



POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Speech Emotion Recognition in the Real World: A Modular Pipeline for Real-Time Deployment and Emotion-Aware Interaction

TESI DI LAUREA MAGISTRALE IN
COMPUTER SCIENCE AND ENGINEERING
INGEGNERIA INFORMATICA

Author: **Pietro Morcavallo**

Student ID: 249070

Advisor: Prof. Alessandro Campi

Co-advisors: Ing. Michele Di Dalmazi

Academic Year: 2025-26

Abstract

Speech Emotion Recognition (SER) has advanced rapidly, yet the field’s most capable models seldom reach the operational settings that motivate them. Production deployments routinely collapse rich model outputs to a single dominant label, bind the temporal behaviour of the system to one specific model, and leave operators with little insight into how uncertain a prediction is. This thesis presents the design, implementation, and evaluation of a modular, real-time SER platform for call-centre monitoring that preserves the full probabilistic richness of model outputs, from inference through to the operator-facing visualisation. The system is built on five contributions. A hexagonal architecture with a dependency-injection container factorises three interchangeable SER backends, a BiLSTM and two Whisper-based checkpoints (categorical and dimensional), behind a uniform port interface, enabling session-level backend selection without any change to the surrounding pipeline. A probabilistic temporal aggregation algorithm turns the raw sliding-window event stream into stable, regularly spaced frames and transcript-enriched utterances while preserving the posterior distribution at every granularity. A distances-to-probabilities inversion grounds the dimensional backend in Russell’s circumplex geometry, mapping continuous valence–arousal estimates to categorical distributions and making the dimensional model a structurally equal citizen of the pipeline. A set of empirically grounded operating points fixes the configuration of the system, and an interpretability-first visualisation layer renders the preserved distributions as a probability-weighted circumplex, an interpolated-colour emotional timeline, and an uncertainty-annotated transcript. The evaluation, built on a synthetic geometric suite and four real Italian call-centre sessions, confirms the intended geometric properties of the conversion exactly, shows that aggregation reduces inter-frame jitter by about 70% without collapsing the distributions, and demonstrates the full pipeline running in real time (real-time factor ≤ 0.75) on a commodity laptop. Together these results show that distributional, interpretability-first SER is deployable today on consumer hardware.

Keywords: speech emotion recognition; real-time streaming; probabilistic representation; uncertainty; circumplex model; modular architecture; interpretability

Abstract in lingua italiana

Il riconoscimento delle emozioni dal parlato (Speech Emotion Recognition, SER) ha conosciuto rapidi progressi, ma i modelli più capaci raramente raggiungono i contesti operativi che ne motivano lo sviluppo. I sistemi in produzione tipicamente riducono le ricche uscite dei modelli a una singola etichetta dominante, vincolano il comportamento temporale del sistema a uno specifico modello e offrono agli operatori scarsa visibilità sul grado di incertezza di ciascuna predizione. Questa tesi presenta la progettazione, l'implementazione e la valutazione di una piattaforma SER modulare e in tempo reale per il monitoraggio di call center, che preserva l'intera ricchezza probabilistica delle uscite dei modelli, dall'inferenza fino alla visualizzazione destinata all'operatore. Il sistema si fonda su cinque contributi. Un'architettura esagonale con un container di dependency injection fattorizza tre backend SER intercambiabili, una BiLSTM e due checkpoint basati su Whisper (categorico e dimensionale), dietro un'interfaccia di porta uniforme, consentendo la selezione del backend a livello di sessione senza alcuna modifica al resto della pipeline. Un algoritmo di aggregazione temporale probabilistica trasforma il flusso grezzo di eventi a finestra scorrevole in frame stabili e regolarmente spaziate e in enunciati arricchiti dalla trascrizione, preservando la distribuzione a posteriori a ogni livello di granularità. Un'inversione da distanze a probabilità ancora il backend dimensionale alla geometria del circumplesso di Russell, mappando le stime continue di valenza e attivazione in distribuzioni categoriche e rendendo il modello dimensionale un cittadino strutturalmente paritario della pipeline. Un insieme di punti operativi fondati empiricamente fissa la configurazione del sistema, e un livello di visualizzazione orientato all'interpretabilità rende le distribuzioni preservate come un circumplesso pesato per probabilità, una linea temporale emotiva a colori interpolati e una trascrizione annotata con l'incertezza. La valutazione, basata su una suite geometrica sintetica e su quattro sessioni reali di call center in lingua italiana, conferma esattamente le proprietà geometriche attese della conversione, mostra che l'aggregazione riduce il jitter tra frame di circa il 70% senza far collassare le distribuzioni e dimostra l'esecuzione dell'intera pipeline in tempo reale (fattore di tempo reale ≤ 0.75) su un laptop di fascia consumer. Nel complesso, questi risultati mostrano che un SER distribuzionale e orientato all'interpretabilità è oggi

implementabile su hardware di consumo.

Parole chiave: riconoscimento delle emozioni dal parlato; streaming in tempo reale; rappresentazione probabilistica; incertezza; modello del circumplesso; architettura modulare; interpretabilità

Contents

Abstract	i
Abstract in lingua italiana	iii
Contents	v
1 Introduction	1
1.1 Motivation and Context	1
1.2 Thesis Objective	3
1.3 Summary of Contributions	4
1.4 Thesis Organisation	6
2 Background and Related Work	9
2.1 Speech Emotion Recognition: Foundations	9
2.1.1 From Hand-Crafted Features to Deep Learning	9
2.1.2 Categorical and Dimensional Emotion Models	10
2.1.3 Evaluation Metrics	10
2.1.4 Cross-Corpus Degradation	11
2.2 Foundation Models for Speech Emotion Recognition	11
2.2.1 Self-Supervised Speech Representations	11
2.2.2 Whisper as an Emotion Encoder	12
2.3 Dimensional SER: Valence/Arousal Inference	13
2.3.1 The Circumplex Model and Anchor Coordinates	13
2.3.2 The Inversion Problem: From Coordinates to Probabilities	14
2.3.3 Valence–Arousal Asymmetry	16
2.4 Uncertainty Quantification in Neural SER	16
2.4.1 Sources of Uncertainty in Affective Labels	16
2.4.2 Probability Calibration and Expected Calibration Error	17
2.4.3 Out-of-Distribution Detection	18

2.5	Real-Time and Streaming SER	19
2.5.1	Sliding-Window Inference and Temporal Aggregation	19
2.5.2	Deployment in Call-Centre Environments	20
2.6	Emotion Corpora	20
2.6.1	Acted Corpora	20
2.6.2	Naturalistic Corpora	21
2.6.3	Cross-Lingual Coverage and EmoBox	22
3	System Design and Architecture	23
3.1	Architectural Overview	23
3.1.1	Three-Layer Decomposition	23
3.1.2	Structural Dependency Enforcement	24
3.2	Domain Layer	24
3.2.1	Typed Event Model	24
3.2.2	Service Ports	25
3.2.3	SER Stream: Sliding Window and VAD Gating	26
3.2.4	Temporal Aggregation	26
3.3	Application Layer	27
3.3.1	Dependency Injection Container	28
3.3.2	Pipeline Factory	29
3.3.3	Use Cases	29
3.4	Infrastructure Layer: ML Adapters	30
3.4.1	Legacy BiLSTM Adapter	31
3.4.2	Whisper Categorical Adapter	31
3.4.3	Whisper Dimensional Adapter	32
3.4.4	Hume AI Adapter	33
3.4.5	Vosk STT Adapter and Session Isolation	33
3.5	HTTP and WebSocket API	34
3.5.1	FastAPI Application	34
3.5.2	WebSocket Streaming Endpoint	34
3.5.3	Batch Analysis Endpoint	35
3.6	React Frontend Architecture	35
3.6.1	Technology Stack and State Management	35
3.6.2	Page Structure and Component Hierarchy	36
3.6.3	Analysis Visualisations	37
3.6.4	Session Hooks and Lifecycle	38
3.7	Key Design Decisions	39

