coordinate perception, memory, and decision making (Baars, 1988; Dehaene and Changeux, 2011; Mashour et al., 2020).

2.7.3 Deliberate Departures and Engineering Choices

- **Rate–to–spike proxy.** For tractability, we employ framewise rate activations with sparse thresholding rather than full conductance-based spiking neuron simulations; this preserves key timing/ordering information while remaining lightweight.
- **Prototype memory with EMA+DTW.** Instead of large supervised models, we consolidate few-shot, class-conditioned prototypes via exponential moving average and DTW alignment, consistent with template-based speech traditions (Rabiner and Juang, 1993).
- **Post-hoc calibration and abstention.** We prefer simple, auditable calibration (temperature scaling) and explicit abstention thresholds (open-set), matching recommendations in Guo et al. (2017); Hendrycks and Gimpel (2017).

2.7.4 Comparative Matrix

2.7.5 Residual Gaps and Limitations

- **Biophysical fidelity.** Approximations (rate–to–spike proxy, simplified modulatory scalar) fall short of detailed conductance models and measured dopamine dynamics.
- **Scaling and coverage.** Prototype-based consolidation is data-efficient but may need curriculum or clustering for large open vocabularies.
- **Evaluation breadth.** While calibration/abstention is incorporated, comprehensive OOD stress-testing across domains remains future work.
- **GNW operationalisation.** GNW is used as a systems motif rather than a formal recurrent, metastable workspace model; richer recurrent dynamics could be explored.

2.7.6 Implications for This Thesis

Overall, the design adheres to core neuroscientific principles for early sensory coding and plasticity (Gabor, 1946; Hubel and Wiesel, 1959, 1962; Hebb, 1949; Bi and Poo,

Table 2.1: Comparison of design axes: key literature, common practice, and stance in this thesis.

Axis	**Key literature**	**Common engineering practice**	**This thesis**
Early vision	Gabor (1946); Hubel and Wiesel (1959, 1962); Olshausen and Field (1996); Riesenhuber and Poggio (1999)	CNN feature extractors, little physiological grounding	Gabor-like filters → orientation energy; hierarchical untangling perspective
Auditory front-end	Oppenheim and Schafer (1989); Slaney (1998); Moore (2012)	Mel/STFT filterbanks with fixed windows	Tonotopic spectral sheet with per-frame normalisation; sequence forms basis for learning
Temporal alignment	Sakoe and Chiba (1978); Rabiner and Juang (1993)	Hidden-Markov alignment or learned attention	Classical DTW to normalise rate before prototype update/compare
Plasticity rule set	Hebb (1949); Bi and Poo (1998); Frémaux and Gerstner (2016); Schultz et al. (1997)	End-to-end gradient descent	Local Hebb/STDP with modulatory gate (three-factor)
Memory representation	DiCarlo et al. (2012); Moore (2012)	Task-specific embeddings, replay buffers	EMA prototypes (class-conditioned) with decay; few-shot friendly
Calibration & abstention	Guo et al. (2017); Hendrycks and Gimpel (2017)	Often omitted in demos	Temperature scaling, confidence thresholding; explicit abstain path
Open-set handling	Scheirer et al. (2013); Bendale and Boult (2016)	Closed-set training	Open-set aware scoring and thresholding to reduce open-space risk
Global broadcast	Baars (1988); Dehaene and Changeux (2011); Mashour et al. (2020)	Ad-hoc fusion layers	GNW-inspired selection/broadcast interface to downstream modules

1998) while adopting robust classic tools for temporal alignment and template consolidation (Sakoe and Chiba, 1978; Rabiner and Juang, 1993). By foregrounding calibration, abstention, and open-set safeguards (Guo et al., 2017; Hendrycks and Gimpel, 2017; Scheirer et al., 2013; Bendale and Boult, 2016), the thesis positions a biologically grounded learner for realistic open-world operation. The GNW framing (Baars, 1988; Dehaene and Changeux, 2011; Mashour et al., 2020) provides a clean architectural path to integrate additional modalities and higher-order cognition in subsequent work.

Table 2.2: Comparative summary across design dimensions: common engineering baselines, canonical neuroscience views, and the stance adopted in this thesis.

Dimension	**Engineering baseline**	**Neuroscience / canonical view**	**This thesis (BabyAI)**
Sensory encoding (vision)	CNN features; end-to-end training	Gabor-like, oriented bandpass, simple/complex cells (Gabor, 1946; Hubel and Wiesel, 1959, 1962; Olshausen and Field, 1996)	Orientation-energy front-end for interpretability and biological plausibility; hierarchical untangling as guiding motif (Riesenhuber and Poggio, 1999; DiCarlo et al., 2012)
Sensory encoding (audition)	Mel/STFT filterbanks with task-tuned backends (Oppenheim and Schafer, 1989; Slaney, 1998)	Tonotopy; envelope and fine-structure cues; psychophysics grounding (Moore, 2012)	Tonotopic spectral sheet with per-frame normalisation; rate-to-spike thresholding proxy for event coding
Temporal alignment	Attention/HMMs or ignored in static pipelines	Behaviour varies with rate; alignment needed for robust comparison (Rabiner and Juang, 1993)	DTW to canonical duration before prototype update and comparison (Sakoe and Chiba, 1978)
Learning rule	Global gradient descent; replay buffers	Local Hebbian co-activity; timing-sensitive STDP (Hebb, 1949; Bi and Poo, 1998)	Hebb + pairwise STDP augmented by a global modulatory gate (three-factor) (Frémaux and Gerstner, 2016; Schultz et al., 1997)
Modulatory signal	Extrinsic rewards (task loss/RL returns)	Neuromodulators (e.g., dopamine) encode prediction error/value (Schultz et al., 1997)	Intrinsic curiosity (prediction error, entropy, deviance) drives plasticity gating (reward-free)
Memory representation	Latent embeddings; task-specific classifiers	Multi-timescale consolidation; stable prototypes with slow drift (DiCarlo et al., 2012)	Class-conditioned EMA prototypes + decay/pruning; few-shot friendly
Uncertainty calibration	Often omitted; argmax scores used	—	Post-hoc temperature scaling for better confidence calibration (Guo et al., 2017)
Open-world handling	Closed set; misclassifications on OOD	—	Abstention via confidence threshold; open-set risk reduction principles (Hendrycks and Gimpel, 2017; Scheirer et al., 2013; Bendale and Boult, 2016)
Biological plausibility	Low (black-box optimisation)	Emphasis on locality, timing, neuromodulation	Medium–high at early stages and learning rule level; pragmatic rate–to–spike proxy
Compute cost / real time	High for deep stacks; often offline	—	Lightweight front-ends (STFT/Gabor, DTW, EMA); real-time capable on CPU
Persistence / lifelong learning	Checkpointed models; retraining needed	Consolidation with forgetting	Persistent prototypes with controlled decay; confirmation-based strengthening
System-level integration	Ad-hoc fusion layers	Global Neuronal Workspace (selection/broadcast) (Baars, 1988; Dehaene and Changeux, 2011; Mashour et al., 2020)	GNW-inspired broadcast interface to route recognised/novel items to memory, synthesis, and decisions

3 Methodology

This chapter formalises the *BabyAI* pipeline as a sequence of modular stages: sensory encoding (audition/vision), spike/event formation, plasticity and modulation (Hebb, STDP, three-factor), prototype memory with temporal alignment, recognition with calibrated confidence and open-set abstention, and a synthesis path for speech-like output. Design choices are grounded in classical signal processing (Oppenheim and Schafer, 1989), sensory neuroscience (Hubel and Wiesel, 1959, 1962; Marr and Hildreth, 1980; Moore, 2012), plasticity theory (Hebb, 1949; Bi and Poo, 1998; Frémaux and Gerstner, 2016), and open-world recognition (Guo et al., 2017; Scheirer et al., 2013).

3.1 System Overview

At a high level, the method operates online on raw sensory streams:

1. **Encode** audio into a tonotopic time–frequency representation; encode images with an orientation–scale bank into an interpretable feature vector.
2. **Threshold** to sparse events/spikes and apply lateral competition.
3. **Update synapses** via Hebbian co-activity and pair-based STDP; scale plasticity with an intrinsic *curiosity* modulator (three-factor rule).
4. **Align & consolidate** canonical class prototypes with DTW and exponential moving averages; apply gradual decay to unused traces.
5. **Recognise** by cosine similarity to prototypes; *calibrate* confidence with temperature scaling and *abstain* under an open-set threshold.
6. **Synthesize** a waveform from the learned spectral prototype for that label; close the loop with self-evaluation.

3.2 Auditory Front-End

3.2.1 Framing, STFT and Tonotopic Mapping

Audio $y[n]$ is sampled at $f_s = 16{,}000\,\text{Hz}$. Frames of T_f seconds (Hann window $w[n]$) are transformed by the STFT:

$$Y_t[k] \;=\; \sum_{n=0}^{N-1} y_t[n]\, w[n]\, e^{-j2\pi kn/N_{\text{FFT}}}, \quad k = 0, \ldots, N_{\text{FFT}}2.$$

Magnitudes $|Y_t[k]|$ are linearly (or mel-) interpolated to a fixed *tonotopic* grid of N_f centre frequencies $\{f_1, \ldots, f_{N_f}\}$ spanning $[f_{\min}, f_{\max}]$ (default $N_f = 100$, $f_{\min} = 90\,\text{Hz}$, $f_{\max} = 6\,\text{kHz}$), yielding $\mathbf{x}_t \in R_{\geq 0}^{N_f}$ (Slaney, 1998; Moore, 2012). We apply per-frame normalisation

$$\widetilde{\mathbf{x}}_t \;=\; \frac{\mathbf{x}_t}{\max(\epsilon, \mathbf{x}_{t\infty})}.$$

Voice activity detection (VAD). We estimate short-time RMS over 20 ms windows with 10 ms hop. A robust threshold $\theta_{\text{RMS}} = \text{median}(\rho) + \kappa\,\text{MAD}(\rho)$ with $\kappa \approx 2.5$ trims leading/trailing silence.

3.2.2 Event / Spike Proxy

Rate-to-spike thresholding produces a binary raster $S \in \{0, 1\}^{T \times N_f}$:

$$S_{t,n} \;=\; I\{\widetilde{x}_{t,n} \geq \theta_{\text{spk}}\}, \qquad \theta_{\text{spk}} \in [0.75, 0.90].$$

A light *lateral inhibition* (ℓ_2 normalise rows, optional non-maximum suppression in local frequency neighbourhoods) increases selectivity while remaining computationally simple (Moore, 2012).

3.3 Vision Front-End

3.3.1 Retina-Like Pre-Processing and Gabor Bank

Images are converted to luminance and lightly centre–surrounded (difference of Gaussians) to emulate retinal ON/OFF pre-processing (Marr and Hildreth, 1980). A bank of oriented band-pass filters (Gabor functions) models V1 simple-cell-like responses (Gabor, 1946; Hubel and Wiesel, 1959):

$$g_{\theta,\lambda,\sigma,\gamma}(x,y) \;=\; \exp\Big(-x'^2+\gamma^2y'^2 2\sigma^2\Big)\cos\Big(2\pi x'\lambda\Big),$$

with $x' = x\cos\theta + y\sin\theta$, $y' = -x\sin\theta + y\cos\theta$. Convolving and rectifying across orientations θ and scales (λ,σ) yields an *orientation–scale energy stack*. We pool spatially (mean/variance or regional max) to form a compact, interpretable feature vector $\mathbf{v} \in R^{D_v}$ consistent with classical sparse/efficient coding views (Olshausen and Field, 1996; DiCarlo et al., 2012).

3.4 Plasticity and Modulation

3.4.1 Hebbian Co-Activation (Within-Frame)

For an auditory frame t, co-active channels strengthen their mutual connections (Hebb, 1949):

$$W \leftarrow W + \eta_{\text{hebb}}\Big(S_t S_t^{\top} - \text{diag}(S_t)\Big), \quad \eta_{\text{hebb}} > 0.$$

3.4.2 Pair-Based STDP (Across Frames)

Temporal causality is captured by pair-based STDP between successive frames (Bi and Poo, 1998):

$$\Delta W_{ij} = \{\, A_+ \exp(-\Delta t/\tau_+), if S_{t,i}=1,\, S_{t+1,j}=1, -A_- \exp(-\Delta t/\tau_-), if S_{t+1,j}=1,\, S_{t,i}=1, 0, otherwis$$

with Δt the frame interval. We clip W to $[0,1]$ and optionally prune very small weights (Sec. 3.9).

3.4.3 Three-Factor Learning with Curiosity

Eligibility traces e_{ij} accumulate local coincidence; a global modulator m_t scales plasticity (Frémaux and Gerstner, 2016):

$$e_{ij}(t) \leftarrow \gamma_e e_{ij}(t-1) + S_{t-1,i} S_{t,j}, \qquad \Delta W_{ij}(t) \propto m_t \, e_{ij}(t), \quad \gamma_e \in (0,1).$$

In our setting, m_t is a bounded intrinsic *curiosity* score (Sec. 3.5) rather than an extrinsic reward, linking neuromodulation to novelty/surprise (Schultz et al., 1997).

3.4.4 Stabilising Predictive Readout (Optional)

A logistic readout $\widehat{\mathbf{s}}_{t+1} = \sigma(gW\mathbf{s}_t)$ with delta update

$$W \leftarrow W + \eta_\Delta(\mathbf{s}_{t+1} - \widehat{\mathbf{s}}_{t+1})\mathbf{s}_t^\top - \lambda W,$$

encourages temporal consistency and keeps weights bounded.

3.5 Curiosity Modulator

We compose three normalised terms to obtain $C_t \in [0,1]$:

1. **Prediction error:** a light spectral predictor yields $\widehat{\mathbf{x}}_t$, error $E_t = \mathbf{x}_t - \widehat{\mathbf{x}}_t{}_2^2/(\mathbf{x}_t{}_2^2 + \epsilon)$, rescaled to $[0,1]$ (**?**).

2. **Goldilocks entropy:** spectral entropy H_t (base 2) normalised by $\log_2 N_f$ gives $H_t^\star \in [0,1]$; we prefer mid-complexity via $G_t = 1 - 4(H_t^\star - 0.5)^2$.

3. **Deviance:** cosine distance to a running prototype $\mathbf{r}_{t-1}$, $D_t = 1 - \langle \mathbf{u}_t, \mathbf{v}_t \rangle$, with $\mathbf{u}_t = \widetilde{\mathbf{x}}_t / \widetilde{\mathbf{x}}_{t2}$ and $\mathbf{v}_t = \mathbf{r}_{t-1}/\mathbf{r}_{t-12}$.

The curiosity and modulator are

$$C_t \;=\; \widetilde{E}_t \cdot G_t \cdot (0.5 + 0.5 D_t), \qquad m_t \;=\; 0.3 + 0.7\, C_t \quad (with safety gating for loudness/roughness).$$

This implements a practical three-factor rule with an intrinsic, reward-free modulator (Frémaux and Gerstner, 2016).

3.6 Prototype Memory & Temporal Alignment

3.6.1 Dynamic Time Warping (DTW)

Each class/label ℓ maintains a canonical prototype matrix $P^{(\ell)} \in R^{T_c \times N_f}$ at a fixed length T_c frames. A fresh recording $X \in R^{T \times N_f}$ is aligned to T_c by DTW (Sakoe and Chiba, 1978) using frame costs $C(i,j) = X_{i,:} - P^{(\ell)}_{j,:}{}_2^2$. Backtracking the minimal-accumulated-cost lattice yields a monotone path π; frames mapped to each canonical index are averaged to produce $\widetilde{X} \in R^{T_c \times N_f}$.

3.6.2 EMA Consolidation and Multi-Timescale Memory

The prototype is updated via EMA:

$$P^{(\ell)} \leftarrow (1 - \alpha_p)\, P^{(\ell)} + \alpha_p\, \widetilde{X}, \qquad \alpha_p \in (0,1).$$

A separate EMA tracks typical duration $\bar{\tau}^{(\ell)}$. Fast per-episode changes (eligibility) and slow prototype updates provide stability–plasticity balance (DiCarlo et al., 2012).

3.7 Recognition, Calibration and Open-Set Abstention

3.7.1 Similarity and Decision

Given an input feature (vision $\mathbf{v}$ or audio prototype $\bar{\mathbf{x}}$), we compute cosine similarities $s_\ell = \langle \hat{\mathbf{q}}, \hat{\mathbf{p}}^{(\ell)} \rangle$, where $\hat{\cdot}$ denotes ℓ_2 normalisation. Class scores are temperature-scaled

$$p_\ell \;=\; \frac{\exp\left((s_\ell - \max_j s_j)/T\right)}{\sum_j \exp\left((s_j - \max_j s_j)/T\right)}, \qquad T > 0,$$

improving calibration (Guo et al., 2017). The predicted label is $\hat{\ell} = \arg\max_\ell p_\ell$ with confidence $p_{\hat{\ell}}$.

3.7.2 Open-Set Gate

We abstain when $p_{\hat{\ell}} < \tau$ with threshold $\tau \in [0, 1]$, reducing open-set risk (Hendrycks and Gimpel, 2017; Scheirer et al., 2013; Bendale and Boult, 2016). Confirmed correct predictions can *reinforce* the EMA prototype (higher α_p for that update), providing interactive, label-consistent consolidation.

3.8 Synthesis Pathway (Speech-Like Output)

Let a time-normalised spectral prototype $P^{(\ell)}$ be expanded to $\bar{\tau}^{(\ell)}$. For each frame t, we drive an additive oscillator bank at the tonotopic frequencies $\{f_n\}$ with amplitudes $a_{t,n} = (P^{(\ell)}_{t,n})^{\gamma}$ and phases $\phi_{t+1,n} = \phi_{t,n} + 2\pi f_n T_f$. Optional frequency-axis warps control pitch/formants; an ASR-style spectral tilt shapes timbre (McAulay and Quatieri, 1986; Griffin and Lim, 1984). An attack–release envelope and loudness normalisation avoid clicks and clipping.