3.7.1	Structural Typing over Inheritance	39
3.7.2	PCM Broadcast Architecture	39
3.7.3	Aggregator Concurrency Model	40
3.7.4	The <code>safe_until</code> Guard	40
3.7.5	Session Isolation and Stateful Recognisers	40
3.7.6	Dimensional Inversion as a Unified Representation	40
4	Original Contributions	43
4.1	Probabilistic Emotion Representation	43
4.1.1	The Value of Full Posterior Preservation	43
4.1.2	End-to-End Distribution Preservation	44
4.1.3	Shannon Entropy as an Uncertainty Proxy	45
4.2	Weighted Aggregation Algorithm	45
4.2.1	The Aggregation Problem	45
4.2.2	Frame Aggregation via Uniform Window Averaging	46
4.2.3	Utterance Enrichment via Overlap-Weighted Averaging	46
4.3	Dimensional SER via Circumplex Inversion	47
4.3.1	Motivation: Bridging Continuous and Categorical Representations	48
4.3.2	The Distances-to-Probabilities Algorithm	49
4.4	Heuristic Design Decisions for Operational Robustness	50
4.4.1	Heuristic Design Strategies	51
4.5	Interpretability-First Visualisation	52
4.5.1	The Probability-Weighted Circumplex	52
4.5.2	The Interpolated-Colour Emotional Timeline	52
4.5.3	Transcript Panel with Per-Utterance Uncertainty Annotation	53
5	Experimental Evaluation	55
5.1	Experimental Setup	55
5.1.1	Evaluation Material	55
5.1.2	Evaluation Metrics	56
5.1.3	Software Configuration and Reproducibility	57
5.2	Geometric Validation of the Distances-to-Probabilities Algorithm	58
5.2.1	Anchor Concentration	58
5.2.2	Neutral Radial Behaviour	59
5.2.3	Bimodality at Anchor Midpoints	60
5.2.4	Full-Plane Coherence and Smoothness	61
5.2.5	Summary	62
5.3	Platform Demonstration: Interchangeable Backends in Real Time	62

5.3.1	Backend Interchangeability	63
5.3.2	Real-Time Performance	65
5.3.3	Streaming Cadence and Release Latency	65
5.4	Validation of the Probabilistic Temporal Aggregation	66
5.4.1	Jitter Reduction	67
5.4.2	Preservation of Distributional Structure	68
5.4.3	The Safe-Until Guard	69
5.4.4	Summary	70
5.5	Configuration Study: Recommended Operating Points	70
5.5.1	VAD Gating Threshold	71
5.5.2	Window Duration	72
5.5.3	Hop Duration	73
5.5.4	Conversion Parameters of the Dimensional Backend	73
5.5.5	Recommended Configuration	76
5.6	Fidelity of the Interpretability-First Visualisation	76
5.6.1	Circumplex Centroid Faithfulness	77
5.6.2	The Interpolated-Colour Timeline	78
5.6.3	Transcript Panel with Uncertainty Annotation	80
5.7	Discussion	81
5.8	Scope and Generalisation of the Results	83
6	Conclusions and Future Work	87
6.1	Summary of Contributions	87
6.2	Future Work	88
6.2.1	Near-Term Technical Consolidation	88
6.2.2	A Quantitative Call-Centre Analytics Instrument under Privacy Constraints	90
6.2.3	Evaluating Cold-Approach Communication: Three Application Do- mains	91
6.2.4	Toward Closing the Recognition Gap: Emotional TTS as a Catalyst	92
6.3	Closing Remarks	93
	Bibliography	94
	List of Acronyms	103

List of Figures	105
List of Tables	109
Acknowledgements	111

1 | Introduction

1.1. Motivation and Context

Spoken interaction has become the primary interface in a growing range of high-stakes services, from customer support and telemedicine to voice-based virtual assistants and mental health monitoring. In these contexts, understanding not only *what* users say, but also *how* they say it, is increasingly recognised as a critical capability for delivering adaptive and effective responses [19, 54]. Speech Emotion Recognition (SER), the automated inference of affective states from acoustic and prosodic cues, has been studied extensively over the past two decades, with applications spanning call-centre analytics, clinical psychology, e-learning, and human–robot interaction [1, 32].

The commercial landscape of contact-centre emotion analytics has converged on a single dominant architecture: sentiment inference performed on Automatic Speech Recognition (ASR) transcripts, not on the acoustic signal of the voice itself. Platforms such as Amazon Contact Lens, Google Cloud Contact Centre AI, and NICE Enlighten AI reduce emotional state to a coarse polarity label (Positive / Negative / Neutral) or to a scalar score derived from lexical content alone. NICE’s own product documentation states explicitly that its sentiment scores “*are not influenced by voice characteristics, such as tone, volume, and speed*” [44], a formulation that encapsulates the structural divergence between the commercial paradigm and the acoustic SER research tradition [16, 64]. Acoustic-native systems, most notably Cogito (acquired by Verint in October 2024) and CallMiner Eureka, constitute a market minority, yet even these deliver hard categorical labels or opaque composite scalars rather than calibrated probability distributions. The sole commercially deployed system to have exposed a genuinely probabilistic, multi-dimensional emotion output was Hume AI’s Expression Measurement API, which reported continuous scores across 48 acoustic-emotional dimensions grounded in the data-driven semantic space theory of emotion [15], derived from large-scale cross-cultural annotation studies. Its retirement, with new API jobs disabled from May 2026 and the service fully offline by June 2026 [27], leaves the commercial landscape without a single deployed system capable of producing the distributional, calibration-aware output that the research community has consistently

advocated [24, 46]. From the practitioner side, a large-scale survey of chief marketing officers and chief customer officers found that accurately identifying customer emotions during interactions is the primary challenge in AI-enabled customer care [26], reinforcing the alignment between the research community’s agenda and a concrete commercial need.

From a modelling perspective, recent advances in self-supervised speech representation learning have fundamentally changed the state of the art in SER. Models such as wav2vec 2.0 [3], HuBERT [25], WavLM [12], and Whisper [50] learn rich acoustic representations from large corpora of unlabelled or weakly labelled audio, which can then be fine-tuned for downstream emotion classification with comparatively small annotated datasets [45]. These foundation models achieve substantially higher accuracy than hand-crafted feature pipelines on standard benchmarks and learn richer, more transferable acoustic representations than earlier acoustic classifiers, whose categorical decision boundaries are tied to the specific recording conditions and annotation conventions of their training corpora.

At the same time, the community has recognised that categorical SER models generalise poorly outside their training distribution. Differences in acoustic conditions, speaking style, and annotation conventions introduce systematic distribution shifts that degrade performance [5, 39, 55], motivating the exploration of dimensional representations, in particular the valence–arousal model of Russell [51], as a complementary representation grounded in the geometry of affect rather than in corpus-specific acoustic statistics.

Despite these modelling advances, a persistent gap separates laboratory benchmarks from operational deployment. Most SER systems are designed for offline, segment-level processing; few address the engineering challenges of streaming audio at low latency [41]. Moreover, collapsing a rich posterior distribution to its argmax discards information critical for capturing ambiguous and mixed affective states [24, 46], and neural classifiers are systematically overconfident [23]. A recent systematic survey of 108 SER publications finds that approximately 17 per cent cite call-centre or customer-service applications as their primary motivation, yet rely on acted, controlled-environment corpora that do not reflect the acoustic conditions of real contact-centre audio [64]. Bridging this gap requires not only more accurate models but also a carefully engineered system architecture that exposes those models to practitioners in a usable form: a platform that handles streaming audio at low latency, preserves the full probabilistic richness of model outputs end-to-end, and presents those outputs to operators in a way that supports informed, interpretable decision-making.

The chapters that follow describe the design, implementation, and evaluation of an in-

egrated real-time system that makes state-of-the-art SER models accessible for practical deployment, preserving distributional information throughout the pipeline and rendering it as actionable intelligence for human operators.

1.2. Thesis Objective

The commercial and academic landscape surveyed in Section 1.1 establishes a clear gap: open, reusable systems that process live contact-centre audio through an acoustically grounded SER backend at low latency and expose calibrated, distribution-aware outputs to practitioners in a form suitable for operational use remain scarce. The market has converged on text-based sentiment analysis; the sole platform to have offered probabilistic acoustic-level emotion scoring has been retired; and the academic literature, despite producing capable models, stops short of the engineering infrastructure that deployment requires. The first concrete requirement this thesis must address is therefore the construction of that infrastructure: an open, end-to-end streaming SER pipeline that processes live audio in real time and can serve as a deployable platform rather than a research artefact.

A second requirement follows directly from the structure of SER model outputs. Every existing system, commercial and academic, reduces the posterior distribution produced at each inference step to its argmax, reporting a single hard label and discarding the secondary emotions, confidence gradients, and ambiguity that the distribution encodes [46, 58]. Neural classifiers are systematically overconfident outside their training distribution [23], making the discarded uncertainty information precisely the signal an operator needs to calibrate their trust in a prediction. The system must therefore preserve the full posterior distribution at every stage of inference, without collapsing it at any intermediate processing step.

This requirement for distribution preservation extends across the temporal dimension of a live session. Frame-level predictions must be combined into stable, utterance-level outputs; the standard approach of majority voting erases distributional structure in the process [41], producing a stream of static labels with no memory of how the emotional signal has evolved. The consequence is that the temporal arc of an interaction, whether the caller’s tone escalated, when frustration peaked, whether an apparently angry utterance reflects a sustained pattern or an isolated spike, remains entirely invisible. No existing contact-centre analytics system provides a longitudinal view of emotional dynamics within a call: operators are presented with a snapshot of the current predicted state rather than with instruments for reading the affective trajectory of the full interaction. The

system must therefore solve the aggregation problem in a way that preserves distributional information across time, and must render that information in a form legible to the operator in real time.

A final structural requirement concerns the heterogeneity of available SER backends and the need for the system to remain open to future models. Categorical classifiers produce distributions over incompatible label sets, seven categories in EmoDB, nine in MSP-Podcast, whose boundaries do not coincide [10]; dimensional regression models produce continuous valence–arousal coordinates [51] that are incommensurable with categorical outputs by construction. Conversions between the two families have been proposed in isolation [28, 30], but no common representation is routinely used to carry both within a single runtime pipeline. Beyond this representational gap, existing SER systems are architecturally monolithic: changing the backend requires rebuilding the pipeline, a rigidity that forecloses both the flexibility of runtime backend selection and the ability to integrate the new foundation-model checkpoints that the field continues to produce. The architecture must therefore abstract away backend heterogeneity behind a common interface, enabling any categorical or dimensional backend, present or future, to be integrated without modifying the surrounding system.

The objective of this thesis is to address these four requirements in combination through the design, implementation, and empirical evaluation of a modular, open-architecture real-time SER pipeline. The system must ingest live audio at low latency with voice-activity gating, route the signal through configurable interchangeable backends spanning both categorical and dimensional model families under a unified probabilistic representation, preserve full posterior distributions through every stage of inference and temporal aggregation, and expose the resulting signals through an operator-facing interface in which uncertainty, emotional trajectory, and temporal dynamics are directly readable.

1.3. Summary of Contributions

At its core, this thesis contributes an integrated, real-time SER platform whose defining principle is the preservation of the full posterior emotion distribution end-to-end, from inference through temporal aggregation to operator-facing display, across heterogeneous categorical and dimensional backends unified under a single probabilistic representation. The five contributions below develop this principle across the research, engineering, and design layers of the system. They map onto the four requirements of Section 1.2 as follows: the open streaming platform (C1) answers the infrastructure requirement; the distribution-preserving aggregation algorithm (C2) answers the twin requirements of pos-

terior preservation at inference and across time; the unified probabilistic representation (C3) answers the backend-heterogeneity requirement; and the interpretability-first visualisation (C5) answers the requirement that this temporal, distributional information be rendered legibly to the operator. The remaining contribution, the heuristic operating points (C4), is cross-cutting: rather than adding an architectural capability, it provides the controlled empirical characterisation and the recommended operating points that secure the reliable behaviour on which the other four depend.

C1. An open, modular real-time SER platform with interchangeable backends.

The thesis provides the deployable streaming infrastructure whose absence both the academic and commercial landscapes have demonstrated. The platform integrates audio ingestion, denoising via RNNoise [61], VAD-gated sliding-window SER inference, Vosk-based automatic speech recognition [2], and temporal aggregation within a single hexagonal-architecture system [14, 40] that processes the acoustic signal directly. Multiple SER backends, categorical and dimensional, locally deployed or cloud-based, are registered behind a common port interface and selectable at session time, so that new backends can be integrated without modifying any surrounding component. The analysis window length w_s and hop interval h_s are exposed as first-class session-level parameters, enabling systematic characterisation of how temporal configuration affects streaming behaviour. The Hume AI Expression Measurement API was retired in May 2026 and was therefore unavailable for the evaluation experiments; the adapter implementation is preserved in the codebase for reproducibility.

C2. A probabilistic temporal aggregation algorithm for distribution-preserving SER.

Rather than reducing model outputs to hard labels, the system propagates full posterior distributions through every pipeline stage. At the frame level, inference window outputs are retained as complete categorical distributions; at the utterance level, a weighted overlap-based fusion algorithm combines successive frame distributions into stable, multi-granularity event streams that preserve secondary emotions, confidence gradients, and temporal transitions end-to-end [41, 53]. This contribution directly addresses both the single-frame information loss produced by argmax collapse and the absence of a principled streaming aggregation strategy that maintains distributional structure across time.