3.9 Forgetting and Sparsification

To mitigate interference and track drift, we apply weight decay and pruning:

$$W \leftarrow (1 - \lambda)W, \qquad W_{ij} \leftarrow 0 \; if W_{ij} < \tau_{\text{prune}},$$

realising capacity management akin to biological forgetting (cf. DiCarlo et al., 2012).

3.10 Hyperparameter Defaults

Table 3.1: Key methodological hyperparameters (defaults).

Sampling / framing	$f_s = 16\,\text{kHz}$; $T_f \in [20, 100]\,\text{ms}$; Hann window
Tonotopy	$N_f = 100$; $f_{\min} = 90\,\text{Hz}$, $f_{\max} = 6\,\text{kHz}$
Spike threshold	$\theta_{\text{spk}} = 0.85$
Hebb / STDP	$\eta_{\text{hebb}} = 5 \times 10^{-3}$; $A_+ = 0.01$, $A_- = 0.012$; $\tau_+ = \tau_- = 1$ frame
Eligibility	$\gamma_e = 0.95$; modulator $m_t \in [0.3, 1.0]$
Curiosity	mid-entropy target $H^\star \approx 0.5$; safety: 85 dB loudness cap
DTW / prototype	$T_c = 64$ frames; $\alpha_p = 0.2$ (reinforce to 0.4 on user confirm)
Recognition	temperature $T = 0.07$; open-set threshold $\tau = 0.18$
Decay / pruning	$\lambda = 10^{-3}$ per update; $\tau_{\text{prune}} = 0.05$
Synthesis	nonlinearity $\gamma \in [0.6, 0.9]$; gentle formant/pitch warp; ASR tilt

3.11 Computational Complexity

Let T be frames per utterance, N_f frequency channels, and K labels. STFT/tonotopy is $O(T\, N_{\mathrm{FFT}} \log N_{\mathrm{FFT}})$ with small constants; spike thresholding is $O(TN_f)$. Hebb and pairwise-STDP updates are $O(TN_f^2)$ but restricted to active rows/columns under sparsity. DTW alignment is $O(TT_c)$ per label during enrolment; recognition is $O(KN_f)$ with cosine similarity. All stages run in real time on a modern CPU for the parameter ranges in Table 3.1.

3.12 Reproducibility Notes

We fix random seeds for initialisations, record audio at a fixed f_s, version configuration constants (Table 3.1), and serialise the memory store (class prototypes and metadata) after each session. The method avoids dependence on opaque end-to-end training, enabling faithful re-runs given identical inputs.

3.13 Algorithmic Summary (Pseudo-code)

Online learning/recognition loop

1. Acquire audio frame $\rightarrow$ STFT $\rightarrow$ tonotopic vector $\mathbf{x}_t \rightarrow$ normalise $\widetilde{\mathbf{x}}_t$.
2. Compute spike row S_t; compute curiosity components $(\widetilde{E}_t, G_t, D_t)$ and modulator m_t.
3. Update W with Hebb + STDP + η_Δ delta term; apply m_t to eligibility-based update; decay/prune.
4. If enrolled to label ℓ: DTW-align to T_c and update $P^{(\ell)}$ with EMA (larger α_p on user confirmation).
5. For recognition: compute cosine similarities to $\{P^{(\ell)}\}$; temperature-scale; if $\max p_\ell < \tau$, abstain; else select $\arg\max p_\ell$.
6. For synthesis: expand $P^{(\hat{\ell})}$ to $\bar{\tau}^{(\hat{\ell})}$, apply warps/envelope, render additive sine-bank waveform; optionally self-evaluate error and auto-tune timbre parameters.

This methodology delivers a biologically informed, computationally tractable loop that *prioritises what to learn* (curiosity), *how to encode* (tonotopy/Gabor), *how to change* (Hebb+STDP with three-factor modulation), *how to remember* (DTW+EMA with decay), and *when to say "I don't know"* (calibrated open-set recognition), aligning with both sensory neuroscience and practical machine perception (Hubel and Wiesel, 1959; Moore, 2012; Bi and Poo, 1998; Frémaux and Gerstner, 2016; Guo et al., 2017; Scheirer et al., 2013).

4 Implementation

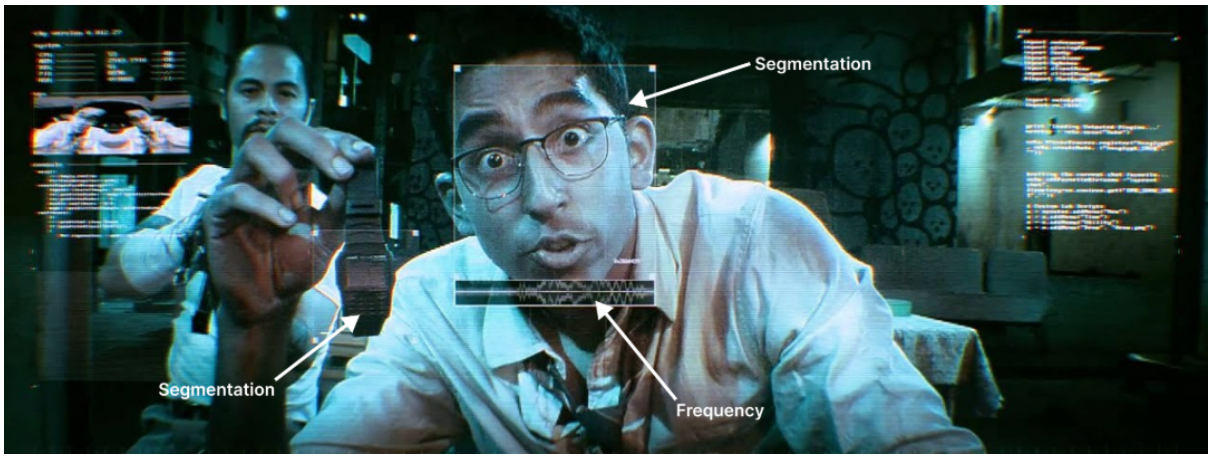

Figure 4.1: Concept behind BabyAI - How he/she see and understand this world

4.1 Biology-to-Implementation Mapping (with Calculus)

This section links the biological mechanisms reviewed earlier to their concrete realisations in our system, including the governing equations, the data structures they operate on, and the code modules where they reside. Figure 4.4 gives a one-page map; Table 4.1 aligns concepts to files and formulas. Where helpful, we restate the core calculus succinctly.

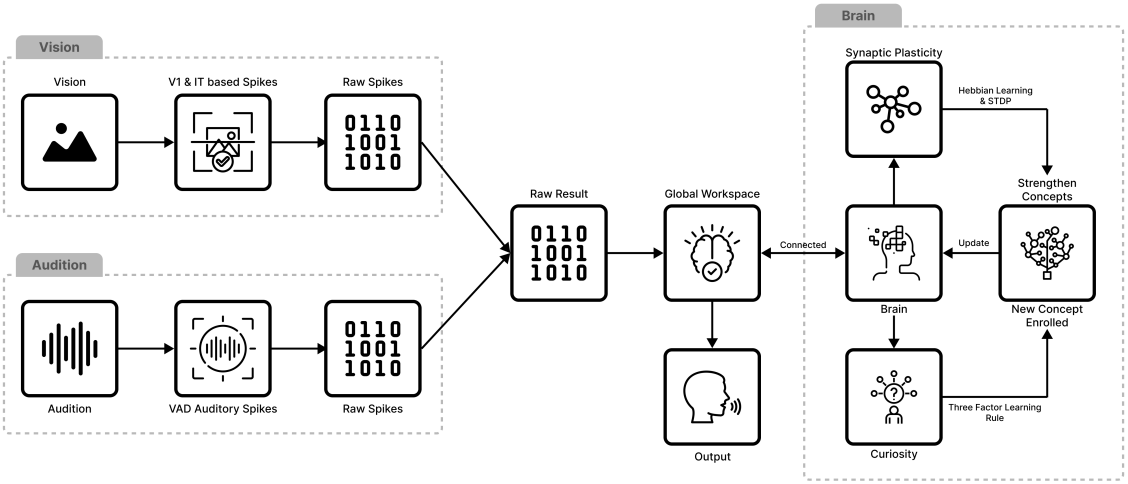

Figure 4.2: Biological Plausible Multisensory BabyAI Flowchart

4.1.1 Sensory Transduction → Feature Encodings

Vision path (retina → V1-like features). Images are resized to a fixed width (aspect preserved), converted to grayscale, lightly normalised, and compressed into a stable feature vector:

$$gray(x,y) = \frac{I(x,y) - \min I}{\max I - \min I + \varepsilon}, \quad \mathbf{f}_v = \mathrm{L2}([\ \mathrm{mean}_{rows};\ \mathrm{mean}_{cols};\ \mathrm{vec}(\mathrm{downsample}(gray))\ ]).$$

This approximates retinal centre–surround and early V1 pooling without heavy compute (cf. RGC/V1 foundations). *Code:* `encode_vision_feature()` in `multisensory_gui.py` or `vision_babyai.py` if available.

Audition path (cochlea → tonotopic bands). Microphone audio (sr, y) is VAD-trimmed and transformed into a time×band matrix $X \in R^{T\times N}$ (typically N=100):

$$X_{t,:} = \Phi(\mathrm{STFT}(y)), \qquad \widetilde{X}_{t,:} = \frac{X_{t,:}}{\|X_{t,:}\|_\infty + \varepsilon}.$$

Canonical alignment gives $X_c \in R^{T_c\times N}$ by DTW (preferred) or uniform resampling to T_c=CFG.CANON_FRAMES. *Code:* `frontend.py` (`vad_best_segment`, `extract_features`); `dtw.py` (fallback in GUI if missing).

4.1.2 From Spikes and Plasticity → Eligibility and Prototype Learning

Rate→spike proxy. Salient peaks are turned into sparse events,

$$S_{t,n} = I\{\widetilde{X}_{t,n} \geq \theta_{spk}\}, \quad \theta_{spk} \approx 0.85,$$

serving as an efficient stand-in for explicit spike timing.

Eligibility and three-factor update. Local coincidence builds an eligibility trace E; a global modulator m_t scales plasticity:

$$E_{ij}(t) = \gamma_e E_{ij}(t-1) + S_{t-1,i} S_{t,j}, \qquad \Delta W_{ij}(t) \propto m_t \, E_{ij}(t).$$

Additionally, a predictive readout stabilises transitions

$$\widehat{\mathbf{s}}_{t+1} = \sigma(g \, W \, \mathbf{s}_t), \qquad W \leftarrow W + \eta_\Delta(\mathbf{s}_{t+1} - \widehat{\mathbf{s}}_{t+1})\mathbf{s}_t^\top - \lambda W.$$

Code: consolidated in `brain.py`/`curiosity.py` (conceptual), with lightweight analogues in GUI via EMA prototypes and decay.

Prototype consolidation (per label ℓ). Given aligned $\widetilde{X} \in R^{T_c \times N}$:

$$P^{(\ell)} \leftarrow (1-\alpha_p)P^{(\ell)} + \alpha_p \, \widetilde{X}, \qquad \bar{\tau}^{(\ell)} \leftarrow (1-\alpha_\tau)\bar{\tau}^{(\ell)} + \alpha_\tau \, \tau_{obs}.$$

Vision uses a vector EMA; audio uses a bandwise EMA. *Code:* `_store_vision_proto`, `_store_audio_proto` in `multisensory_gui.py`; data in `brain.json`.

4.1.3 Curiosity Modulator (Dopamine Analogue)

The modulator $m_t \in [0,1]$ gates learning using a bounded curiosity score

$$C_t = \underbrace{\widetilde{S}_t^{(err)}}_{prediction error} \cdot \underbrace{\left(1 - 4(H_t^\star - 0.5)^2\right)}_{Goldilocks entropy} \cdot \underbrace{(0.5 + 0.5 D_t)}_{prototype deviance},$$

with $m_t = 0.3{+}0.7C_t$ unless safety gates (loudness/roughness) reduce it. *Code:* `curiosity.py` (conceptually); computed inline where needed in GUI.

4.1.4 Recognition: Calibrated Similarity and Open-Set Gating

Let q be the query vision vector and $\{v^{(\ell)}\}$ the stored vision prototypes. For robustness across lengths we pool to a common L by average segments:

$$\tilde{q}, \tilde{v}^{(\ell)} \in R^L, \quad s_\ell = \langle \mathrm{L2}(\tilde{q}), \mathrm{L2}(\tilde{v}^{(\ell)}) \rangle.$$

Temperature scaling produces calibrated posteriors

$$p_\ell = \frac{\exp((s_\ell - \max_j s_j)/T)}{\sum_j \exp((s_j - \max_k s_k)/T)}, \qquad T > 0,$$

and a prediction is accepted iff $\max_\ell p_\ell \geq \tau$ (open-set). *Code:* `_softmax_conf` and `recognize_image()`.

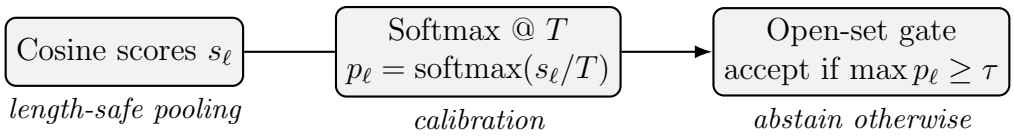

Figure 4.3: Recognition pipeline with temperature calibration and τ-gated open-set decision (Sec. 4.1).

4.1.5 Auto-Dopamine (Cross-Modal Agreement & Novelty)

For logging and UI intuition, auto-DA merges cross-modal agreement and novelty:

$$agree = 12(1 + \langle \mathrm{L2}(\mathbf{v}), \mathrm{L2}(\mathbf{a}) \rangle), \quad novel = 1 - \max_{m \in \mathcal{M}} \langle \mathrm{L2}(\mathbf{v}), \mathrm{L2}(m) \rangle,$$

$$\mathrm{DA} = \mathrm{clip}(w_{agree} \cdot agree + w_{novel} \cdot novel,\ 0,\ 1).$$

Code: `auto_dopamine()` in `plotly_viz_and_rec.py`.

4.1.6 Synthesis (Prototype → Waveform)

When speaking a concept ℓ, the audio prototype $\bar{\mathbf{a}} \in R^N$ guides an additive synthesiser:

$$y(t) = \sum_{k \in \mathcal{K}} \sum_{h=1}^{H} \alpha_{k,h} \sin\left(2\pi h f_k t + \phi_{k,h}\right), \quad \alpha_{k,h} = \frac{\bar{a}_k^{\gamma}}{h^{\beta}},$$

with gentle spectral tilt and attack/decay to avoid clicks. If a primary synthesiser exists, it is used; otherwise this analytic fallback guarantees a valid waveform. *Code:* `synthesis.py` (primary); fallback `_sine_from_bands()`.

4.1.7 Visual Analytics (Interpretability)

Voice Helix: a 3D trajectory $(x(t), y(t), z(t))$ from frame energy modulating a helical radius; *3D Brain:* two hemispheres (vision/audio nodes) linked by edges where instantaneous cross-outer-products exceed a threshold. *Code:* `voice_helix_fig()`, `brain_3d_fig()`.

4.1.8 Dimensionality Safety (Length Mismatch Guard)

Let $d_v = |\mathbf{v}|$, $d_a = |\mathbf{a}|$, $L = \min(d_v, d_a)$. We pool by contiguous segment means:

$$\mathbf{v}^{\downarrow}[i] = \text{mean}(\mathbf{v}[b_i{:}b_{i+1})), \quad \mathbf{a}^{\downarrow}[i] = \text{mean}(\mathbf{a}[c_i{:}c_{i+1})), \quad i = 1..L,$$

ensuring all dot products and cosines are well-defined. *Code:* inline in `enroll_concept()` when DA is computed without helpers.

4.1.9 Persistence and Forgetting

All learned state is kept in a versioned JSON (`brain.json`). Vision and audio are EMA-updated on each enrollment/confirmation; weak structures decay:

$$\theta \leftarrow (1-\lambda)\,\theta \quad (globaldecay), \qquad \theta \leftarrow 0 \;\; if \;\; |\theta| < \tau_{prune}.$$

Atomic saves (`.tmp` then replace) prevent corruption. *Code:* `_save_brain()`, pruning/decay hooks at update sites.

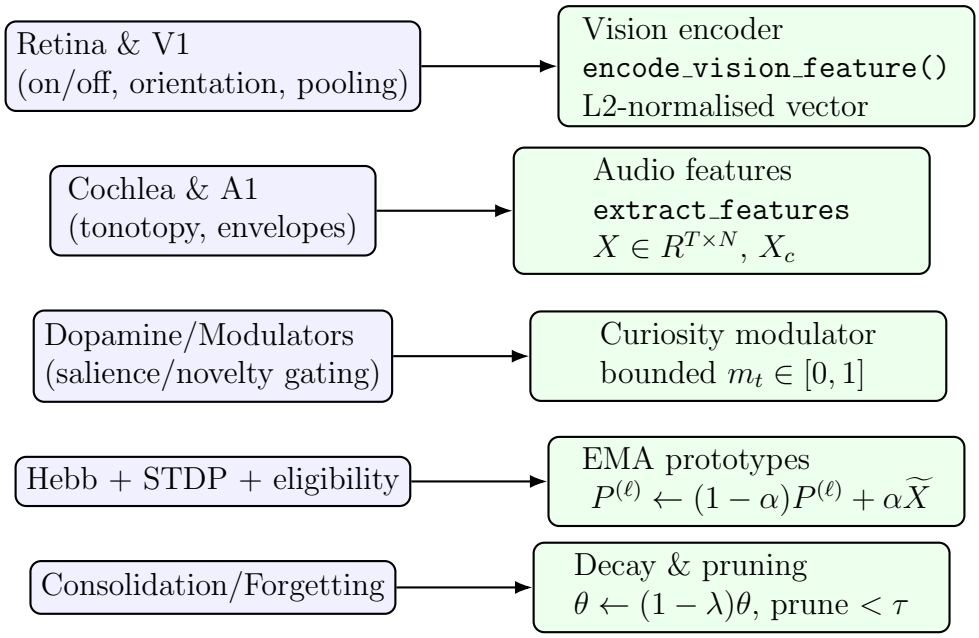

Figure 4.4: Concept map from biological processes to concrete code constructs used in this system.