C3. A unified probabilistic representation for categorical and dimensional backends via the distances-to-probabilities algorithm.

To resolve the representational incompatibility between categorical and dimensional model families, a third SER backend exploits the valence and arousal regression heads of a Whisper-large-v3 model fine-tuned on MSP-Podcast [10, 20, 37, 50]. A geometric *distances-*

to-probabilities inversion maps continuous valence–arousal coordinates to a categorical probability distribution by nearest-anchor weighting in the circumplex plane [30], with dedicated handling of the Neutral and Other categories. The contribution is the resulting unified representation: the conversion places dimensional backend outputs in the same probabilistic space as the categorical backends, enabling unified display and runtime switching without modifying the upstream pipeline or the downstream interface.

- C4. Heuristic design decisions for reliable operation.** The thesis documents and establishes through controlled experiments the engineering decisions required to make the pipeline reliable in operational conditions: the VAD gating threshold θ_{vad} , which controls the trade-off between jitter reduction and speech-region coverage; the window-duration selection strategy for the Whisper encoder, which governs the zero-padding behaviour of the encoder; the hop-duration characterisation, which relates temporal configuration to observed event rates; and the conversion parameters (τ, s) governing the dimensional backend. Empirical evaluation of these parameters establishes recommended operating points grounded in observed pipeline behaviour [39, 55].
- C5. An interpretability-first visualisation *design* for uncertainty and emotional trajectory.** The operator-facing tools for reading both instantaneous model uncertainty and the longitudinal emotional trajectory of a call, each rendered directly from the preserved posterior rather than from a reduced label. The Emotion Circumplex renders the probability-weighted centroid of all active emotion anchors with contribution lines proportional to probability mass, making distributional spread directly readable [30, 51]. The interpolated-colour emotional timeline maps each frame’s posterior onto a single perceptual channel: hue encodes the dominant affect and saturation its certainty, so that confident frames read as vivid colours and ambiguous ones desaturate toward grey. Laid out over the whole call, this renders the emotional trajectory and its uncertain stretches readable at a glance. The transcript panel couples each utterance to its overlap-weighted probability distribution with audio synchronisation, enabling retrospective inspection at utterance granularity. No visual element reduces the distribution to a dominant label at any display stage.

1.4. Thesis Organisation

The remainder of this thesis is organised as follows.

Chapter 2 reviews the scientific and engineering foundations of the work. It covers the

evolution of SER from hand-crafted feature pipelines to deep architectures, categorical and dimensional emotion models, self-supervised foundation models and their role as SER backbones, probability calibration and uncertainty quantification, real-time and streaming SER in call-centre environments, and the benchmark corpora used throughout the evaluation.

Chapter 3 describes the overall system design and architecture. It presents the hexagonal architecture adopted, the audio ingestion and preprocessing pipeline, the SER backends, the Vosk speech-to-text (STT) module, the temporal aggregation and inter-process communication (IPC) event bus, the API and WebSocket streaming protocol, and the React frontend architecture.

Chapter 4 details the original contributions of the thesis: the design of probabilistic emotion representation, the weighted aggregation algorithm, the dimensional SER backend and the *distances-to-probabilities* algorithm, the heuristic design decisions adopted to ensure operational robustness, and the interpretability-first visualisation framework.

Chapter 5 presents the experimental evaluation, which tests each of the five contributions in turn. It validates the *distances-to-probabilities* algorithm (C3) through a fully synthetic geometric test suite; demonstrates the modular platform (C1) by processing one session through three interchangeable backends in real time on commodity hardware; validates the probabilistic temporal aggregation (C2) on the evaluation corpus; consolidates the heuristic design decisions (C4) into empirically grounded operating points for the five configuration parameters; and verifies that the interpretability-first visualisation components (C5) are faithful renderings of the underlying probability distributions. It closes with a discussion synthesising the evidence by contribution and a delimitation of the scope and generalisation of the results.

Chapter 6 synthesises the contributions of the thesis, situates them with respect to the existing literature, and outlines directions for future work.

2 | Background and Related Work

This chapter surveys the scientific and engineering foundations on which the work rests. Section 2.1 reviews the evolution of SER from hand-crafted feature pipelines to modern deep architectures. Section 2.2 examines self-supervised speech encoders as universal backbones for downstream emotion tasks. Section 2.3 covers dimensional affect modelling and the coordinate-conversion framework underlying the dimensional inference module. Section 2.4 addresses uncertainty quantification and probability calibration for neural classifiers. Section 2.5 discusses real-time and streaming SER in contact-centre deployments. Section 2.6 describes the benchmark corpora used in the evaluation.

2.1. Speech Emotion Recognition: Foundations

2.1.1. From Hand-Crafted Features to Deep Learning

Automatic recognition of speaker affect from the acoustic signal has been studied since the early 2000s. The canonical pipeline of the first decade extracted a fixed-size descriptor per utterance, typically comprising *mel-frequency cepstral coefficients* (MFCCs), pitch contours, energy envelopes, voice-quality measures, and statistical functionals (mean, variance, range) computed over these low-level descriptors [1, 19]. These so-called *Interspeech Computational Paralinguistics Challenge* feature sets, curated jointly by Schuller and Batliner [54], became a de-facto standard for comparable benchmarking and still appear as baselines in contemporary work.

The arrival of deep learning fundamentally changed the field. Convolutional and recurrent architectures, in particular *Long Short-Term Memory* (LSTM) networks, replaced hand-crafted descriptors by learning task-relevant representations directly from the raw or lightly processed signal. Trigeorgis *et al.* [60] demonstrated that an end-to-end convolutional-recurrent model trained on raw waveforms could match or surpass well-engineered feature pipelines on standard benchmarks, questioning the need for domain-expert acoustic descriptors. The subsequent introduction of *self-supervised pre-training* on tens of thousands of hours of unlabelled speech, discussed in Section 2.2, pushed per-

formance further and shifted the field towards a fine-tuning paradigm [3, 25].

2.1.2. Categorical and Dimensional Emotion Models

Two representational frameworks coexist in affective computing.

Categorical models describe emotion as membership in a finite set of discrete classes. The classical set adopted in acted-speech research derives from Ekman’s *basic emotions* [18]: anger, disgust, fear, happiness, sadness, and surprise, though operational corpora vary in coverage and labelling granularity. Categorical outputs are natural for most downstream applications, such as routing, alerting, or quality scoring, because they map directly to actionable states.

Dimensional models embed emotion in a continuous low-dimensional space, most commonly the *valence–arousal* (V–A) plane proposed by Russell [51]. Valence encodes hedonic tone (negative $\leftrightarrow$ positive), and arousal encodes the level of activation (calm $\leftrightarrow$ excited). Scherer [52] argued that this representation is more faithful to the continuous, graded nature of human affect and avoids the artificial discretisation that obscures ambiguous or mixed emotional states. The seminal empirical work of Bänziger and Scherer [5] on intonation established acoustic correlates of V–A coordinates in natural speech, connecting the psychological framework to measurable signal properties.

The two paradigms are not mutually exclusive. Wagner *et al.* [62] showed that transformer-based fine-tuning dramatically reduces the persistent “valence gap”, the large gap between arousal and valence accuracy that has plagued dimensional SER for over a decade, suggesting that the two spaces are more bridgeable than previously assumed. This thesis exploits this duality: the dimensional module produces a (V, A) point that is then inverted to a categorical probability distribution (Section 2.3), combining the psychological richness of the continuous representation with the interpretability of named emotion labels.

2.1.3. Evaluation Metrics

Categorical SER performance is most commonly reported as *Unweighted Average Recall* (UAR), the mean per-class recall computed without regard for class frequency. UAR is preferred over overall accuracy on corpora where class distributions are highly skewed - a characteristic feature of naturalistic datasets in which *neutral* and *sadness* dominate [37]. *Weighted Average Recall* (WAR) and macro F_1 are reported alongside UAR for completeness.

Dimensional SER performance is evaluated via the *Concordance Correlation Coefficient* (CCC) introduced by Lin [34]:

$$\text{CCC}(x, y) = \frac{2 \rho \sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \quad (2.1)$$

where ρ is the Pearson correlation between predicted and ground-truth sequences, σ their standard deviations, and μ their means. Unlike pure correlation, CCC penalises systematic biases in mean prediction, making it sensitive to models that capture the trend of emotional variation without matching its absolute level, a common failure mode in valence prediction [62].

2.1.4. Cross-Corpus Degradation

A recurring empirical observation in SER is that models trained on one corpus generalise poorly to another, even when both corpora purportedly cover the same emotion categories. Schuller *et al.* [55] systematically documented this cross-corpus degradation and identified its principal causes: mismatch in speakers, recording conditions, emotional intensity, and annotation conventions. These causes are instances of the broader phenomenon of *dataset shift*, the systematic divergence between training and deployment distributions analysed in depth by Quiñonero-Candela *et al.* [49]. Devillers *et al.* [16, 17] documented the annotation and machine-learning challenges posed by real call-centre speech, showing that spontaneous emotions in this domain are expressed with subtler and more ambiguous prosodic contours than in scripted portrayals, a characteristic that substantially raises the difficulty of both labelling and automatic recognition.

These observations establish cross-corpus and domain mismatch as a well-known challenge in the field and motivate the choice, in this thesis, of realistic, conversational-length material that reflects the conditions a monitoring deployment actually observes (Chapter 5).

2.2. Foundation Models for Speech Emotion Recognition

2.2.1. Self-Supervised Speech Representations

The past five years have been dominated by *self-supervised learning* (SSL), a paradigm that extracts powerful representations from large unlabelled corpora by solving pretext

tasks, then fine-tunes on small labelled downstream datasets. Three encoder families have become the reference systems in SER.

wav2vec 2.0 [3] frames SSL as masked prediction in a discretised latent space: a convolutional encoder maps raw waveform frames to latent vectors, a subset of which are masked, and the model is trained to identify the correct discretised code among distractors via a contrastive objective. Fine-tuned on as few as ten minutes of labelled speech, wav2vec 2.0 matched the performance of fully supervised systems trained on hundreds of hours. Pepino *et al.* [45] were among the first to demonstrate that these representations transfer effectively to SER, and more recent work continues to refine the paradigm: Wang and Yang [63], for instance, pair a fine-tuned wav2vec 2.0 encoder with a neural controlled-differential-equation classifier for emotion recognition.

HuBERT [25] iterates the approach: offline cluster assignments of MFCC features or preceding HuBERT representations are used as pseudo-labels, and the model learns by masked prediction of these hidden units. Unlike wav2vec 2.0, HuBERT does not require a contrastive quantiser at training time, which simplifies the architecture and improves stability across different training regimes.

WavLM [12] extends HuBERT with a denoising objective: corrupted speech frames are reconstructed alongside the masked prediction task, providing representations that are simultaneously semantic and noise-robust. On the SUPERB benchmark [66], WavLM consistently ranks among the strongest models across diverse downstream tasks including SER, speaker verification, and source separation.

The EMO-SUPERB benchmark [65] specialises SUPERB for affective computing, providing a standardised comparison of SSL encoder quality across multiple SER corpora and label sets. It confirms that WavLM and HuBERT representations outperform wav2vec 2.0 on most evaluation sets, though the gap narrows when the downstream task involves dimensional outputs rather than categorical classification.

2.2.2. Whisper as an Emotion Encoder

Radford *et al.* [50] introduced *Whisper*, an encoder–decoder architecture trained with weak supervision on approximately 680,000 hours of multilingual web audio. Although designed for automatic speech recognition (ASR), Whisper’s encoder acquires broad acoustic representations that transfer to paralinguistic downstream tasks. Gong *et al.* [22] showed that Whisper representations retain audio-event and speaker characteristics that pure ASR training would not be expected to preserve, attributing this to the diversity of the training data rather than task-specific design.

For emotion recognition, the broad and diverse pretraining of Whisper is particularly attractive: exposure to a wide variety of recording conditions, speaking styles, and acoustic environments may reduce the domain mismatch that plagues models trained on narrow, homogeneous corpora. Wagner *et al.* [62] demonstrated CCC scores of 0.638 on MSP-Podcast valence using transformer fine-tuning, the highest reported at the time of publication.

This thesis uses two publicly released checkpoints from the same fine-tuning effort [20], both built on top of Whisper-large-v3 and trained on MSP-Podcast but serving distinct roles in the system. The first, `tiantiaf/whisper-large-v3-msp-podcast-emotion`, provides a nine-class categorical classification head (Anger, Contempt, Disgust, Fear, Happiness, Neutral, Sadness, Surprise, Other) and is used by the Whisper categorical SER backend described in Section 3.4.2. The second, `tiantiaf/whisper-large-v3-msp-podcast-emotion-dim`, provides continuous valence, arousal, and dominance regression heads (sigmoid-activated, trained on the MSP-Podcast dimensional annotations) and underlies the dimensional SER backend and the *distances-to-probabilities* algorithm introduced in Section 2.3.2 and described in full in Section 3.4.3. Shendabadi *et al.* [57] recently explored attentive pooling strategies on top of Whisper representations to further improve SER accuracy, providing empirical support for encoder-only emotion inference without decoder involvement.

2.3. Dimensional SER: Valence/Arousal Inference

2.3.1. The Circumplex Model and Anchor Coordinates

Russell’s *circumplex model of affect* [51] organises emotions on a circle in the valence–arousal plane: high-arousal positive emotions (*joy, excitement*) in the upper-right quadrant, high-arousal negative emotions (*anger, fear*) in the upper-left, low-arousal positive emotions (*calm, contentment*) in the lower-right, and low-arousal negative emotions (*sadness, boredom*) in the lower-left. The model predicts that neighbouring emotions share acoustic and expressive features, while emotions in opposite quadrants are maximally dissimilar, a prediction well supported by prosodic data [5, 52].

Jing *et al.* [30] formalised a conversion procedure from discrete emotion labels to dimensional coordinates. Starting from the circumplex geometry, they assigned each discrete category a canonical polar angle θ corresponding to that emotion’s prototypical position on the wheel. Their primary aim was annotation enrichment: given a recording labelled only with a categorical class, the method infers a plausible dimensional rating, allowing

datasets with categorical-only labels to contribute to dimensional model training. This thesis adopts the same polar positions from this framework and computes the anchor coordinates as $(V, A) = (\cos \theta, \sin \theta)$, placing each emotion on the unit circle; the resulting values are listed in Table 2.1 and shown geometrically in Figure 2.1.