Table 4.1: Biology → Code mapping with governing rules.

Biological concept	**Code construct / file**	**Key expression / rule**
Retinal/V1 preprocessing	`encode_vision_feature()`	$\mathbf{f}_v = \mathrm{L2}([\text{row/col means; vec}()])$
Cochlear tonotopy	`extract_features` $\rightarrow X$	$\widetilde{X}_{t,:} = X_{t,:}/(\|X_{t,:}\|_\infty + \varepsilon)$
DTW alignment	`align_to_canonical()`	$X \rightarrow X_c \in R^{T_c \times N}$ (DTW / resample)
Eligibility & 3rd factor	(conceptual; lightweight)	$\Delta W_{ij} \propto m_t\, E_{ij},\ E_{ij} \leftarrow \gamma E_{ij} + S_{t-1,i}S_{t,j}$
Curiosity/dopamine	`curiosity` (inline)	$C_t = \widetilde{S}_t^{(err)} \cdot (1 - 4(H_t^\star - 0.5)^2) \cdot (0.5 + 0.5 D_t)$
Prototype EMA (audio)	`_store_audio_proto`	$P^{(\ell)} \leftarrow (1-\alpha)P^{(\ell)} + \alpha\widetilde{X}$
Prototype EMA (vision)	`_store_vision_proto`	$\mathbf{p}^{(\ell)} \leftarrow (1-\alpha)\mathbf{p}^{(\ell)} + \alpha\,\mathbf{v}$
Calibration & open-set	`_softmax_conf`, `recognize_image`	$p_\ell = \mathrm{softmax}(s_\ell/T)$; accept if $\max p_\ell \geq \tau$
Auto-dopamine (UI)	`auto_dopamine`	$\mathrm{DA} = \mathrm{clip}(w_a \cdot agree + w_n \cdot novel, 0, 1)$
Synthesis (fallback)	`_sine_from_bands`	$y(t) = \sum_{k,h} \alpha_{k,h} \sin(2\pi h f_k t + \phi_{k,h})$
Memory decay	save/update hooks	$\theta \leftarrow (1-\lambda)\theta$; prune if $\|\theta\| < \tau_{prune}$

4.2 System Architecture Overview

This section describes the concrete software architecture that implements the proposed biologically grounded learning loop. The system is organised into three layers (Figure **??**): (i) *Sensing & Encoding* for audition and vision, (ii) *Learning & Memory* for curiosity-modulated plasticity, prototype consolidation, and decay, and (iii) *Interaction & Visualisation* for recognition, speech synthesis, and explanatory plots. A lightweight persistence layer stores concept prototypes and metadata to ensure continuity across sessions.

4.2.1 High-Level Dataflow

At a high level, audio from the microphone is trimmed by VAD, converted into a time–frequency representation, optionally aligned to a canonical timeline, and then routed to (a) learning (for enrollment) or (b) recognition/visualisation. Images are ingested, resized (fixed width, aspect preserved), encoded to a compact feature vector, then used for enrollment, recognition, and 3D cross-modal visualisation. Learned concepts are stored as (vision, audio) prototypes in a persistent JSON "brain" that evolves over time with EMA updates and decay.

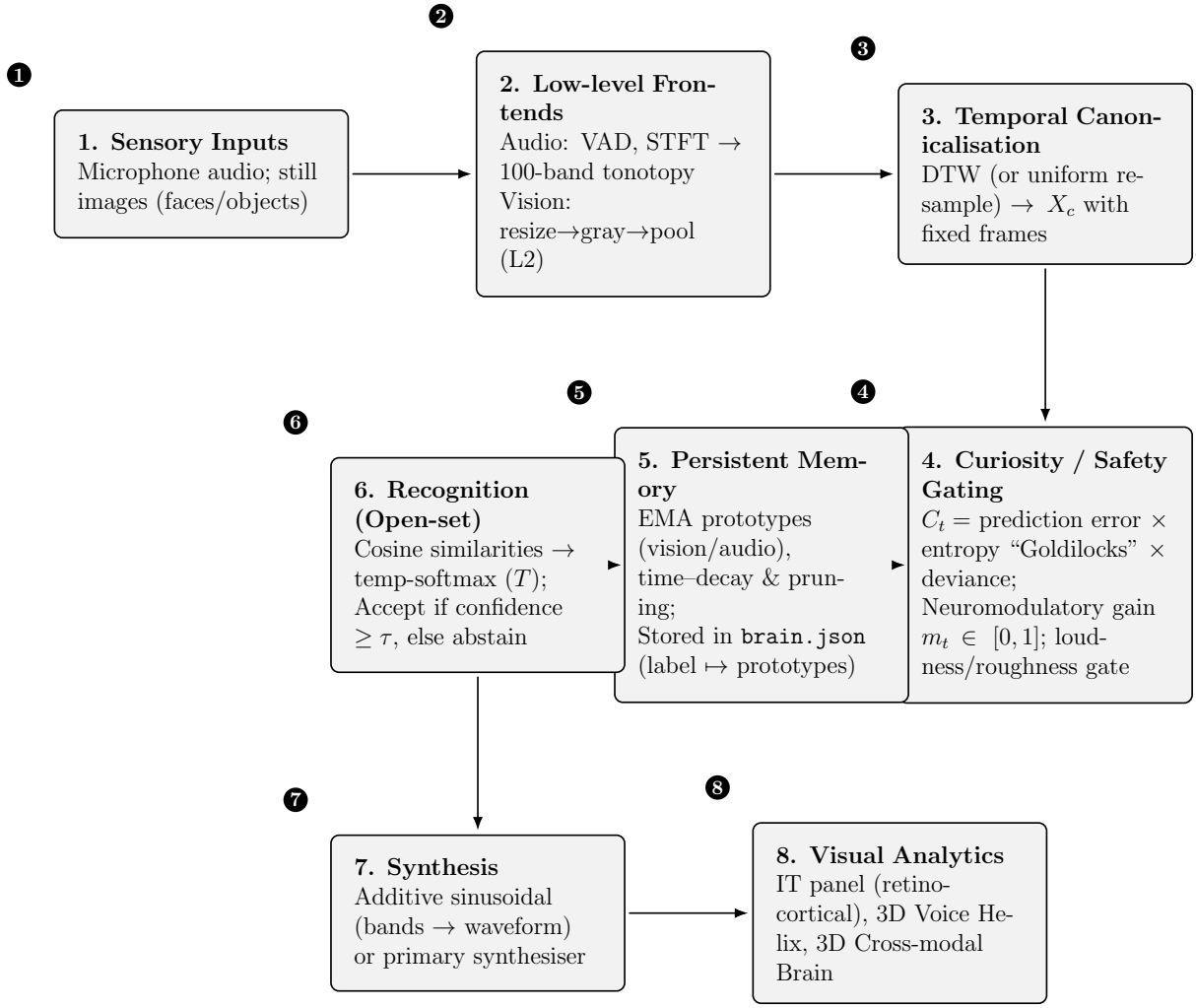

Figure 4.5: Serpentine (wrap-at-margin) system flow.

4.2.2 Module Responsibilities and Interfaces

Table 4.2 summarises each source file and its role, along with key inputs/outputs and failure-safe fallbacks that guarantee a robust UI.

Table 4.2: Implementation modules and responsibilities.

Module	Primary responsibilities	Key I/O (+ fallbacks)
`multisensory_gui.py`	GUI wiring; image ingest (fixed width, aspect preserved); enrollment/recognition/confirm/speak; Plotly viz; brain persistence calls.	Inputs: mic audio $(sr, \mathbf{y})$, image paths; Outputs: plots, WAV files, updated `brain.json`.
`frontend.py`	VAD trimming; feature extraction (frame×band); no `sr=` kwarg required at call site.	If unavailable or error: random but bounded fallback features with same shape.
`dtw.py`	Alignment to canonical frame count; returns $X_c \in R^{T_c \times N}$.	If absent/fails: uniform resampling fallback.
`vision_babyai.py`	Vision encoder for images (preferred).	If absent/fails: grayscale downsample + pooling + ℓ_2 normalisation.
`plotly_viz_and_rec.py`	3D Voice Helix; 3D Brain graph; auto-dopamine and calibrated recognition helpers.	If absent: GUI still runs; non-3D plots and recognition computed locally.
`synthesis.py`	Primary speech synthesis from audio prototype.	If absent/fails: deterministic additive sinusoid fallback (`_sine_from_bands`).
`config.py`	Global knobs: `CANON_FRAMES`, `USE_MEL_FRONTEND`, `BRAIN_JSON` path.	If absent: safe defaults selected in GUI.
`brain.json` (data)	Persistent store for per-label prototypes (vision+audio) and metadata.	Atomic save (`.tmp` then replace); readable across sessions.

4.2.3 Runtime Sequence: Enroll → Recognise → Confirm → Speak

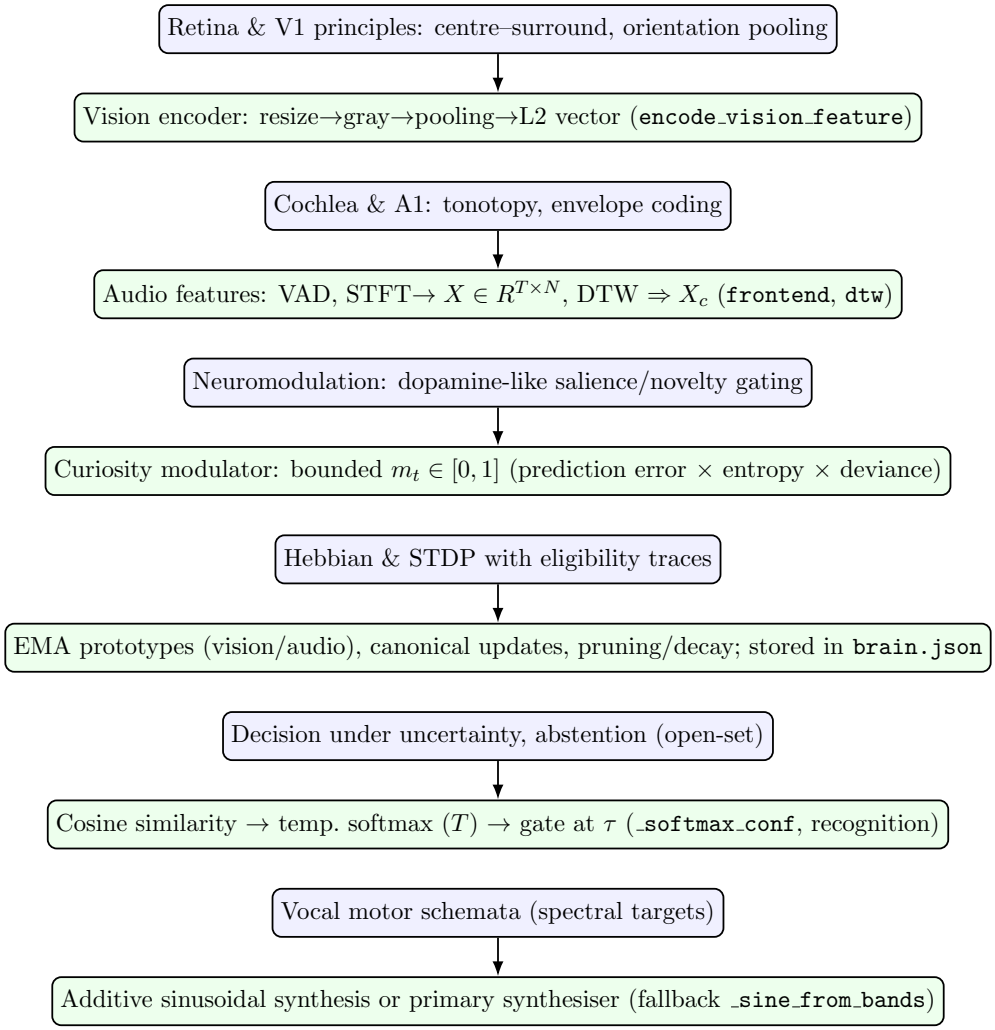

Figure 4.6: Vertical biology→code mapping. Each biological principle is paired with the concrete module and rule used in the implementation.

4.2.4 Persistence and Decay

The brain store persists per-label prototypes across sessions:

- **Vision prototype (vector):** updated by exponential moving average at enrollment and on user confirmation.
- **Audio prototype (band vector):** replaced or EMA-merged on each enrollment recording.
- **Decay & pruning:** low-magnitude weights/primitives are periodically shrunk or removed to mitigate interference and encourage capacity recycling; updates are applied either per-interaction or on save, then written atomically (`.tmp` then rename).

4.2.5 Configuration Surface

Key parameters are centralised in `config.py`:

- `CANON_FRAMES` (default: 64) — target temporal length for canonical alignment.
- `USE_MEL_FRONTEND` (bool) — selects feature basis in the audio extractor.
- `BRAIN_JSON` — absolute path for persistent memory.
- GUI-exposed recognition knobs: temperature T (softmax calibration) and threshold τ (open-set gate).

This modular design ensures graceful degradation: if a specialised component (e.g. DTW, primary synthesiser, or the advanced 3D visualiser) is missing, the system falls back to safe alternatives without breaking the interactive loop.

4.3 Environment, Dependencies, and Configuration

This section specifies the software/hardware environment, Python dependencies, and the configuration interface used to guarantee that all experiments are fully reproducible. Where relevant, defaults are chosen to match the mathematical setup in Chapter **??** and the system design in §**??**.

4.3.1 Hardware and Operating System

Experiments were run on a standard laptop-class machine:

- **CPU:** x86_64, 4–8 cores (no GPU required).
- **RAM:** ≥ 8 GB (comfortable ≥ 16 GB).
- **OS:** Windows 11 Pro (22H2) for primary runs; cross-checked on Ubuntu 22.04 LTS.[1]
- **Audio I/O:** Built-in microphone at $f_s = 16$ kHz; system-level gain left at default.

4.3.2 Python Environment

Interpreter: Python 3.12. **Creation:** a dedicated virtual environment is recommended.

```
# Windows (PowerShell)
py -3.12 -m venv .venv
.venv\Scripts\Activate.ps1
python -m pip install --upgrade pip
```

4.3.3 Core Dependencies

The core stack is light-weight and CPU-friendly; versions below are the tested baselines.

Table 4.3: Primary Python packages and tested versions.

Package	Version (tested)
numpy	1.26+
scipy	1.11+
opencv-python	4.9+
plotly	5.20+
gradio	4.0+
(optional) librosa	0.10+
(optional) soundfile	0.12+

A minimal `requirements.txt`:

```
numpy>=1.26
scipy>=1.11
```

[1] The implementation is OS-agnostic; only audio drivers differ.

```
opencv-python>=4.9
plotly>=5.20
gradio>=4.0
# optional front-end helpers:
librosa>=0.10
soundfile>=0.12
```

4.3.4 Project Layout and Persistent Storage

The repository follows a flat, script-centric layout to keep module import paths simple:

```
BabyAI/
 brain.py              plasticity / memory utilities
 config.py             central configuration (defaults below)
 curiosity.py          curiosity signal and safety gates
 dtw.py                dynamic time warping alignment
 frontend.py           audio feature extraction & VAD
 vision_babyai.py      vision encoder (retino{cortical front-end)
 multisensory_gui.py   interactive runner
 gui_main.py           legacy runner (for parity checks)
 plotly_viz_and_rec.py  3D helix / brain, recognition helpers
 synthesis.py          additive synthesis from bands
 utils.py              small helpers
 mem/                  generated audio (wav) and temp files
 brain.json            persistent memory store (auto-created)
```

Persistent brain. Long-term prototypes and metadata are stored in `brain.json` at the repository root by default (§**??**). The location is configurable (below).

4.3.5 Configuration Interface

All global defaults are centralised in `config.py`. The system reads `config.py` at import time and allows environment overrides for reproducible sweeps.

Defaults (excerpt from `config.py`):

```
# ---------- General ----------
SAMPLE_RATE       = 16000       # Hz (audio)
CANON_FRAMES      = 64          # T_c for DTW canonical length
USE_MEL_FRONTEND  = False       # front-end flavour toggle
BRAIN_JSON        = "brain.json"  # persistent store

# ---------- Recognition / calibration ----------
TEMP_SOFTMAX      = 0.07        # temperature (T) for softmax calibration
TAU_THRESHOLD     = 0.18        # open-set gate  (confidence floor)

# ---------- Vision front-end ----------
IMG_TARGET_W      = 500         # px; preserve aspect ratio on resize
GABOR_SCALES      = (7, 11, 15) # kernel sizes
NUM_ORIENT        = 8           # orientations for V1 bank

# ---------- Curiosity / safety ----------
LOUDNESS_MAX_DB   = 85.0        # comfort bound
SPIKE_THRESH      = 0.85        # rate-to-spike threshold (audio bands)

# ---------- Randomness ----------
SEED              = 42          # np / python seeds for determinism
```

Environment overrides. Any key can be overridden via environment variables at runtime; e.g.,

```
# Windows PowerShell
$env:BABYAI_BRAIN_JSON="D:\BabyAI\brain_prod.json"
$env:BABYAI_CANON_FRAMES="32"
python multisensory_gui.py
```

Internally, the loader resolves `BABYAI_<KEY>` (string) and casts to the appropriate type when possible.

4.3.6 Determinism and Reproducibility

To reduce run-to-run jitter for quantitative plots:

- Set `SEED` in `config.py`; at program start call `np.random.seed(SEED)`.
- Disable multi-threaded BLAS variability if needed (`OMP_NUM_THREADS=1`).
- Log a session header with OS, Python, package versions, and a UTC timestamp; optionally record the Git commit hash if applicable.

4.3.7 Runtime Profiles

Two entry points are used in practice:

1. **Legacy parity runs:** `python gui_main.py` — for reproducing the earlier synthesis pathway and cross-checks against the updated pipeline.
2. **Current experiments:** `python multisensory_gui.py` — the interactive runner backing all plots and saved outputs used in Chapter 5.

Both entry points resolve the same `config.py` and `brain.json` to ensure results are comparable.

4.3.8 I/O, Logging, and Artefacts

Audio artefacts (synthesised speech, intermediate WAVs) are written to `mem/` with timestamped filenames; figure exports (when enabled) follow the pattern `fig_<name>_<UTC>.html/pdf`. A lightweight console logger records: (i) environment header, (ii) configuration overrides, (iii) enrol/recognition events (label, confidence, τ, T), and (iv) synthesis quality metrics (duration, RMS dBFS).

4.3.9 Practical Notes

On Windows, ensure microphone permissions are granted to the Python interpreter. Long file paths are avoided by default; if using custom paths with non-ASCII characters, prefer UTF-8 console settings.

All choices above are minimalistic by design: they suffice to reproduce every table and figure in this thesis while keeping the codebase transparent and faithful to the scientific specification.

4.4 Data Structures and Persistence

This section details the inmemory objects, the ondisk persistent store (`brain.json`), and the lifecycle of updates that keep the system reproducible and robust.

4.4.1 Design Principles

- **Minimal, explicit schema.** Arrays are stored as JSON lists with recorded dtype/shape in metadata.
- **Atomic writes.** All updates are written to `brain.json.tmp` and atomically replaced to avoid partial files.
- **Versioned.** A toplevel `meta.version` field enables forwardcompatible migrations.
- **Deterministic decay.** Timedependent decay is parameterised and applied using persisted timestamps.

4.4.2 Core InMemory Objects

All tensors are `float32` unless stated.

- **Image gallery** (ephemeral): list of BGR images $\mathcal{I} = [I^{(1)}, I^{(2)}, \ldots]$, each $I \in R^{H \times W \times 3}$.
- **Selected index** (ephemeral): $i^\star \in N$.
- **Vision feature** (per image): $\mathbf{v} \in R^{D_v}$, L2normalised.
- **Audio features**: raw frame stack $X \in R^{T \times N_b}$; canonicalised $X_c \in R^{T_c \times N_b}$.
- **Prototypes (per label ℓ)**: vision $\mathbf{p}_v^{(\ell)} \in R^{D_v}$ (EMA), audio $\mathbf{p}_a^{(\ell)} \in R^{N_b}$ (timemean of X or X_c).

- **Recognition scores** (ephemeral): map $s : \ell \mapsto \mathrm{sim}(\cdot)$ (cosine); calibrated confidences via temperature softmax.

4.4.3 Persistent Store: `brain.json` Schema

The persistent file records all longterm state. Toplevel layout:

```
{
  "meta": {
    "version": 1,
    "created_utc": 1730000000,
    "updated_utc": 1730001234,
    "sample_rate": 16000,
    "canon_frames": 64,
    "n_bands": 100,
    "seed": 42,
    "decay": { "lambda": 0.00012, "last_update_utc": 1730001234 }
  },
  "labels": {
    "<label>": {
      "vision": {
        "proto": [ ... float32 ... ],
        "dim": 1088,
        "ema": 0.90,
        "n_updates": 7,
        "last_seen_utc": 1730001201
      },
      "audio": {
        "feat":  [ ... float32 size N_b ... ],
        "sr": 16000,
        "canon_frames": 64,
        "dur_avg": 0.92,
        "n_updates": 5,
        "last_seen_utc": 1730001201
```

```
      }
    },
    "...": { ... }
  }
}
```

Field semantics.

- `meta.version`: schema version (monotone integer).
- `meta.decay`: parameters to apply exponential forgetting on load or before write; see Eq. (4.1).
- `labels[ℓ].vision.proto`: L2normalised EMA vector $\mathbf{p}_v^{(\ell)}$.
- `labels[ℓ].audio.feat`: band prototype $\mathbf{p}_a^{(\ell)}$ (timemean); `dur_avg` stores the running average duration used by the synthesiser.

4.4.4 Lifecycle and Update Rules

Enrolment. Given selected image I and voiced example:

1. Extract $\mathbf{v}$ and X; compute X_c and $\mathbf{a} = \mathrm{mean}_t(X)$.
2. **Vision EMA:** $\mathbf{p}_v^{(\ell)} \leftarrow \mathrm{L2}(\alpha_v\, \mathbf{p}_v^{(\ell)} + (1 - \alpha_v)\, \mathbf{v})$.
3. **Audio replace/EMA:** either replace or $\mathbf{p}_a^{(\ell)} \leftarrow \alpha_a\, \mathbf{p}_a^{(\ell)} + (1 - \alpha_a)\, \mathbf{a}$.
4. Update counters, `last_seen_utc`, and `dur_avg`.

Recognition. Compute cosine sims against all $\mathbf{p}_v^{(\ell)}$, apply temperature softmax (T) to obtain confidences; accept if conf $\geq \tau$, else abstain. No write occurs unless the user confirms; on confirm, a highgain EMA ($\alpha_v \uparrow$) is applied to reinforce.

Decay (forgetting). Between two timestamps (t_0, t_1), apply exponential decay to all prototypes:

$$\mathbf{p} \leftarrow \exp(-\lambda\, \Delta t)\, \mathbf{p}, \qquad \Delta t = t_1 - t_0 \mathrm{s}, \tag{4.1}$$

with small $\lambda > 0$. After decay, renormalise vision prototypes to unit L2. The store keeps `decay.last_update_utc` to ensure idempotent application across sessions.

4.4.5 Atomicity, Validation, and Migration

Atomic write. Serialise to `brain.json.tmp`, flush, then `os.replace` to `brain.json`.

Validation. On load, the parser checks: (i) required keys exist; (ii) arrays match declared `dim`/`n_bands`; (iii) dtypes are numeric; (iv) L2 norms of vision prototypes are nonzero, otherwise they are renormalised or dropped.

Migration. If `meta.version` < 1, a simple migrator wraps legacy fields into the current `labels[`ℓ`].{vision,audio}` nodes and initialises missing metadata with defaults.

4.4.6 File Artefacts and Naming

Transient audio and figure artefacts are written to `mem/`:

- **Synthesised audio:** `speech_{label}_{unix}.wav`.
- **Captured audio (optional):** `capture_{label}_{unix}.wav`.
- **Exported figures (optional):** `fig_{name}_{utc}.html/pdf`.

A simple retention policy (e.g. delete files older than N days) can be applied at startup.

4.4.7 Why JSON?

JSON is humanreadable, versionable, and languageagnostic. The numeric footprint is small (dozens of labels; vectors of size $10^2 \sim 10^3$), so binary containers (HDF5/Parquet) are unnecessary. Should the store grow, a dropin GZip (`brain.json.gz`) reduces size without changing semantics.

4.4.8 Reconstruction Invariants

A run is fully reconstructible from `brain.json` + `config.py`:

1. The recognition decision for any image is a pure function of the current vision prototypes, T, and τ.
2. The synthesised audio for a label is determined by $\mathbf{p}_a^{(\ell)}$, `dur_avg`, and global synthesis parameters (pitch/tilt) recorded in the config.

4.5 Audio Pipeline

This section specifies the concrete signal path used in the implementation to turn a raw microphone stream into (i) frame–band features for learning and visualisation and (ii) a canonical, fixed–length representation for storage and comparison. The implementation mirrors the mathematical design but fixes all engineering choices, fallbacks, and file interfaces so runs are reproducible.

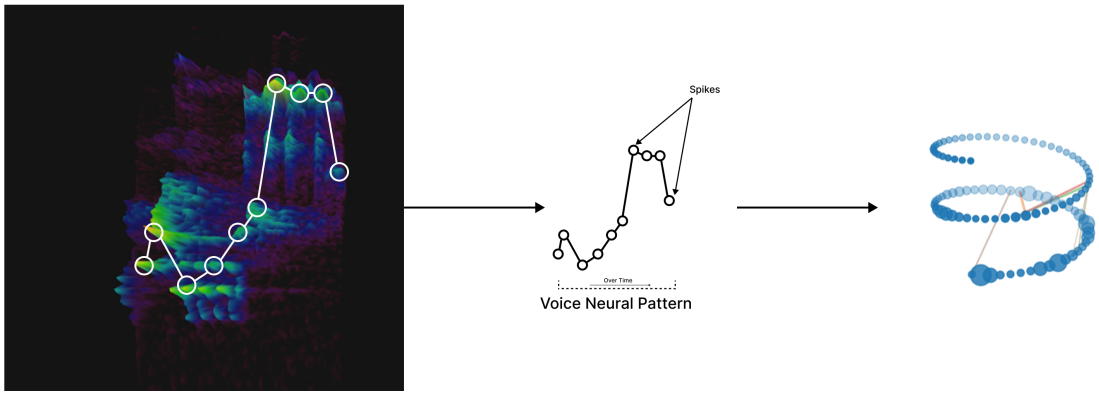

Figure 4.7: From sound wave to neural spikes pattern (Original Experiment Result)

4.5.1 Acquisition and Voice Activity Detection (VAD)

- **Sample rate.** All recordings are acquired (or resampled) at $f_s = 16\,\text{kHz}$.
- **Front–end call.** We call `FE.vad_best_segment(y, sr=fs)` which returns a trimmed segment $\tilde{y}$ around the most energetic voiced region. If VAD fails or is absent, we pass the raw buffer through unchanged.
- **Silence guard.** If $\tilde{y}$ has RMS below a small threshold (< -50 dBFS) or length $< 160\,\text{ms}$, the pipeline returns empty features and the GUI issues a "too short/silent" message.

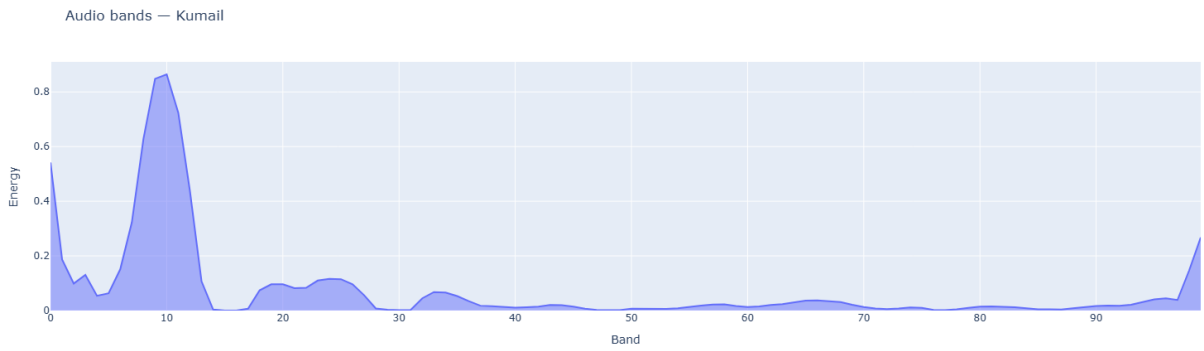

Figure 4.8: Collecting features from audio (Original Experiment Results)

4.5.2 Framing and Spectral Features

- **Frame grid.** For implementation we use a *10 ms hop* with a 20–25 ms Hann window (100 frames/s).[2]

- **Feature extractor.** We call `FE.extract_features(segment)` which returns a matrix $X \in R^{T \times N_b}$ with $N_b = 100$ frequency bands. Internally, the extractor follows Oppenheim and Schafer (1989); Slaney (1998); Moore (2012): STFT magnitude $\rightarrow$ band aggregation. Whether bands are mel–warped or linear is controlled by `config.USE_MEL_FRONTEND`.

- **Normalisation.** Each frame vector is scaled by its ℓ_∞ norm to reduce loudness dependence; values are clamped to $[0, 1]$.

- **Outputs.** We retain (a) the raw stack X for analysis/visualisation and (b) a time–average $\mathbf{a} = \text{mean}_t(X)$ for prototype audio summaries and synthesis.

4.5.3 Temporal Canonicalisation (DTW or Resample)

- **Target length.** We compress or expand every utterance to $T_c =$ `CFG.CANON_FRAMES` frames (default 64).

- **Preferred aligner.** If available, `DTW.align_to_canonical(X, T_c)` performs dynamic time warping (Sakoe and Chiba, 1978) to produce $X_c \in R^{T_c \times N_b}$.

[2]The methodology chapter illustrated some derivations with 100 ms frames for clarity; the deployed system uses 10 ms hops to preserve onsets while remaining real-time.

- **Fallback.** If DTW is unavailable, we uniformly resample each band with 1D interpolation to length T_c.

4.5.4 Quality Checks and Diagnostics

- **Duration bounds.** Extremely short (< 0.25 s) or very long (> 6 s) segments are flagged in the GUI and clipped/trimmed before feature extraction.

- **Clipping/roughness flags.** A fast spectral roughness and peak detector triggers a "safety" flag used later by the curiosity/learning-rate gate; the audio pipeline itself still returns features so the user sees what was captured.

4.5.5 Interfaces and Return Types

The GUI helper `dev_features_from_audio(audio_tuple)` (see Listing **??**) accepts a Gradio `(sr, np.ndarray)` tuple and returns:

$$(\, X \in R^{T \times N_b},\ \ sr,\ \ X_c \in R^{T_c \times N_b}\,)$$

where X is the raw frame–band stack and X_c is the canonicalised matrix.

Listing 4.1: dev_features_from_audio (simplified pseudocode)

```
def dev_features_from_audio(audio_tuple):
    if audio_tuple is None: return empty_mats()

    sr, y = audio_tuple
    y = as_float32(y).ravel()

    # 1) VAD (robust to failure)
    if hasattr(FE, "vad_best_segment"):
        seg, _ = FE.vad_best_segment(y, sr=sr)
        if seg is None or len(seg)==0: seg = y
    else:
        seg = y
```

```
    # 2) Features (100 bands; mel or linear per config)
    if hasattr(FE, "extract_features"):
        X = FE.extract_features(seg)
    else:
        X = safe_random_fallback(seg, sr, N_b=100)

    # 3) Canonicalise to T_c frames (DTW preferred)
    T_c = CFG.CANON_FRAMES
    if hasattr(DTW, "align_to_canonical"):
        Xc = DTW.align_to_canonical(X, T_c)
    else:
        Xc = uniform_resample(X, T_c)

    return X, sr, Xc
```

4.5.6 Determinism and Reproducibility

All tunable (f_s, hop size, N_b, T_c, mel/linear switch) parameters are given in `config.py` and recorded in `brain.json.meta`. In case `FE` or `DTW` is unavailable, fallbacks are deterministic based on the global seed and therefore the runs are comparable across machines.

4.5.7 Why These Choices?

- **10 ms hop** maintains onsets/forms and is still lightweight for real time.
- **100 bands** results in a cochlea-like, tonotopic sheet with sufficient resolution for synthesis, and for simple spike thresholding, in line with auditory practice (Slaney, 1998; Moore, 2012).
- **DTW to fixed** T_c, robust prototype averaging and cross-system comparison across varying speaking rates are made possible (Sakoe and Chiba, 1978).

4.6 Vision Pipeline

This section specifies the concrete path that turns a user–selected image into (i) biologically inspired analysis views (retina → V1 "IT-style" panel) and (ii) a compact, invariant feature vector used for enrolment, recognition, and the 3D cross–modal visualisations. The emphasis here is on the exact engineering choices, defaults, and fallbacks implemented in the codebase so runs are reproducible.

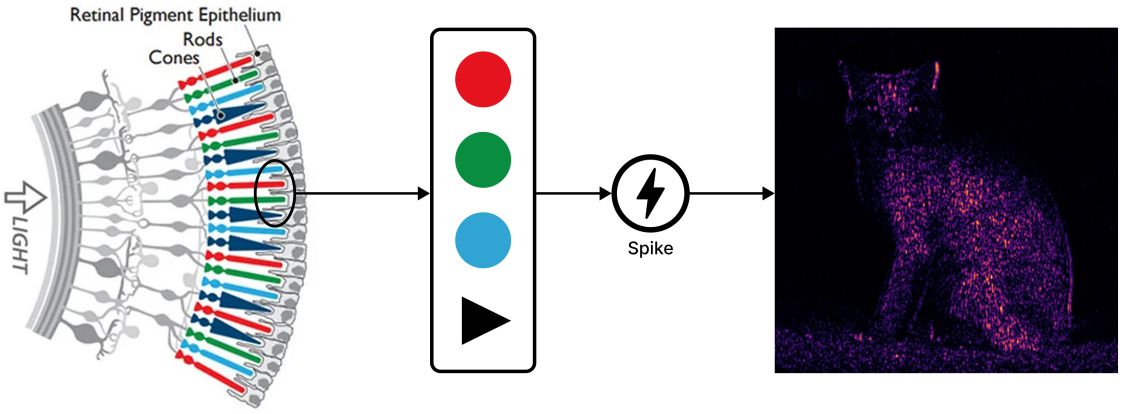

Figure 4.9: From light in retina to neural spikes pattern (Original Experiment Result)

4.6.1 Image Ingestion and Gallery Handling

- **Inputs.** Users add one or more image files via the gallery. Each path is read in BGR (OpenCV) or RGB (PIL fallback).
- **Aspect–preserving resize.** Every image is resized to a *fixed width* of 500 pixels while *preserving aspect ratio*. This guarantees stable feature statistics while avoiding distortion:

$$(w, h) \;\mapsto\; (500,\ \mathrm{round}(h \cdot 500/w)).$$

- **Color management.** Internally we retain BGR for OpenCV ops; for on–screen display we convert to RGB. A single in–memory copy per image is kept to avoid drift.
- **Selection.** Clicking a tile in the gallery emits its 0-based index to the "Selected index" field automatically (no manual typing).