Table 2.1: Circumplex anchor coordinates from Jing *et al.* [30]; **bold** = adopted by the dimensional SER backend.

Emotion	Valence V	Arousal A
Pleasure	$\cos(\pi/20)$	$\sin(\pi/20)$
Happiness	$\cos(3\pi/20)$	$\sin(3\pi/20)$
Pride/Elation	$\cos(5\pi/20)$	$\sin(5\pi/20)$
Excitement	$\cos(7\pi/20)$	$\sin(7\pi/20)$
Surprise	$\cos(9\pi/20)$	$\sin(9\pi/20)$
Anger	$-\cos(9\pi/20)$	$\sin(9\pi/20)$
Hate	$-\cos(7\pi/20)$	$\sin(7\pi/20)$
Contempt	$-\cos(5\pi/20)$	$\sin(5\pi/20)$
Disgust	$-\cos(3\pi/20)$	$\sin(3\pi/20)$
Fear	$-\cos(\pi/20)$	$\sin(\pi/20)$
Boredom	$-\frac{1}{2}$	0
Disappointment	$-\cos(\pi/20)$	$-\sin(\pi/20)$
Shame	$-\cos(3\pi/20)$	$-\sin(3\pi/20)$
Regret	$-\cos(5\pi/20)$	$-\sin(5\pi/20)$
Guilt	$-\cos(7\pi/20)$	$-\sin(7\pi/20)$
Sadness	$-\cos(9\pi/20)$	$-\sin(9\pi/20)$
Compassion	$\cos(9\pi/20)$	$-\sin(9\pi/20)$
Relief	$\cos(7\pi/20)$	$-\sin(7\pi/20)$
Admiration	$\cos(5\pi/20)$	$-\sin(5\pi/20)$
Love	$\cos(3\pi/20)$	$-\sin(3\pi/20)$
Contentment	$\cos(\pi/20)$	$-\sin(\pi/20)$
Neutral	0	0

2.3.2. The Inversion Problem: From Coordinates to Probabilities

Jing *et al.* [30] address the *forward* direction of the label–dimension correspondence: given a recording annotated with a discrete emotion class, infer plausible valence–arousal coordinates so that categorically labelled datasets can contribute to dimensional model training.

This thesis requires the complementary *inverse*: given a continuous model output (V, A) , compute a probability distribution over discrete labels. This inversion is non-trivial. In-

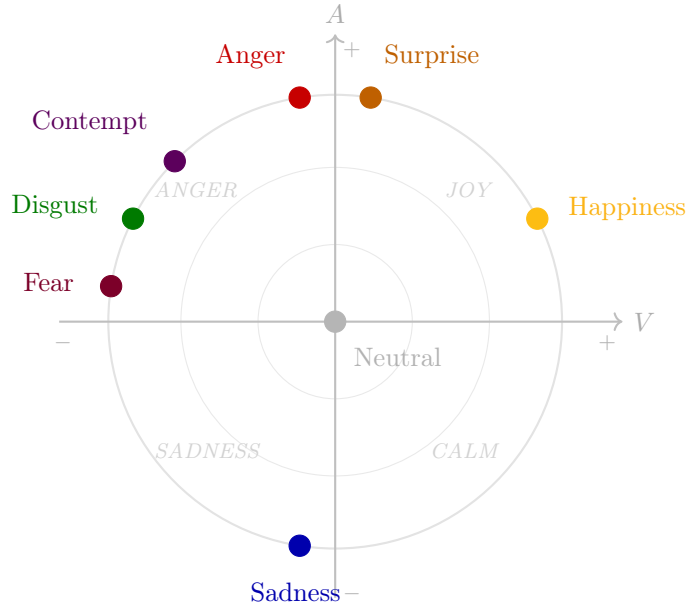


Figure 2.1: Russell’s circumplex model of affect [51] with the eight emotion anchors adopted from Jing *et al.* [30] (coordinates in Table 2.1). *Neutral* is placed at the origin; peripheral emotions lie on the unit circle ($r = 1$).

finitely many distributions are consistent with any given continuous point, and the most natural baseline, assigning all mass to the nearest anchor (Voronoi assignment), discards all distributional information about neighbouring emotions, reducing the representation to a hard-boundary classification equivalent to argmax . A principled formulation must exhibit four geometric properties across the circumplex. Two are baseline correctness conditions that any usable mapping should meet: (a) at each peripheral anchor the corresponding emotion must be the dominant mode, with residual mass routed to its geometric neighbours rather than scattered across the label space; and (d) the mapping must remain coherent over the entire valence–arousal plane, varying smoothly without spurious modes. The other two are the *structural requirements* that rule out the naive formulations and genuinely constrain the construction: (b) *Neutral* must be handled via the radius from the circumplex origin rather than by distance to its anchor, because the origin is equidistant from all peripheral anchors and a distance-based formula would assign near-uniform peripheral probability precisely where a central prediction should concentrate mass on *Neutral*; and (c) predictions at midpoints between two adjacent peripheral anchors should yield bimodal distributions rather than collapsing to the geometric centroid, which would misrepresent genuine uncertainty between two neighbouring affective states. These four properties are the behaviours the synthetic suite of Chapter 5 verifies in turn.

The *distances-to-probabilities* algorithm that satisfies these requirements is derived and evaluated in Section 4.3.2. As a construction, mapping a continuous point to a categorical distribution by a temperature-scaled softmax over distances to fixed anchors is closely related to radial-basis-function and soft-assignment kernels; the contribution here is not the kernel form itself but its use as a shared, streaming representation that places dimensional outputs in the same probabilistic space as the categorical backends, with circumplex-specific handling of *Neutral* and a geometric constraint that fixes the admissible temperature. Ispas *et al.* [28] independently validated the value of joint categorical and dimensional modelling in a multimodal setting, reinforcing the motivation for bridging the two representation spaces.

2.3.3. Valence–Arousal Asymmetry

An empirical regularity of dimensional SER is that arousal is substantially easier to predict than valence [41, 62]. Arousal correlates with prominent acoustic features (energy, speaking rate, fundamental frequency range) that are robustly encoded by any feature extractor. Valence, by contrast, depends heavily on subtle spectral qualities and on linguistic content: the words spoken, not only how they are spoken [59]. On MSP-Podcast, the gap between arousal CCC (≈ 0.65) and valence CCC (≈ 0.45) for strong baselines persisted for years and was only meaningfully reduced by transformer fine-tuning that learns linguistic representations alongside acoustic ones [62].

This asymmetry has direct operational implications for the dimensional backend. The system maps V–A coordinates to emotion categories; any systematic bias in valence will affect the distribution of probability mass across positive and negative affective states (*Happiness*, *Calm* versus *Anger*, *Fear*). The temporal properties of the pipeline output on realistic speech are characterised empirically in Chapter 5, where the window duration sensitivity analysis provides insight into how the acoustic context provided to the model shapes the resulting probability distributions.