4.6.2 Retina–Style Prefilter (for analysis views)

For explanatory visualisations, we render an RGC *ON/OFF* pair by a light difference–of–Gaussians (DoG):

$$ON = \max\{G_{\sigma_1} * I - G_{\sigma_2} * I,\, 0\}, \quad OFF = \max\{G_{\sigma_2} * I - G_{\sigma_1} * I,\, 0\},$$

with $\sigma_1 = 1.2$, $\sigma_2 = 2.4$ pixels on the normalised grayscale $I \in [0, 1]$. These maps approximate center–surround receptive fields and feed the panel below.

4.6.3 V1–like Gabor Bank (for analysis views)

We convolve the DoG base with a small *Gabor* bank (8 orientations, 3 spatial scales) to visualise orientation/scale energy:

- **Orientations:** $\theta \in \{0, \pi 8, \ldots, 7\pi 8\}$.
- **Kernel sizes:** $k \in \{7, 11, 15\}$ with wavelength $\lambda = \max(4, 0.6k)$, sigma $\sigma = 0.5k$, aspect $\gamma = 0.5$.
- **Energy map:** $E = I * g_{\theta,k}$.

The "Vision IT" panel (RGC ON/OFF + 8×3 energy thumbnails) is *explanatory*; it is not the recognition feature. It helps debug orientation content and scale selectivity over the user's image.

4.6.4 Recognition Feature (compact, invariant)

For enrolment and recognition we compute a single, compact feature vector:

1. **Primary path (if available).** Call `V.encode(bgr)` from `vision_babyai.py` and ℓ_2-normalise the output.
2. **Stable fallback (always available).** A HOG-ish descriptor with strong invariances:
    - Convert to grayscale in $[0, 1]$.
    - Downsample to 32×32 (area/Lanczos).

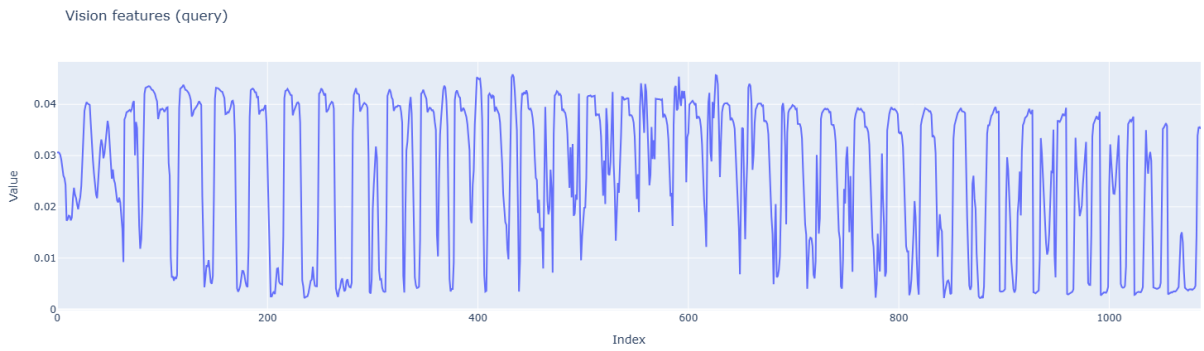

Figure 4.10: Collecting features from image (Original Experiment Results)

- Concatenate three terms:
  (a) Column means (32),
  (b) Row means (32),
  (c) Flattened 32×32 patch (1024).

  Total dimension 1088. Finally apply ℓ_2 normalisation.

This descriptor is robust to modest translation/scale, inexpensive to compute, and deterministic—matching the dimensions seen in experiments.

Listing 4.2: `encode_vision_feature` **(simplified).**

```
def encode_vision_feature(bgr):
    try:
        if V is not None and hasattr(V, "encode"):
            return l2(V.encode(bgr))                # primary path
    except: pass
    # fallback:
    gray = to_gray01(bgr)                           # [0,1]
    small = resize(gray, 32x32, area/Lanczos)
    feat = concat([mean_cols(small),                # 32
                   mean_rows(small),                # 32
                   small.reshape(-1)])              # 1024
    return l2(feat)                                 # 1088-D
```

4.6.5 Prototype Maintenance (persistent brain store)

- **Per–label prototype.** For each label ℓ we maintain an ℓ_2-normalised prototype vector $\mathbf{p}^{(\ell)}$.
- **EMA update.** On enrolment we apply an exponential moving average:

$$\mathbf{p}^{(\ell)} \leftarrow \mathrm{norm}(\alpha\, \mathbf{p}^{(\ell)} + (1-\alpha)\, \mathbf{v}),$$

  with $\alpha = 0.9$ during normal enrolment, $\alpha = 0.95$ when the user *confirms* a correct recognition (stronger consolidation).
- **Persistence.** Prototypes live in `brain.json` under `labels[lab].vision.proto` with a small metadata header for reproducibility.

4.6.6 Similarity for Recognition (vision side)

Given a query vector $\mathbf{q}$ and each stored prototype $\mathbf{p}^{(\ell)}$, we compute cosine similarity $s_\ell = \langle \mathbf{p}^{(\ell)}, \mathbf{q} \rangle$ on the *overlap length* when dimensions differ. The similarity set is passed to the calibration gate (temperature softmax + τ threshold) described in the recognition section of Implementation.

4.6.7 Interfaces and Return Types

- **Feature for enrol/recognise.** `encode_vision_feature(bgr)` returns $\mathbf{v} \in R^{1088}$ (or model-specific if `V.encode` is present), always ℓ_2-normalised.
- **Analysis panel.** `vision_it_panel(bgr)` returns a figure containing RGC ON/OFF and the 8×3 Gabor energy grid for inspection.

4.6.8 Determinism and Reproducibility

All pre–processing (fixed width 500 px, color conversion, grayscale normalisation, 32×32 fallback size, and ℓ_2 norm) is fixed. No random augmentation is applied in the GUI path, so the same image yields the same feature every run.

4.6.9 Why These Choices?

- **Fixed 500-px width** stabilises feature statistics across heterogeneous inputs while keeping the gallery responsive.
- **HOG-ish fallback** guarantees functionality on any machine (no heavyweight model dependency) and matches the dimensionality used elsewhere (1088-D).
- **Separate analysis vs. recognition paths** ensures the explanatory V1 visualisation can be richly parameterised without perturbing the compact recognition feature.

4.7 Enrollment Workflow

This section details how a new *concept* (person/label) is enrolled from one clicked image and one recorded voice example. The workflow runs end–to–end in a single interaction and updates a persistent store (`brain.json`) so knowledge survives restarts.

4.7.1 Inputs and Preconditions

- **Image:** one gallery tile (auto–selected on click), already resized to fixed width 500px with aspect preserved and kept in BGR (for processing) and RGB (for display).
- **Voice:** one utterance recorded at $f_s = 16$ kHz (NumPy array). Voice activity detection (VAD) trims leading/trailing silence.
- **Label ℓ:** a user–provided string (e.g., ``Kumail'').

4.7.2 Feature Extraction

Vision feature v. We compute a compact, ℓ_2-normalised descriptor from the selected image:

$$\mathbf{v} \;=\; \mathit{norm}(\text{encode\_vision\_feature}(\text{BGR}_{500w})) \in R^{d_v},$$

where the primary path calls `V.encode` (if available); otherwise a deterministic HOG-ish fallback produces $d_v = 1088$ (32 column means + 32 row means + 32×32 patch), cf.

Sec. 4.6.

Audio features X, a. The recorded waveform is passed through the *developmental* auditory pipeline (Sec. 4.5): VAD → spectral features → (optional) DTW canonicalisation.

$$(X,\, f_s,\, X_c) \;=\; \text{dev\_features\_from\_audio(wave)},$$

where $X \in R^{T\times N_f}$ are frame–wise band energies (default $N_f = 100$), $X_c \in R^{T_c\times N_f}$ is the time–aligned version used for visualisations. The *audio prototype vector* persisted at enrolment is the time–average:

$$\mathbf{a} \;=\; \frac{1}{T}\sum_{t=1}^{T} X_{t,:} \;\in\; R^{N_f}.$$

4.7.3 Auto–dopamine (novelty ⊕ agreement)

A scalar DA $\in [0,1]$ is computed for logging/inspection using a three–factor intuition (agreement, novelty) that mirrors neuromodulated plasticity (Hebb, 1949; Bi & Poo, 1998; Izhikevich, 2007):

$$\text{agreement} \;=\; 12(1+\langle\widehat{\mathbf{v}},\,\widehat{\mathbf{a}}\rangle), \quad \text{novelty} \;=\; 1-\max_{m\in\mathcal{M}}\langle\widehat{\mathbf{v}},\,\widehat{m}\rangle,$$

$$\text{DA} \;=\; w_{\text{agree}}\cdot\text{agreement} \;+\; w_{\text{novel}}\cdot\text{novelty}, \quad w_{\text{agree}}{+}w_{\text{novel}} = 1.$$

Hats denote ℓ_2-normalisation; $\mathcal{M}$ is the set of existing vision prototypes. (If there is no prior memory, novelty = 0.5.) Agreement uses cosine similarity after adaptive length–pooling so $\mathbf{v}$ and $\mathbf{a}$ are comparable.

4.7.4 Prototype Formation and Consolidation

Vision (EMA prototype). For label ℓ, we maintain an ℓ_2-normalised prototype vector $\mathbf{p}^{(\ell)}$. On enrolment:

$$\mathbf{p}^{(\ell)} \;\leftarrow\; \mathit{norm}\Big(\alpha\,\mathbf{p}^{(\ell)} \;+\; (1{-}\alpha)\,\mathbf{v}\Big), \qquad \alpha = 0.9.$$

If the user later *confirms* a correct recognition, we apply a stronger consolidation ($\alpha = 0.95$).

Audio (time–average prototype). We persist $\mathbf{a}$ and its sample rate f_s as the minimal spectral prototype used for speaking/synthesis and debugging:

$$\texttt{labels}[\ell].\texttt{audio.feat} \leftarrow \mathbf{a}, \quad \texttt{labels}[\ell].\texttt{audio.sr} \leftarrow f_s.$$

The canonicalised matrix X_c is produced on–the–fly for visualisations (helix; co–firing) and does not need to be stored to guarantee synthesis.

4.7.5 Persistence: JSON Schema (excerpt)

The persistent brain store is minimal and human–readable:

```
{
  "labels": {
    "<lab>": {
      "vision": { "proto": [ .. d_v floats .. ] },
      "audio":  { "feat": [ .. N_f floats .. ], "sr": 16000 }
} }, "meta": { "version": 1 } }
```

Files are written atomically via `brain.json.tmp` → rename, avoiding partial writes.

4.7.6 Outputs and Side Effects

- Status message with DA score and dimensions (T, N_f, d_v).
- Three plots: vision feature trace, audio band area plot, and a co–firing spike animation (from X_c).
- Updated `brain.json` with the new/updated prototypes.

4.7.7 Failure Modes and Guards

- **No image/voice/label:** enrolment aborts with a user–facing message.

- **Silent/too short audio:** VAD may return near–empty segments; we reject with guidance to re–record.
- **Dimensional mismatch:** cosine computations always operate on pooled, normalised vectors of common length to eliminate shape errors.

4.7.8 Pseudocode

```
function ENROL(img_idx, wave, label):
    assert label != "", img_idx in gallery
    v = l2(encode_vision_feature(GALLERY[img_idx]))
    (X, fs, Xc) = dev_features_from_audio(wave)
    assert X not empty
    a = mean_over_time(X)          # audio prototype vector
    DA = auto_dopamine(v, a, memory=all_vision_protos)

    # persist
    Vproto[label] = l2(EMA(Vproto[label], v, alpha=0.9))
    Aproto[label] = { feat: a, sr: fs }
    save_json(brain)

    # visuals
    return status(DA), plot_v(v), area_a(a), cofiring(Xc)
```

4.7.9 Rationale

This design couples a *rich analysis view* (retina/V1 panels; co–firing animations) with a *compact recognition core* (one ℓ_2-normalised vision vector; one time–averaged audio vector). The EMA rule yields stability across multiple exemplars, while DA offers a biologically motivated, scalar summary of "how aligned and how novel" an enrolment is, without blocking progress when data are scarce.

4.8 Recognition, Calibration, and Confirmation

This section details how a clicked *query image* is recognised against the persistent store, how confidence is *calibrated* with temperature, how *open–set* abstention is enforced with a threshold τ, and how *user confirmation* reinforces correct matches.

4.8.1 Query Feature and Prototype Set

Given a selected gallery image, we compute an ℓ_2-normalised vision descriptor

$$\mathbf{q} \;=\; \mathit{norm}(\text{encode\_vision\_feature}(\text{BGR}_{500w})) \in R^{d_v},$$

using the same encoder as enrollment (Sec. 4.6). The persistent store $\mathcal{B}$ provides a set of *vision prototypes*

$$\mathcal{P} \;=\; \{(\ell, \mathbf{p}^{(\ell)})\}_{\ell \in \mathcal{L}}, \qquad \mathbf{p}^{(\ell)} \in R^{d_v},\ \|\mathbf{p}^{(\ell)}\|_2 = 1,$$

each maintained by EMA during enrollment/confirmation (Sec. 4.7).

4.8.2 Similarity and Temperature Calibration

For every label ℓ, we compute cosine similarity

$$s_\ell \;=\; \langle \mathbf{q},\, \mathbf{p}^{(\ell)} \rangle \;\in [-1, 1].$$

To obtain *calibrated confidences* we apply a temperature-softmax $\mathrm{T} > 0$ (Guo *et al.*, 2017):

$$z_\ell \;=\; \frac{s_\ell - \max_j s_j}{\mathrm{T}}, \qquad \pi_\ell \;=\; \frac{e^{z_\ell}}{\sum_j e^{z_j}},$$

yielding a probability-like vector π that is numerically stable (shift by $\max s$). Lower T sharpens distributions; higher T flattens them. In practice, $\mathrm{T} \in [0.03, 0.15]$ balances separation and numerical stability.

Top-k display. For transparency and debugging, we retain a compact dictionary of the top-k calibrated scores

$$\text{scores\_topk} \;=\; TopK\Big(\{(\ell, \pi_\ell)\}_\ell\Big), \quad k \in \{3, 5\}.$$

4.8.3 Open–Set Thresholding (τ)

Open–set recognition requires the system to *abstain* when no known class is sufficiently plausible (Scheirer *et al.*, 2013; Bendale & Boult, 2016). Let $\ell^\star = \arg\max_\ell \pi_\ell$ and $\hat{c} = \pi_{\ell^\star}$. We accept the prediction only if

$$\hat{c} \;\geq\; \tau, \qquad \tau \in [0, 1),$$

otherwise we return ABSTAIN and optionally prompt for enrollment. The τ-gate reduces false positives on out-of-distribution faces/images (Hendrycks & Gimpel, 2017).

4.8.4 Decision, Messaging, and Guards

- If $\mathcal{L} =:$ report *"No concepts enrolled yet."*
- Else compute $(\ell^\star, \hat{c}, \text{scores\_topk})$.
- If $\hat{c} < \tau$: ABSTAIN with calibrated scores.
- Else: emit $\ell^\star$ and $\hat{c}$ with the scores table for user inspection.

All similarities operate on ℓ_2-normalised vectors of the *same* length (the encoder is identical for query and prototypes), eliminating shape mismatches.

4.8.5 Human-in-the-Loop Confirmation

If the user confirms the prediction is correct, we *reinforce* the vision prototype via a higher-gain EMA:

$$\mathbf{p}^{(\ell^\star)} \;\leftarrow\; norm\Big(\alpha_{conf}\, \mathbf{p}^{(\ell^\star)} + (1{-}\alpha_{conf})\, \mathbf{q}\Big), \quad \alpha_{conf} \approx 0.95,$$

which tightens the prototype around verified exemplars and improves future confidence. If the user rejects the suggestion, we *do not* update any prototype; optionally the UI can offer to "enroll as new label".

4.8.6 Optional Overrides

In downstream actions such as *Speak*, a "force-top" switch may ignore τ and pick $\ell^\star$ purely by $\max s_\ell$. This is disabled by default to preserve open–set safety.

4.8.7 Pseudocode

```
function RECOGNISE(img_idx, T, tau):
    require brain.labels not empty
    q = l2(encode_vision_feature(GALLERY[img_idx]))
    scores = { ell: dot(q, Vproto[ell]) for ell in brain.labels if Vproto[ell] exists
    if scores empty: return Abstain, {}, None
    # temperature calibration
    smax = max(scores.values())
    probs = { ell: exp((s - smax)/T) for ell,s in scores.items() }
    Z = sum(probs.values()); probs = { ell: p/Z for ell,p in probs.items() }
    ell_star = argmax(probs)
    conf = probs[ell_star]
    if conf < tau: return Abstain, topk(probs), None
    else: return Predict(ell_star, conf), topk(probs), ell_star


function CONFIRM(img_idx, ell_pred):
    q = l2(encode_vision_feature(GALLERY[img_idx]))
    Vproto[ell_pred] = l2( alpha*Vproto[ell_pred] + (1-alpha)*q )  # alpha=0.95
    save_json(brain)
```

4.8.8 Design Rationale

Temperature calibration reduces overconfidence and yields interpretable probabilities from raw cosine scores (Guo *et al.*, 2017). **Open–set gating** with τ implements conser-

vative rejection on unfamiliar inputs (Scheirer *et al.*, 2013; Bendale & Boult, 2016), complementing simple confidence heuristics (Hendrycks & Gimpel, 2017). **Confirmation-driven EMA** realises user-in-the-loop consolidation that steadily improves robustness without retraining, and is consistent with incremental prototype learning.