2.4. Uncertainty Quantification in Neural SER

2.4.1. Sources of Uncertainty in Affective Labels

Kendall and Gal [31] establish the canonical taxonomy of uncertainty in deep learning. *Aleatoric* uncertainty (also called *data uncertainty*) is irreducible: it is inherent in the observation process itself and cannot be reduced by collecting more training data or improving the model. It captures the fundamental randomness or ambiguity present in

the input. *Epistemic* uncertainty (also called *model uncertainty*) arises from gaps in the model’s knowledge due to limited or unrepresentative training data; it can in principle be reduced by exposing the model to more data from the relevant distribution.

In SER, both forms of uncertainty are prominent and have concrete interpretations. Aleatoric uncertainty is structurally high because human emotion is inherently ambiguous: a speaker may express two emotions simultaneously, and listeners frequently disagree on the dominant one [58]. Steidl *et al.* [58] showed that a classifier reaching approximately 60% raw accuracy on a four-class spontaneous-emotion task already matches the average performance of five human annotators, and that the “correct” label for many utterances depends on which cognitive frame the annotator adopts, a consequence of irreducible aleatoric uncertainty in the labelling process itself. Peterson *et al.* [46] demonstrated that preserving annotator disagreement as a soft-label distribution, rather than collapsing it to a single majority class, improves the robustness of downstream classifiers, consistent with the view that disagreement encodes genuine aleatoric signal rather than annotation noise. In the same spirit, Lotfian and Busso [38] train a soft-label SVM on the annotator label distribution rather than on a single hard class, showing that learning from the soft targets improves categorical emotion prediction on naturalistic speech. Epistemic uncertainty manifests when the system encounters speech styles, languages, or acoustic conditions not represented in the training corpus: the model is asked to generalise beyond its experience, and its predictions become unreliable regardless of its calibration on in-distribution data.

Shannon entropy of the output distribution

$$H(p) = - \sum_e p_e \log_2 p_e \quad (2.2)$$

(in bits) serves as a practical scalar proxy for total uncertainty in this work. A high-entropy output either reflects genuine affective ambiguity (*aleatoric*) or indicates that the model lacks relevant prior experience (*epistemic*). A near-uniform distribution (high entropy) is less informative than one concentrated on a single emotion class, regardless of which class that is.

2.4.2. Probability Calibration and Expected Calibration Error

A model is *calibrated* if its output probability for a class equals the empirical frequency with which that class is the ground-truth label, i.e., $\Pr(\hat{y} = y \mid \hat{p} = c) = c$ for all confidence levels c . Guo *et al.* [23] showed that modern deep networks, despite achieving high

accuracy, are systematically *overconfident*: the maximum softmax probability is higher than the actual accuracy on a binned reliability diagram. They further demonstrated that *temperature scaling*, which divides logits by a learned scalar $T > 1$ before softmax, is a simple and remarkably effective post-hoc remedy that preserves accuracy while substantially reducing miscalibration.

The *Expected Calibration Error* (ECE) [23] quantifies miscalibration as the average deviation between confidence and accuracy over M equally spaced confidence bins:

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)|, \quad (2.3)$$

where B_m is the set of samples whose predicted confidence falls in bin m , and n is the total number of samples. In the experimental evaluation (Chapter 5), Shannon entropy of the output distribution is reported as a scalar proxy for prediction uncertainty, characterising how concentrated or diffuse each prediction is; Expected Calibration Error, which requires per-utterance ground-truth labels, is a natural complement once an annotated corpus is available (Section 6.2).

Chou *et al.* [13] were the first to evaluate calibration systematically for multi-label SER classifiers: they applied a multi-label variant of temperature scaling with emotion-dependent temperature parameters and achieved ECE reductions of 15–20% while preserving classification accuracy. Schröder *et al.* [53] extended this line of work by benchmarking multiple uncertainty quantification strategies on realistic in-the-wild SER scenarios and found that temperature scaling remained competitive with more complex Bayesian and ensemble approaches, while being deployable in real-time systems without architectural modifications.

2.4.3. Out-of-Distribution Detection

At deployment, a SER system may receive audio that is acoustically unlike anything in its training distribution: a heavily accented speaker, a noisy channel, a non-speech signal, or a language the model has never encountered. In these cases, even a calibrated model may produce overconfident outputs for incorrect classes.

Hendrycks and Gimpel [24] established the *maximum softmax probability* (MSP) as a foundational baseline for out-of-distribution (OOD) detection: inputs with low maximum softmax probability are more likely to be OOD than correctly classified in-distribution examples. Despite its simplicity, MSP provides a surprisingly strong reference point against which more complex detectors must be compared. Maximum softmax probability and

entropy (equivalent in the one-hot case) are the standard scalar proxies for flagging such inputs; in the present system the maximum softmax probability is preserved on every event as the `confidence` field, while entropy-based suppression of low-confidence predictions is identified as future work. Beyond such scalar proxies, recent work brings distribution-free guarantees to SER: Jia *et al.* [29] combine conformal prediction with risk control to emit calibrated prediction *sets* carrying a user-specified coverage guarantee, a principled route to quantifying predictive uncertainty that complements the maximum-softmax-probability signal preserved here.

2.5. Real-Time and Streaming SER

2.5.1. Sliding-Window Inference and Temporal Aggregation

Batch SER operates on complete, segmented utterances. Real-time SER must produce predictions continuously, before the speaker has finished a turn. The pipeline adopts a *sliding window* strategy: audio is segmented into overlapping windows of fixed duration w seconds with hop $h < w$, and the SER model is invoked once per hop.

Metallinou *et al.* [41] adopt temporal windowing for continuous emotion tracking in a multimodal setting, combining prosodic and body-language cues within successive context windows. Although the modality differs from the speech-only setting of this thesis, the rationale for fixed temporal granularity (capturing short-term affect dynamics while smoothing frame-level noise) is analogous.

Window length is a key hyper-parameter [35]: shorter windows capture rapid affect changes but lack sufficient temporal context for models with large receptive fields, while longer windows introduce prediction latency and risk smoothing over rapid emotional transitions.

Once per-window predictions are available, the system aggregates them using a *weighted average* of probability distributions, where each window’s weight is proportional to its temporal overlap with the utterance interval under evaluation.

Lin and Busso [35] propose a comparable decomposition into temporal chunks for utterance-level SER: variable-length utterances are split into fixed-duration segments, each processed independently, with predictions aggregated at the end. Their framework uses a learned attention mechanism rather than overlap-weighted averaging, but the underlying principle (reducing a variable-length sequence to a fixed set of independently scored segments) is shared with the approach adopted here. Mirsamadi *et al.* [42] ad-

dressed the same aggregation problem with a deep bidirectional long short-term memory (BiLSTM) network equipped with a *local attention* mechanism that learns to assign higher weight to emotionally salient frames, achieving approximately 1–2% weighted accuracy improvement over mean pooling on IEMOCAP. Against these learned poolers, the present system deliberately adopts the unlearned, overlap-weighted variant: it requires no additional training and runs deterministically online, trading the modest accuracy gain a learned attention pooling can offer for transparency and zero training cost, in keeping with the systems orientation of this work.