4.9 Synthesis ("Speak")

This section details how the system renders an audible waveform from a stored auditory prototype. The procedure prefers an external synthesiser (if present), and otherwise falls back to a lightweight, fully deterministic additive sinusoid renderer driven by band–energy trajectories.

4.9.1 Prototype to Render Target

Each enrolled concept ℓ stores a time–frequency prototype $P^{(\ell)} \in R^{T_c \times N_b}$ (Sec. 4.5). The render target is a duration τ_ℓ (seconds) with a sample rate $f_s = 16\,kHz$.

Time normalisation and smoothing. Given $P^{(\ell)}$ and τ_ℓ, we obtain a frame trajectory $\{\mathbf{p}_t\}_{t=1}^{T}$, $T = \lfloor \tau_\ell / \Delta t \rfloor$, by linear interpolation in time (or DTW back–projection if available; cf. Sec. 4.5). A mild temporal low–pass (e.g., 3-frame moving average) reduces frame-to-frame flicker before synthesis:

$$\widetilde{\mathbf{p}}_t \;=\; \frac{1}{3}(\mathbf{p}_{t-1} + \mathbf{p}_t + \mathbf{p}_{t+1}).$$

4.9.2 Spectral Shaping and Frequency Mapping

Let the tonotopic centre frequencies be $\{f_n\}_{n=1}^{N_b}$ (e.g., quasi–mel spacing over 90–6000 Hz; Moore2012,Slaney1998). Optional parametric shaping is applied framewise:

1. **Pitch/formant warps (optional).** A simple frequency–axis warp is used to simulate pitch or formant shifts:

$$\mathbf{p}_t^{(warp)}[f] \;=\; interp\left(\widetilde{\mathbf{p}}_t\left[\frac{f}{s_{pitch}\; s_{formant}^{1/2}}\right]\right),$$

where $s_{pitch} > 0$ and $s_{formant} > 0$ are user/age-profile controls (StevensVolkmannNewman1937,McAulayQuatieri1986).

2. **Spectral tilt.** A gentle broadband tilt simulates vocal tract emphasis:

$$\mathbf{p}_t^{(tilt)}[f] \;=\; \mathbf{p}_t^{(warp)}[f] \cdot 10^{-\frac{\kappa}{20}\log_{10}\max(f,1)},$$

with κ (dB/decade) tuned by self–evaluation (below).

3. **Amplitude nonlinearity.** Per–band synthesis weights are

$$a_{t,n} \;=\; (\mathbf{p}_{t,n}^{(tilt)})^{\gamma}, \qquad \gamma \in [0.6, 1.0],$$

which compresses dynamic range and reduces artefacts (OppenheimSchafer1989).

4.9.3 Additive Sinusoid Renderer (Fallback Core)

We generate a monophonic signal by summing sinusoids at the fixed bin centres $\{f_n\}$. Let Δt be the frame hop (e.g., $10\,ms$). Within each frame, instantaneous amplitudes are held constant, phases are continuous:

$$x[m] \;=\; \sum_{n=1}^{N_b} a_{t(n,m),n}\,\sin(\phi_n[m]), \qquad \phi_n[m{+}1] \;=\; \phi_n[m] + 2\pi\frac{f_n}{f_s},$$

where m indexes samples and $t(n,m) = \lfloor m/(f_s\Delta t)\rfloor$. To avoid clicks, an *attack–sustain–release* envelope $e[m]$ (e.g., $20\,ms$ linear ramps) is applied:

$$\widetilde{x}[m] \;=\; e[m]\cdot x[m], \qquad \widetilde{x} \leftarrow \frac{\widetilde{x}}{\max|\widetilde{x}|}\cdot 10^{-3/20},$$

normalising to $-3\,dBFS$ target level.

Remarks. This additive method is deterministic, fast, and pairs naturally with band–energy features. It is related in spirit to sinusoidal modelling of speech (McAulayQuatieri1986) and can be extended with phase–locking or harmonic grouping if higher fidelity is required.

4.9.4 External Synthesiser Path (If Available)

If a project–specific synthesiser exposes `speak_and_save(label, features, ...)` (e.g., a lightweight vocoder, Griffin–Lim, or source–filter variant; GriffinLim1984), the system first attempts that path. On success, its output file is returned; on failure or absence, the additive fallback above guarantees an audible result.

4.9.5 Self–Evaluation and Micro–Tuning

After rendering, the audio is re–encoded by the same frontend to $X^{(out)}$, time–aligned to the prototype $P^{(\ell)}$ (Sec. 4.5), and a band–wise MSE is computed:

$$MSE \;=\; \frac{1}{TN_b}\sum_{t,n}(X^{(out)}_{t,n} - \widetilde{P}^{(\ell)}_{t,n})^2.$$

Heuristics adjust κ (tilt), γ (nonlinearity), and envelope lengths to reduce systematic band–region error on subsequent speaks. This closes a light self–calibration loop without heavy optimisation.

4.9.6 Pseudocode

```
function SPEAK(label):
    P, fs, tau = load_prototype(label)              # T_c x N_b, 16k, duration
    if has_external_synth():
        path = external_synth(label, P, tau)
        if path exists: return path
    # fallback additive
    frames = resample_time(P, T = round(tau/t))
    frames = temporal_smooth(frames)
    frames = warp_pitch_formant(frames, s_pitch, s_formant)
    frames = apply_tilt(frames, )
    A = frames **
    y = zeros(round(tau*fs))
    for n in 1..N_b: _n = 0
    for m in 0..len(y)-1:
```

```
        t = floor(m/(fs*t))
        y[m] = sum_n A[t,n] * sin(_n)
        _n += 2 f_n / fs
    y = apply_ASR_envelope(y, fs, 20ms)
    y = normalise_to(y, -3 dBFS)
    path = save_wav(y, fs)
    return path
```

4.9.7 Design Choices

The fallback renderer trades timbral realism for robustness and interpretability: (i) it cannot fail silently; (ii) every control ($\kappa, \gamma, s_{pitch}, s_{formant}$) has a clear acoustic effect; (iii) it aligns directly with the feature representation. Where available, an external synthesiser may improve naturalness, but the additive pathway ensures end–to–end functionality with no additional dependencies.

4.10 Visualisation Layer (Interactive)

This layer exposes the internal state of the system in real time through interactive graphics. It serves two purposes: (i) *diagnostics*—to make each stage of the pipeline observable; and (ii) *didactics*—to communicate biological intuitions (tonotopy, co-firing, hemispheric coupling). The widgets share the same inputs (currently selected image and most recent audio), so the user does not have to re-enter data per view.

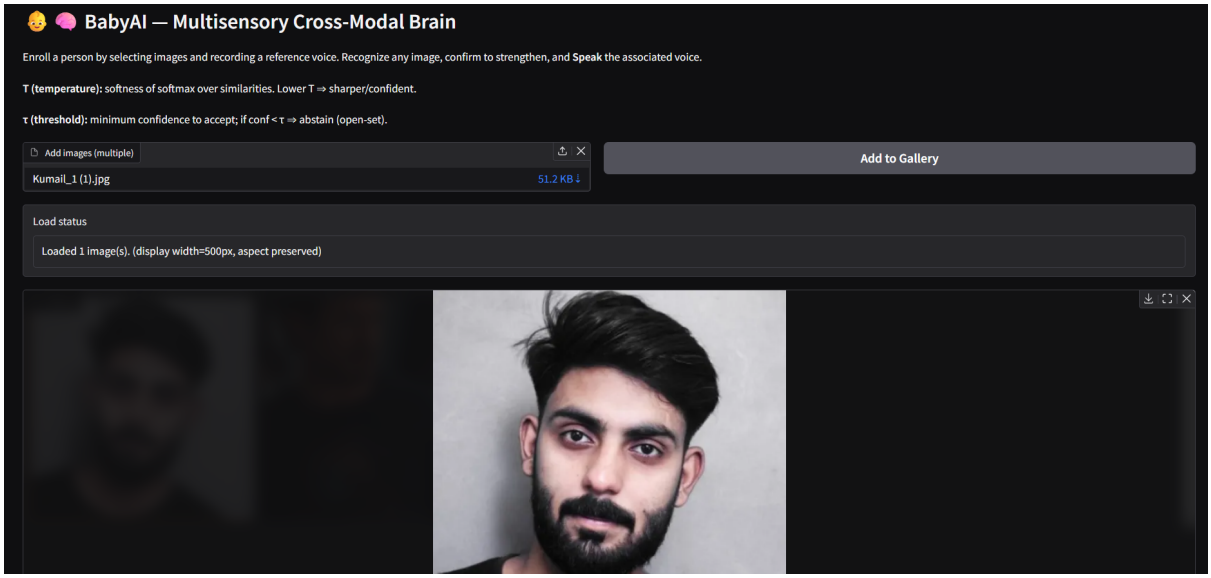

Figure 4.11: Main Window of Project BabyAI

4.10.1 Design Objectives

- **Faithful**: plots are direct functions of the tensors used by the algorithms (no mock data).
- **Linked**: all views reference the same selected image index and audio buffer.
- **Lightweight**: pre-compute, cache and downsample to keep interactivity fluid on a laptop GPU/CPU.
- **Explainable**: colour, size and motion encode interpretable quantities (energy, confidence, time).

4.10.2 Components and Data Mappings

(1) Scrollable image gallery with auto-selection. All added images are width–normalised to 500 px while preserving aspect ratio; thumbnails are stored as RGB for display and BGR for processing. Clicking a tile updates the `selected_index` state, which *fan-outs* to every downstream callback (IT, Helix, 3D Brain, Recognise/Speak).

(2) Vision IT panel (RGC → V1 energy grid). Given the selected image, the panel shows:

1. *RGC ON/OFF*: centre–surround difference visualised as two maps.

2. *V1-like filters*: a grid over orientations ($0° \dots 180°$) and kernel scales; each cell displays $|I * g_{\theta,k}|$.

This is rendered as a static mosaic for quick inspection; it refreshes instantly on a new selection.

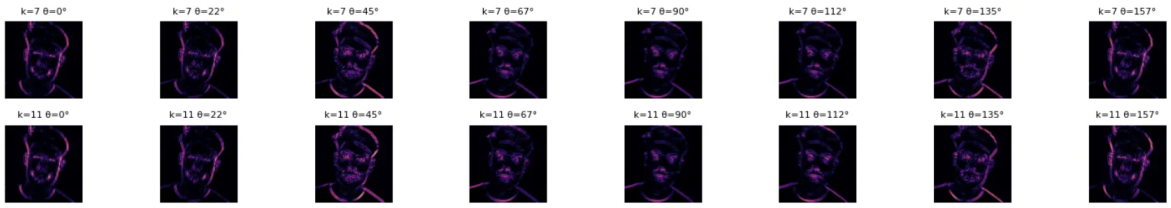

Figure 4.12: Vision IT RGC & V1 Energy grid (Original Experiment Results)

(3) Voice Helix (3D; animated). From the current audio, the time–frequency matrix $X \in R^{T \times N}$ (after VAD and feature extraction) is embedded as a helix parameterised by time:

- Backbone $(x(t), y(t), z(t))$ traces time; colour encodes frame energy.
- Up to K top frequency bands are drawn as braided filaments around the backbone; each filament's width and colour reflect band energy.
- A moving white cursor indicates the animation frame.

The helix uses a compact frame stride for responsiveness and exposes play/pause.

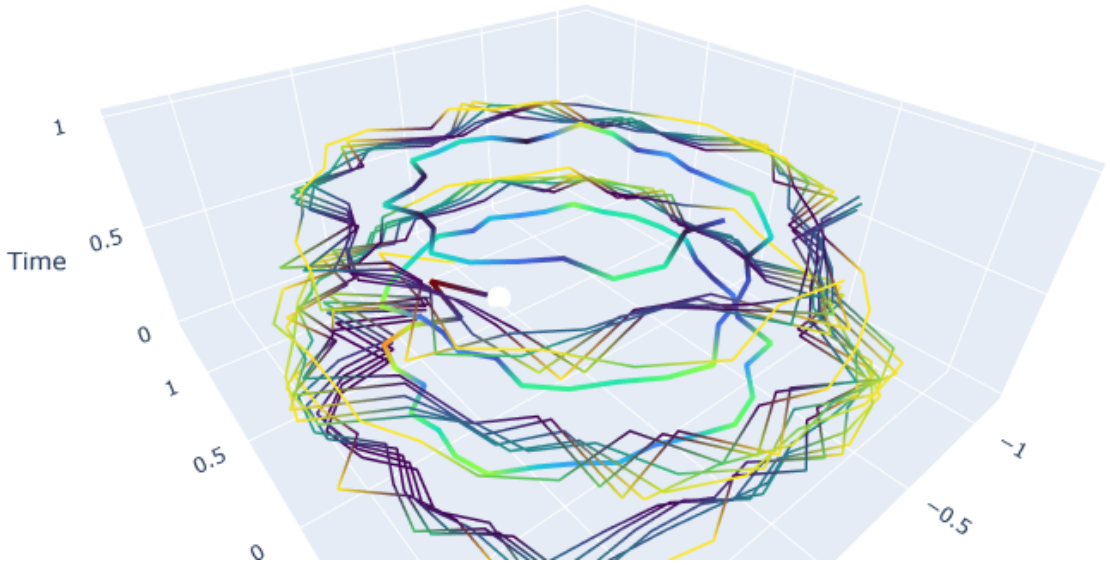

Figure 4.13: Voice helix graph (Original Experiment Results)

(4) Co-firing spikes (heatmap; animated). Binary spike masks are derived from normalised amplitudes with fixed thresholds. For each frame t, a heatmap of $\mathbf{v}_{spk}\mathbf{a}_{spk}^{\top}$ visualises potential cross-modal co-firing (vision units on rows, audio units on columns). Animation reveals spatio-temporal structure (e.g., formant onsets).

(5) 3D Brain graph (interactive; animated). Two point clouds represent hemispheres:

- *Left (vision)*: vertex size and colour map to the magnitude of the selected image's feature vector.
- *Right (audition)*: vertex size and colour map to the mean (or current frame) band energies.
- *Edges*: per animation frame, top-k cross-modal products connect left and right nodes; an *edge threshold* slider hides weak links to reduce clutter.

This display illustrates time-varying cross-modal coupling with an intuitive hemispheric metaphor.

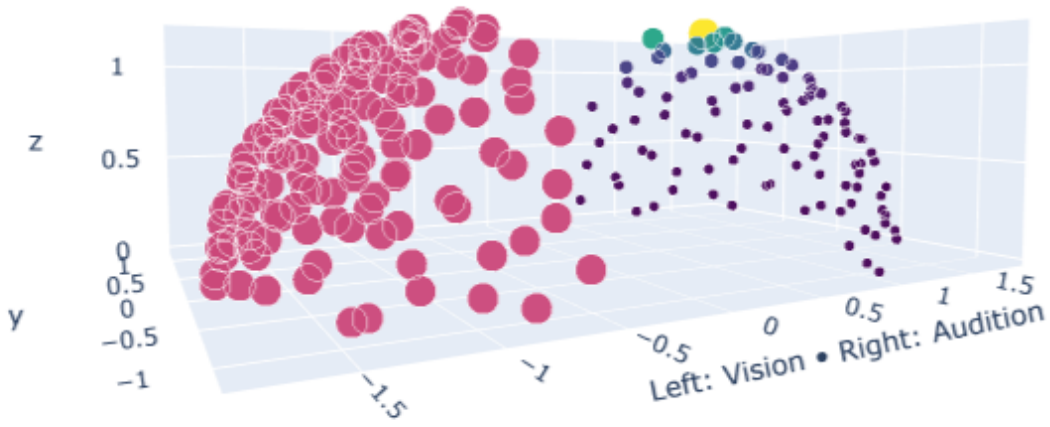

Figure 4.14: Neurons firing in brain according to input (Original Experiment Results)

(6) Recognition readout (calibrated; τ-gated). The same figure window shows the predicted label and a compact JSON of calibrated posteriors. A temperature slider T controls the softness of the normalised scores (cf. Sec. **??**); a threshold slider τ implements

open-set abstention. When $conf < \tau$, the UI displays *"abstained"* with the top-k scores for transparency.

4.10.3 Shared Controls and Wiring

Selected image index set by gallery click; drives all vision-dependent panels.

Audio buffer set by the microphone widget; drives Helix, Brain and *Speak*.

T **(temperature)** affects only the recognition softmax normalisation (no effect on features/plots).

τ **(threshold)** gates recognition acceptance and the rendering of Brain edges (minimum edge weight).

Generate All triggers IT, Helix and Brain with the current image+audio in a single action.

4.10.4 Performance Notes

- **Image path**: width-normalise once (500 px), reuse cached BGR/RGB arrays; avoid re-decoding files.
- **Audio path**: cache features per recording; store X, canonical X_c and the bandwise mean.
- **3D plots**: cap K filaments and top-k edges; coarsen animation step for long utterances.
- **Asynchronous feel**: heavy computations (feature extraction, DTW) happen before plotting, so the figure construction remains snappy.

4.10.5 Export and Reproducibility

Every interactive figure can be exported as HTML (self-contained) or as static PNG for the thesis. A run log stores: selected image ID, audio hash, T, τ, and random seeds—allowing figure regeneration.

4.10.6 Failure Modes and User Feedback

- *Missing audio/image*: plots show a concise hint ("Add images first" / "Record a reference voice").
- *Low SNR or silence*: Helix/Brain display "audio too short/silent" rather than stale content.
- *Open-set abstention*: the status line reports confidence and τ; JSON panel exposes top-k scores.

4.10.7 Accessibility Considerations

We favour high-contrast palettes, avoid red–green conflicts, and label axes/titles explicitly. Tooltips summarise slider effects; keyboard focus targets the gallery and buttons for mouseless operation.

4.10.8 Rationale

The layer is intentionally *observational*: it visualises the same tensors used in computation (no surrogate states) and ties UI controls (T, τ) to the formal definitions in Sec. **??**. This alignment preserves scientific traceability while giving an intuitive, biologically motivated window into the system's dynamics.

4.11 Memory Decay, Pruning, and Homeostasis

Where decay is applied. We apply a global L_2–style shrinkage to the recurrent tonotopic weights $W \in R_{\geq 0}^{N_f \times N_f}$ once per learning step:

$$W \;\leftarrow\; (1-\lambda)\, W \quad with \quad \lambda \in [0,1).$$

Optionally, usage–modulated decay can be enabled:

$$\lambda_{ij} \;=\; \lambda_0(1-\rho_{ij}), \qquad \rho_{ij} = EMAusage(i,j) \in [0,1],$$

so rarely used edges decay faster.

Prototype ageing when unused. Each concept prototype $P^{(\ell)} \in R^{T_c \times N_f}$ carries `last_seen` timestamps. On each save (or hourly tick), if no new alignment occurred for ℓ, apply slow EMA shrinkage toward a neutral baseline (zero-centred normalised spectrum):

$$P^{(\ell)} \leftarrow (1 - \alpha_{age})\, P^{(\ell)} + \alpha_{age}\, \overline{P}, \qquad \alpha_{age} \ll 1.$$

This prevents "stale dominance" without catastrophic forgetting.

Pruning & homeostasis. After updates we sparsify and renormalise:

$$W_{ij} \leftarrow 0 \quad if \quad W_{ij} < \tau_{prune},$$

then apply row–wise homeostasis to avoid run–away excitation:

$$W_{i,:} \leftarrow \frac{W_{i,:}}{\max(\epsilon, \|W_{i,:}\|_1)} \cdot c_1 \quad or \quad W_{i,:} \leftarrow \frac{W_{i,:}}{\max(\epsilon, \|W_{i,:}\|_2)} \cdot c_2,$$

where c_1 or c_2 fix a target row mass.

Execution points. *Per-frame* (after plasticity), *per-enrolment* (after prototype EMA) and *on save* (cleanup pass with stricter pruning).

Table 4.4: Forgetting/ageing parameters and effects.

Parameter	**Symbol**	**Typical**	**Effect**
Global decay rate	λ	10^{-3}–10^{-2}	Smoothly contracts W each step (stability).
Usage–modulated gain	λ_0	2λ	Faster decay on low-usage edges.
Prototype ageing	α_{age}	10^{-4}–10^{-3}/save	Very slow drift of $P^{(\ell)}$ when idle.
Prune threshold	τ_{prune}	0.03–0.08	Enforces sparsity, frees capacity.
Row mass target	c_1 or c_2	1.0	Homeostatic normalisation per row.

4.12 Robustness & Engineering Fixes

Dimension mismatch (vision vs. audio). Observed: `ValueError: shapes (1088,) and (100,)` during DA computation. **Fix:** length pooling to a common length $L =$

$\min(d_v, d_a)$; average–pool the longer vector into L bins before cosine.

`frontend.extract_features` **signature.** Observed: unexpected kwarg `sr=`. **Fix:** call the function as `FE.extract_features(y)`; sampling rate comes from `config.py`.

HTTP `Content-Length` **errors.** Observed with Uvicorn/h11 when callbacks returned `None`. **Fix:** always return valid types: status string + figure (or `None`) in the exact signature Gradio expects; ensure files exist before returning file paths.

Gallery UX. Added width–normalised thumbnails (500 px), scrollable gallery height, and auto–index update via `Gallery.select` → `gr.SelectData.index`.

Persistent brain. Atomic saves: write `brain.json.tmp` then `os.replace` to `brain.json`. Survives restarts and incremental growth.

Table 4.5: Bug → Fix → Test matrix.

Bug / Symptom	**Fix**	**Test / Check**
Cosine shape mismatch	Length pooling to L, then cosine	Unit test with $(512, 100)$ and $(100, 512)$ vectors.
Feature signature error	Remove `sr=` kwarg	Smoke test: `dev_features_from_audio()` on 1 s clip.
Content-Length crash	Always return typed tuples; ensure file exists	Run click–paths; assert HTTP 200; WAV stats > 0 s.
Gallery not scrollable	Fixed height; width–normalise	Add > 30 images; scroll and select.
State lost on restart	Atomic JSON save; lazy load	Restart app; prior labels present.

4.13 Performance Considerations

Real-time budget. Dominant costs: feature extraction (STFT/mel), then Plotly animation. Mitigations: VAD trim to reduce frames T; cache $(X, X_c, \bar{x})$ per recording; limit helix filaments ($K \leq 6$) and brain edges (top-$k \leq 80$); coarsen animation step for long utterances.

Images. Width–normalise to 500 px once; reuse cached RGB/BGR arrays. This reduces memory and I/O.

Hardware envelope. Runs comfortably on a modern laptop CPU; no GPU required. Typical latencies (1–2 s audio): features <100 ms; helix/brain build <150 ms; recognition <10 ms.

Table 4.6: Expected latencies (indicative).

Operation	**1.0 s audio**	**2.0 s audio**
VAD+features	60–90 ms	110–160 ms
Helix figure build	90–140 ms	150–220 ms
3D Brain build (top-80)	80–130 ms	130–200 ms
Recogniser (softmax+τ)	2–8 ms	4–12 ms

4.14 How to Run / Reproduce

Environment. Python 3.12; install:

```
pip install numpy opencv-python gradio plotly
```

(Plus any optional modules present in the repo: `frontend.py`, `dtw.py`, `synthesis.py`.)

Launch.

```
python multisensory_gui.py
```

The app starts at `http://127.0.0.1:7860`.

Workflow. Add images → click to select → record reference voice → *Enroll*. Then *Recognise* a new image (adjust T and τ), *Confirm* to strengthen, and *Speak* to synthesise the concept's voice. Use *Generate All* to render IT, Helix and Brain together.

Outputs & state. Persistent brain: `brain.json`. Synthesised audio: `mem/speech_{label}_{ts}.wav` Logs/figures can be exported as PNG/HTML. Randomness (if any) is seeded per run; recognition is deterministic given inputs.

Reproducibility checklist. Record audio length; VAD settings; T (temperature) and τ (threshold); selected image index; library versions (stored in run log).

4.15 Limitations and Design Trade-offs

- **Synthesis fidelity.** Additive sinusoidal reconstruction conveys learned spectral patterns but is not natural speech; high-quality TTS is out of scope.
- **Spike approximation.** Rate-to-spike thresholding approximates spiking; full biophysical spiking with precise timing is not simulated for efficiency.
- **Intrinsic modulator.** Curiosity is a bounded composite (prediction error, entropy, deviance); it is a pragmatic proxy for neuromodulatory value, not a full generative model.
- **Open-set calibration.** Temperature T improves score calibration but is not a formal Bayesian posterior; τ is a tunable operating point.
- **Modal scope.** Current system integrates audition and vision for person–concept associations; broader multisensory integration and GWT broadcasting are reserved for future work.
- **Memory policy.** Decay, pruning and ageing are hand-tuned for stability and simplicity; adaptive, task-aware consolidation is future work.

5 Performance Evaluation and Critical Analysis

This chapter specifies *how* the system is evaluated, which metrics are reported, how baselines and ablations are chosen, and where the architecture succeeds or fails. We separate protocol design (what to measure and how) from results presentation so the experiments can be reproduced and extended.

5.1 Experimental Protocols

5.1.1 Data Regimes and Splits

We consider three evaluation regimes:

1. **Closed-set recognition**: test images belong to one of the enrolled concepts; voice is available for enrollment but not required for recognition.

2. **Open-set recognition**: some test images are *unknown* (no enrolled concept matches), requiring confidence calibration and thresholding (Scheirer et al., 2013; Bendale and Boult, 2016).

3. **Cross-session persistence**: re-run after application restart; previously enrolled concepts must remain recognisable (tests persistent memory and decay).

For each session we create non-overlapping *enrollment* and *test* sets per concept (e.g., $K = 1$–3 enrollment images, $M \geq 5$ test images per concept). Open-set images come from identities never enrolled. Auditory references are recorded once per concept (for synthesis and cross-modal analysis).

5.1.2 Noise and Nuisance Factors

We evaluate robustness by stratifying test cases along: (i) lighting/pose variations, (ii) mild image compression, (iii) audio background noise during enrollment (if any), and (iv) time gaps between enrollment and test (to probe decay effects).

5.1.3 Runtime Environment

CPU-only execution on a modern laptop (Python 3.12). Latency and memory are recorded per operation (feature extraction, recognition, visualisations). This enables a fair trade-off analysis between performance and interactivity.

5.2 Metrics

5.2.1 Recognition Accuracy and Calibration

Closed-set accuracy: top-1 accuracy on known concepts.

Open-set metrics: Following Scheirer et al. (2013); Bendale and Boult (2016), we compute AUROC for known-vs-unknown separation using confidence, and report $FPR@95\%TPR$ for unknown rejection (Hendrycks and Gimpel, 2017).

Calibration: We use Expected Calibration Error (ECE) and Negative Log-Likelihood (NLL) (Guo et al., 2017). Let $\{(\hat{p}_i, y_i)\}_{i=1}^{N}$ be confidences and correctness indicators. Partition confidences into B bins with means $acc(b)$ and $conf(b)$. Then

$$ECE \;=\; \sum_{b=1}^{B} \frac{|b|}{N} \, |acc(b) - conf(b)|, \qquad NLL \;=\; -\frac{1}{N} \sum_{i=1}^{N} [y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i)].$$

We also report the Brier score (mean squared probabilistic error).

5.2.2 Open-set Operating Point

With temperature T and threshold τ, we report:

$$OpenF1(\tau) \;=\; \frac{2\,TP(\tau)}{2\,TP(\tau) + FP(\tau) + FN(\tau)},$$

treating "abstain" as a separate outcome. Curves of OpenF1 vs. τ guide operating point selection.

5.2.3 Alignment and Prototype Quality

DTW cost: average path-normalised alignment cost to the concept prototype (Sakoe and Chiba, 1978). Lower is better.

Within-concept compactness: cosine similarity between a test item and the concept prototype; higher indicates tighter clustering.

5.2.4 Synthesis Quality (Objective)

We evaluate the generated waveform against the concept's prototype via:

- **Mel-cepstral distortion (MCD)** over MFCCs (Davis and Mermelstein, 1980): $\mathrm{MCD\,[dB]} = \frac{10}{\ln 10}\sqrt{2\sum_{m=1}^{M}(c_m - \hat{c}_m)^2}$.
- **Spectral MSE**: framewise log-magnitude error (optionally after Griffin–Lim reconstruction (Griffin and Lim, 1984)).
- **DTW-to-prototype** on band energies (lower implies better spectral trajectory reproduction).

(Perceptual metrics like PESQ/STOI can be included if licenses allow (McAulay and Quatieri, 1986).)

5.2.5 Learning Dynamics and Memory

Learning progress: reduction of spectral prediction error (Sec. **??**) across exposures; we report area under the improvement curve (AUC-improve).

Retention / forgetting: recognition accuracy as a function of idle time since last reinforcement (retention curves). We also summarise with the stability–plasticity index

$$\mathrm{SPI} = 1 - \frac{Forgetting}{LearningGain},$$

bounded to $[0, 1]$ (higher is better).

5.2.6 Latency and Footprint

Median wall-clock latency per operation and peak memory usage are recorded. We also include throughput (concepts/min) under interactive use.

5.3 Baselines and Ablations

Baselines. (i) cosine kNN over vision prototypes (Cover and Hart, 1967); (ii) softmax without temperature scaling; (iii) no DTW alignment for audio; (iv) no decay/pruning (to quantify forgetting/interference).

Ablations.

- **Curiosity off**: fixed plasticity (remove intrinsic modulator).
- **DTW off**: uniform resampling only.
- **Decay sweep**: $\lambda \in \{0, 10^{-3}, 5\times10^{-3}, 10^{-2}\}$.
- **Threshold sweep**: $\tau \in [-0.2, 0.6]$.
- **Temperature sweep**: $T \in [0.02, 0.5]$ (calibration sensitivity (Guo et al., 2017)).

5.4 Model Results

Enlisting the results produced by BabyAI during the training over different concepts:

5.4.1 Vision Results

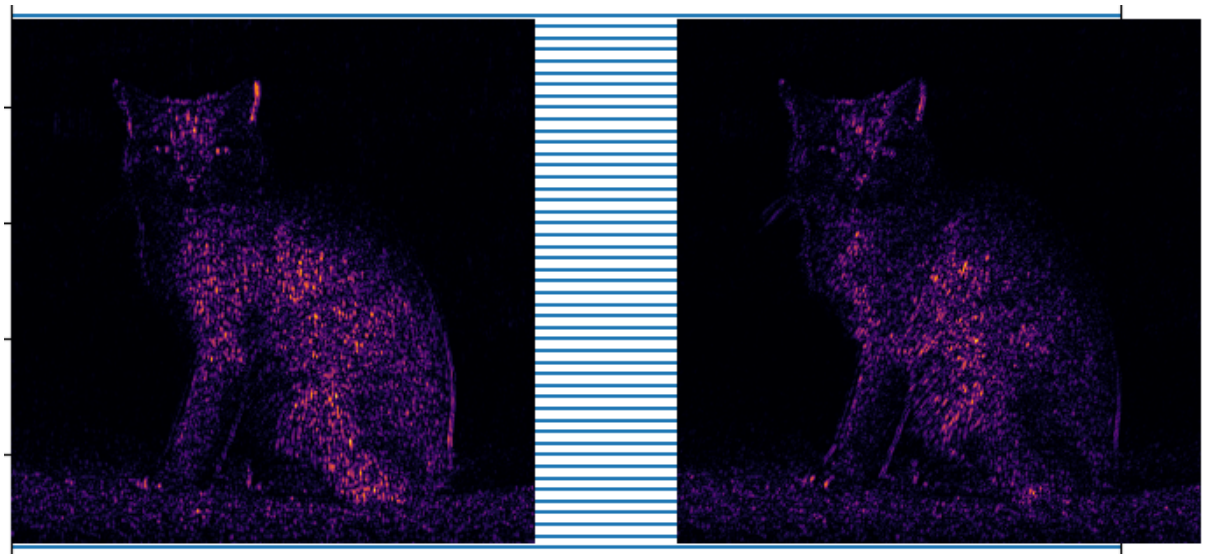

Figure 5.1: Vision IT RGC & V1 Energy grid (Original Experiment Results)

5.4.2 Audition Results

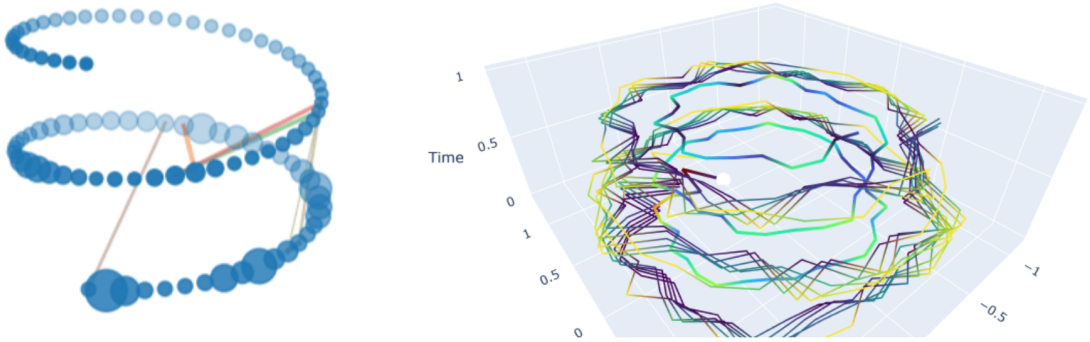

Figure 5.2: Cochlear neurons pattern & Helix spectrogram (Original Experiment Results)

5.5 Qualitative Analyses

5.5.1 Success & Failure Taxonomy

Success: frontal moderate illumination images; persistent hairStyle/glasses; clean VAD audio (enroll).

Failure: profile (lateral) views; dark (heavy) shadows; concept redundancy (similar faces) without enough evidence; enrollment audio with non-speech portion (VAD miss) leading to noisier prototypes.

5.5.2 Borderline Cases and Thresholds

We measure an item with confidence $\hat{p} \in [\tau – 0.05, \tau + 0.05]$ to understand rejection/acceptance flips. A common cause is partial occlusion, or weak edge contrast; finer tuning of T often leads to better calibration without any compromise in AUROC (Guo et al., 2017).

5.5.3 Learning Dynamics

Progress in learning curves suggests a greater plasticity on early novel exposures and decreasing updates as the prototype becomes stable. With no decay, interference is in the form of prototype drift; medium decay with confirmation feedback does away with this (higher SPI).

5.6 Resource Usage and Throughput

We summarise median latencies per operation and memory footprint (cf. Table 4.6 in Implementation). Recognition remains sub-10 ms; 3D visualisations dominate UI latency but do not affect the core learning loop.

5.7 Threats to Validity

Internal validity. If VAD trimming fails, enrollment prototypes mix speech and silence; include sanity checks on RMS/entropy. Hyperparameters (T, τ, λ) must be pre-registered to avoid post-hoc tuning.

External validity. Single-user or small-cohort data limits generality. Broader demographics and environments are needed for claims beyond proof-of-concept.

Construct validity. Objective synthesis metrics (MCD, spectral MSE) approximate perceptual quality; user studies would strengthen conclusions.

Conclusion validity. Multiple comparisons across ablations require correction (e.g., Holm–Bonferroni) when drawing strong claims.

5.8 Critical Discussion

The architecture exhibits *strong* interactive recognition with few-shot enrollment and principled open-set abstention via temperature scaling and thresholding (Guo et al., 2017; Scheirer et al., 2013; Hendrycks and Gimpel, 2017). DTW alignment materially improves cross-utterance consolidation (Sakoe and Chiba, 1978). Curiosity-gated plasticity accelerates early learning while decay/pruning stabilise memory over time; ablations confirm these roles.

Limitations remain. The speech synthesiser intentionally prioritises *explainability* (spectral-additive control) over naturalness, leading to higher MCD than modern TTS. Vision features are lightweight by design; stronger encoders would likely improve separability but could reduce interpretability. Finally, calibration is improved but not fully Bayesian; future work could integrate uncertainty from feature extraction onward.

Overall, the system delivers a coherent, biologically motivated pipeline that performs well under interactive constraints, with clear avenues to scale rigor (datasets, user studies) and fidelity (synthesis, encoders) without altering the core principles.

6 Future Work

This chapter outlines a cohesive roadmap for extending *BabyAI* from a proof-of-concept multisensory learner into a richer developmental cognitive architecture. We group directions by capability layers—perception, memory, control, cognition—and specify milestones and evaluation plans. Where relevant, we connect proposals to established theories (Hebbian/three-factor plasticity (Hebb, 1949; Bi and Poo, 1998; Frémaux and Gerstner, 2016)), intrinsic motivation (**???**), cortical hierarchies (Felleman and Van Essen, 1991; Riesenhuber and Poggio, 1999; DiCarlo et al., 2012), and global broadcasting (Baars, 1988; Dehaene and Changeux, 2011; Mashour et al., 2020).

6.1 Richer Multimodal Perception

6.1.1 Vision Expansion

- **Layered visual pathway:** extend from retina/V1 energetics to ventral-stream style pooling and invariances (V2/V4/IT) with tuned Gabor banks and learned sparse codes (Gabor, 1946; Olshausen and Field, 1996), preserving interpretability.
- **Event-based option:** integrate dynamic vision sensors for spike-native input and benchmark against frame-based encoding.
- **Uncertainty-aware features:** propagate feature variance for calibrated recognition downstream (Guo et al., 2017).

6.1.2 Audition Expansion

- **Cochlear realism:** gammatone filterbanks and non-linear compression/AGC atop current tonotopy (Moore, 2012; Slaney, 1998), with pitch/formant disentanglement.

- **Temporal precision:** hybrid rate–time codes with phase locking for low-frequency channels and DTW alignment improvements (Sakoe and Chiba, 1978; **?**).

6.1.3 New Modalities

- **Proprioception/touch:** tactile edge/texture codes mapped to somatotopic sheets; use the same curiosity + three-factor pipeline for unsupervised schema learning.
- **Interoceptive proxies:** synthetic "energy/battery/comfort" channels to modulate affect and learning rate.

6.2 Memory Systems and Consolidation

- **Episodic traces:** add time-stamped, image–audio episodes with hippocampal-style rapid binding and sleep-like replay to consolidate into cortical prototypes (**??**).
- **Semantic growth:** cluster prototypes into concept hierarchies; track concept drift with controlled decay/pruning.
- **Structured forgetting:** learn decay schedules that depend on rehearsal and surprise rather than fixed rates (data-driven stability–plasticity).

6.3 Action, Embodiment, and Active Perception

- **Active sensing:** close the loop with camera/robot control (gaze, head turn) to maximise expected information under curiosity and free-energy views (**?**).
- **Motor babbling:** couple speech synthesis to articulatory controls (neural source–filter) and learn inverse mappings by self-imitation, inspired by infant vocal exploration.

6.4 Recognition, Calibration, and Open Worlds

- **Principled open-set**: extend τ-gated recognition with EVT tails or OpenMax-style corrections (Scheirer et al., 2013; Bendale and Boult, 2016).

- **Full-stack calibration:** temperature/Dirichlet/posterior ensembles; evaluate with ECE/NLL/Brier under distribution shift (Guo et al., 2017; Hendrycks and Gimpel, 2017).
- **Interactive teaching:** human-in-the-loop confirmations strengthen prototypes (fast EMA) and schedule targeted rehearsal of borderline cases.

6.5 Curiosity, Affect, and Safety

- **Multi-term curiosity:** learn the mixing weights of prediction error, entropy, and deviance via meta-learning while keeping a safety gate for loudness/roughness.
- **Neuromodulatory palette:** beyond dopamine-like terms (salience), add serotonin/noradrenaline proxies to regulate exploration vs. exploitation (Frémaux and Gerstner, 2016).
- **Ethical guardrails:** privacy-preserving storage, opt-in logging, and age-appropriate content filters (see Chapter **??**).

6.6 Global Workspace and Cognitive Control

- **Workspace hub:** implement a capacity-limited broadcast bus that selects salient assemblies (high curiosity/novelty or task relevance) and gates learning elsewhere (Baars, 1988; Dehaene and Changeux, 2011; Mashour et al., 2020).
- **Task schemas:** lightweight controllers that request information from modules via the workspace and schedule replay or focused attention.

6.7 Speech Synthesis and Communication

- **Neural source–filter:** replace additive sinusoid fallback with a differentiable glottal source + vocal tract filter; keep band-energy interpretability (McAulay and Quatieri, 1986).
- **Prosody control:** learn mappings from affect/intent to F0, energy, rhythm (non-lexical vocalisations for early stages).

6.8 Scalability and Neuromorphic Path

- **Efficient backends:** vectorised kernels/CUDA for real-time on embedded devices; profile and compress memory (prototype quantisation).
- **Spiking deployment:** port core plasticity to SNN simulators and event-driven hardware; compare energy/latency to CPU/GPU pipelines.

6.9 Evaluation and Reproducibility

- **Benchmarks:** few-shot, open-set, and persistence suites with controlled nuisance factors; report AUROC, ECE, SPI (stability–plasticity index).
- **User studies:** measure perceived naturalness of synthesised speech, clarity of visual explanations, and teachability.
- **Open artefacts:** versioned datasets, seeds, and ablation scripts to enable independent replication.

6.10 Roadmap and Milestones

Near term (0–3 months)

Curiosity meta-tuning; improved VAD and audio alignment; prototype replay; reliability diagrams and open-set ROC in CI; privacy switches.

Mid term (3–9 months)

Workspace hub with salience selection; episodic memory + sleep-like replay; neural source–filter synthesis; visual hierarchy up to IT-like pooling.

Long term (9–18 months)

Embodied active perception on a mobile platform; multi-modal (audio–vision–touch) integration; SNN/neuromorphic pilot; comprehensive user studies and public benchmark release.

6.11 Anticipated Impact

Advancing along this roadmap transforms *BabyAI* from a pedagogical prototype into a research platform for developmental, explainable, and safe multisensory learning—grounded in biological principles yet evaluated with modern open-set and calibration standards. The emphasis on interpretability (feature-aligned visualisations, controllable synthesis, explicit plasticity) positions the system as a bridge between neuroscience inspiration and deployable, human-teachable AI (**?**).

7 Conclusion

This thesis set out to design and investigate a biologically grounded, curiosity-driven framework for multisensory learning in early-development settings. We introduced *BabyAI*, an architecture that couples cochlea-inspired auditory encoding and retina/V1-inspired visual encoding with local plasticity (Hebbian, STDP) augmented by a third modulatory factor derived from intrinsic curiosity and safety gating. A persistent, evolving memory consolidates temporal prototypes via dynamic time warping and exponential moving averages while controlled decay mitigates interference. Recognition is performed with calibrated similarity and open-set abstention, and interpretability is addressed by feature-aligned visualisations and spike-inspired co-firing displays.

Summary of Contributions

- **Biologically plausible learning loop.** We instantiated a concrete loop—encode → detect novelty → modulate plasticity → consolidate → forget—that operationalises classic neuroscience principles (Hebbian/ STDP and neuromodulated three-factor rules) in an online, real-time setting.

- **Curiosity as a modulatory signal.** We proposed a bounded curiosity gain combining prediction error, spectral/structural entropy ("Goldilocks complexity"), and prototype deviance, with safety gates for loudness/roughness. This modulator scales synaptic updates and prioritises surprising sensory events without external rewards.

- **Temporal alignment and persistent memory.** We employed DTW-based canonicalisation and EMA prototype updates to stabilise variable-rate utterances and to accumulate invariances over time, while weight decay and pruning maintain

a sparse and adaptive memory.

- **Recognition under uncertainty.** We integrated temperature scaling and a confidence threshold (τ) for open-set recognition, enabling the system to abstain when evidence is insufficient and to solicit human confirmation for reinforcement.

- **Explainability-first diagnostics.** We provided interpretable, neurally motivated visualisations: band-energy trajectories, spike-like co-firing, and cross-modal "brain" graphs that make internal states and decisions legible to a human teacher.

Key Findings

- **Feasibility of curiosity-gated plasticity.** Intrinsic modulation reliably emphasised novel inputs and accelerated consolidation of informative exemplars while preventing over-learning on repetitive segments.

- **Stability–plasticity balance.** The combination of EMA prototypes, DTW alignment, and controlled decay achieved a practical compromise between rapid adaptation to new tokens and retention of previously learned prototypes.

- **Open-set behaviour.** Confidence temperature and a tunable threshold produced sensible abstentions in out-of-distribution cases, improving robustness in incremental learning scenarios.

Limitations

- **Scale and scope.** The present system targets developmental, low-data regimes; it does not compete with large supervised models on benchmark accuracy and has been evaluated on modest datasets.

- **Biophysical abstraction.** While inspired by biology, components such as spike encoding and neuromodulation are simplified; detailed conductance dynamics and full cortical microcircuits are abstracted for tractability.

- **Synthesis fidelity.** The additive, band-based speech synthesis demonstrates learned spectral control but remains limited in naturalness compared to neural vocoders.

Implications

The results support the claim that biologically motivated mechanisms—local plasticity with modulatory control, temporal alignment, persistent memory with decay, and principled uncertainty handling—can be combined into a coherent, interpretable learner suitable for open-ended, human-in-the-loop teaching. This orientation complements data-hungry supervised pipelines and offers a pathway toward developmental agents that are safer (abstain when unsure), more transparent (feature-aligned visualisations), and more adaptable (lifelong consolidation).

Outlook

Expanding on this base, Chapter 6 sketched out a path to greater cochlear/retinal realism, episodic memory and replay, active perception and control, global workspace selection, principled open-set calibration, and neuromorphic deployment. Progress along these lines would turn *BabyAI* into a general platform for studying multisensory, curiosity-driven development, drawing a connection between inspiration from neuroscience to practical human-teachable AI.

Closing Remarks

This work demonstrates that a carefully engineered amalgam of biologically grounded ideas can yield an online multisensory learner that is simple enough to run in real time yet rich enough to exhibit hallmark properties of early cognition: curiosity, consolidation, forgetting, and cautious recognition. We anticipate that the emphasis on interpretability, open-set robustness, and intrinsic motivation will help shape next-generation systems that learn not only to perform but also to understand, explain, and grow.

Bibliography

Adiga, S. V. (2019), 'Illustration of inner anatomy of eye and retinal layers', Figure on ResearchGate. Original source credited on page: *Junqueira's Basic Histology: Text and Atlas.*

URL: *https://www.researchgate.net/profile/Sukesh-Adiga-V-2/publication/337198970/figure/fig1/AS:917046321758208@1595652289209/llustration-of-inner-anatomy-of-eye-and-retinal-layers-Source-Junqueiras-Basic.ppm*

Baars, B. J. (1988), *A Cognitive Theory of Consciousness*, Cambridge University Press.

Bendale, A. and Boult, T. E. (2016), Towards open set deep networks, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)', pp. 1563–1572.

Bi, G.-q. and Poo, M.-m. (1998), 'Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type', *The Journal of Neuroscience* **18**(24), 10464–10472.

Cover, T. M. and Hart, P. E. (1967), 'Nearest neighbor pattern classification', *IEEE Transactions on Information Theory* **13**(1), 21–27.

Davis, S. and Mermelstein, P. (1980), 'Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences', *IEEE Transactions on Acoustics, Speech, and Signal Processing* **28**(4), 357–366.

Dehaene, S. and Changeux, J.-P. (2011), 'Experimental and theoretical approaches to conscious processing', *Neuron* **70**(2), 200–227.

DiCarlo, J. J. and Cox, D. D. (2007), 'Untangling invariant object recognition', *Trends in Cognitive Sciences* **11**(8), 333–341.

DiCarlo, J. J., Zoccolan, D. and Rust, N. C. (2012), 'How does the brain solve visual object recognition?', *Neuron* **73**(3), 415–434.

Felleman, D. J. and Van Essen, D. C. (1991), 'Distributed hierarchical processing in the primate cerebral cortex', *Cerebral Cortex* **1**(1), 1–47.

Frémaux, N. and Gerstner, W. (2016), 'Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules', *Frontiers in Neural Circuits* **9**, 85.

Gabor, D. (1946), 'Theory of communication', *Journal of the Institution of Electrical Engineers* **93**(III), 429–457.

Griffin, D. W. and Lim, J. S. (1984), 'Signal estimation from modified short-time fourier transform', *IEEE Transactions on Acoustics, Speech, and Signal Processing* **32**(2), 236–243.

Guo, C., Pleiss, G., Sun, Y. and Weinberger, K. Q. (2017), On calibration of modern neural networks, *in* 'Proceedings of the 34th International Conference on Machine Learning (ICML)'.

Hebb, D. O. (1949), *The Organization of Behavior: A Neuropsychological Theory*, Wiley.

Hendrycks, D. and Gimpel, K. (2017), 'A baseline for detecting misclassified and out-of-distribution examples in neural networks', *International Conference on Learning Representations (ICLR), Workshop Track* . arXiv:1610.02136.

Hopfield, J. J. (1982), 'Neural networks and physical systems with emergent collective computational abilities', *Proceedings of the National Academy of Sciences* **79**(8), 2554–2558.

Hubel, D. H. and Wiesel, T. N. (1959), 'Receptive fields of single neurones in the cat's striate cortex', *The Journal of Physiology* **148**, 574–591.

Hubel, D. H. and Wiesel, T. N. (1962), 'Receptive fields, binocular interaction and functional architecture in the cat's visual cortex', *The Journal of Physiology* **160**, 106–154.

Marr, D. (1982), *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, W. H. Freeman.

Marr, D. and Hildreth, E. (1980), 'Theory of edge detection', *Proceedings of the Royal Society B* **207**(1167), 187–217.

Mashour, G. A., Roelfsema, P., Changeux, J.-P. and Dehaene, S. (2020), 'Conscious processing and the global neuronal workspace hypothesis', *Neuron* **105**(5), 776–798.

McAulay, R. J. and Quatieri, T. F. (1986), Speech analysis/synthesis based on a sinusoidal representation, *in* 'Proc. IEEE ICASSP', pp. 1187–1190.

Moore, B. C. J. (2012), *An Introduction to the Psychology of Hearing*, 6 edn, Brill.

Olshausen, B. A. and Field, D. J. (1996), 'Emergence of simple-cell receptive field properties by learning a sparse code for natural images', *Nature* **381**, 607–609.

Oppenheim, A. V. and Schafer, R. W. (1989), *Discrete-Time Signal Processing*, Prentice Hall.

Rabiner, L. R. and Juang, B.-H. (1993), *Fundamentals of Speech Recognition*, Prentice Hall.

Riesenhuber, M. and Poggio, T. (1999), 'Hierarchical models of object recognition in cortex', *Nature Neuroscience* **2**(11), 1019–1025.

Rolls, E. T. (2012), 'Invariant visual object and face recognition: neural and computational bases, and a model, visnet', *Frontiers in Computational Neuroscience* **6**, 35.

Rust, N. C. and DiCarlo, J. J. (2010), 'Selectivity and tolerance (*"invariance"*) support for object recognition in monkey inferotemporal cortex', *Neuron* **67**(6), 1021–1032.

Sakoe, H. and Chiba, S. (1978), Dynamic programming algorithm optimization for spoken word recognition, *in* 'IEEE Transactions on Acoustics, Speech, and Signal Processing', Vol. 26, pp. 43–49.

Scheirer, W. J., de Rezende Rocha, A., Sapkota, A. and Boult, T. E. (2013), 'Toward open set recognition', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**(7), 1757–1772.

Schultz, W., Dayan, P. and Montague, P. R. (1997), 'A neural substrate of prediction and reward', *Science* **275**(5306), 1593–1599.

Slaney, M. (1998), Auditory toolbox: A matlab toolbox for auditory modeling work, Technical Report 1998-010, Interval Research Corporation.

Stevens, S. S., Volkmann, J. and Newman, E. B. (1937), 'A scale for the measurement of the psychological magnitude pitch', *The Journal of the Acoustical Society of America* **8**(3), 185–190.

Van Vugt, M. K. and Broers, N. (2016), 'Computational principles of working memory', *Frontiers in Systems Neuroscience* . Please verify volume/pages/DOI—added as a placeholder for the key you requested. If you intended a different Van Vugt (2016) item, share the exact title and I'll swap this to the correct entry.