2.5.2. Deployment in Call-Centre Environments

Real call-centre audio presents conditions substantially more challenging than laboratory benchmarks. Devillers *et al.* [16, 17] identified three core challenges that distinguish real call-centre corpora from acted datasets: the contrast between genuinely spontaneous emotional expression and the scripted affect of laboratory recordings; the pervasive presence of *blended* or mixed emotions that resist single-label annotation; and high inter-annotator disagreement reflecting genuine affective ambiguity in conversational settings. These characteristics substantially reduce the ecological validity of UAR scores achieved on acted corpora such as EmoDB [8].

Feng *et al.* [21] introduced CusEmo, a large-scale corpus of real French customer-service call-centre recordings with continuous dimensional annotation (valence and arousal). Working on genuinely spontaneous telephone speech they found that a label sampling rate of 500 ms outperforms finer-grained 40 ms annotation, and report CCC scores of approximately 0.67–0.70 for valence and 0.48–0.51 for arousal, substantially below values typically reported on laboratory corpora, illustrating the domain gap faced by any SER system operating on real call-centre streams.

2.6. Emotion Corpora

2.6.1. Acted Corpora

EmoDB [8] is a German corpus of 535 utterances recorded by ten professional actors portraying seven emotions: anger, boredom, disgust, fear, happiness, neutral, and sadness. Its small size, clean recording conditions, and unambiguous acted portrayals make it an upper bound for categorical SER accuracy on the acted paradigm.

IEMOCAP [9] comprises approximately 12 hours of English dyadic acted and im-

provided interactions between ten actors across five sessions, with annotations for both categorical labels and V–A–dominance dimensions. Despite its scripted nature, IEMO-CAP is the most widely cited benchmark in the SER literature and remains the standard reference for comparing SSL fine-tuning results [12, 25, 45]. Its four-class subset (neutral, happiness, sadness, anger) dominates comparison tables, though this subset suppresses the full class variability of the corpus.

CREMA-D [11] (Crowd-Sourced Emotional Multimodal Actors Dataset) comprises 7,442 audio clips from 91 actors spanning ages 20–74 and diverse ethnic backgrounds (Caucasian, African American, Hispanic, and Asian). Six emotion categories (anger, disgust, fear, happiness, neutral, sadness) were elicited through short sentences and crowd-sourced to over 2,000 raters, making CREMA-D one of the most demographically representative acted corpora in the SER literature. Its short single-sentence clips, however, are unsuited to the platform-level evaluation conducted in this thesis, for the reasons set out in Section 5.1.1.

RAVDESS [36] (Ryerson Audio-Visual Database of Emotional Speech and Song) contributes 24 professional actors performing eight emotions in North American English; like CREMA-D and TESS, its brief acted clips serve as a reference corpus rather than as evaluation material for the streaming pipeline studied here.

2.6.2. Naturalistic Corpora

MSP-Podcast [37] is a large-scale corpus of naturalistic English speech extracted from public podcast recordings. Unlike IEMOCAP and EmoDB, it contains spontaneous rather than acted emotional speech, making it substantially more representative of the conditions encountered in real call centres. At the time of writing, the corpus contains over 100 hours of labelled speech distributed across nine categorical classes and continuous V–A–dominance annotations. MSP-Podcast underpins the Interspeech 2025 Challenge on Speech Emotion Recognition in Naturalistic Conditions [43], a recent community benchmark that consolidates in-the-wild categorical and dimensional SER and confirms the corpus as the reference setting for the deployment regime this thesis targets.

The **TESS** [47] corpus (Toronto Emotional Speech Set) provides Canadian English acted speech from two female speakers across seven emotion categories; like CREMA-D, its very short single-word and single-sentence clips make it a reference corpus rather than evaluation material for this work (Section 5.1.1).

2.6.3. Cross-Lingual Coverage and EmoBox

The EmoBox benchmark [39] provides a unified multilingual, multi-corpus evaluation toolkit covering 32 datasets across fourteen languages. Its analysis confirms that SSL models fine-tuned on English show consistent performance degradation across Italian, German, and non-Indo-European languages, and that this degradation is more severe for valence than for arousal [62]. Focusing on the encoder side, Bellver *et al.* [7] evaluate wav2vec 2.0 and Whisper as cross-lingual feature extractors across eight languages and find that Whisper’s multilingual pre-training yields the strongest features on Italian and other non-English languages, evidence that directly supports the choice of a Whisper backbone for the Italian deployment targeted in this thesis.

The cross-corpus degradation documented by EmoBox illustrates the broader challenge that any acoustic SER model faces when its deployment audio differs in language or recording conditions from its training corpus.

3 | System Design and Architecture

This chapter describes the design and implementation of the real-time call-centre emotion monitoring system. Section 3.1 introduces the three-layer architecture and its structural enforcement. Sections 3.2–3.4 describe each layer from the inside out, following the direction of the dependency rule. Section 3.5 covers the HTTP and WebSocket transport layer. Section 3.6 describes the React frontend. Section 3.7 discusses key design decisions and the engineering trade-offs behind them.

3.1. Architectural Overview

3.1.1. Three-Layer Decomposition

The system is structured according to the hexagonal architecture [14] and the dependency rule of Clean Architecture [40]. All source-code imports travel in a single direction, from outer layers towards the domain core, and never in the reverse direction. Figure 3.1 illustrates the three layers and the interfaces between them.

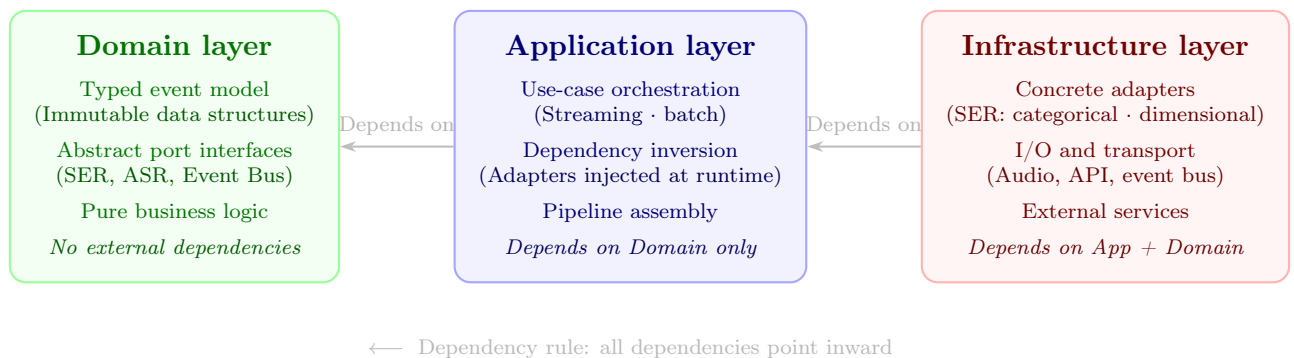


Figure 3.1: Three-layer hexagonal architecture [14, 40]. All source-level imports flow inward; the domain core carries no external dependencies.

The **domain layer** (`src/app/domain/`) contains pure Python types and func-

tions with no deep-learning-framework or web-framework dependencies: no `import tensorflow`, `import torch`, or `import fastapi` statement appears in this package, which relies only on NumPy and lightweight signal-processing utilities. The layer defines: the typed event dataclasses (`EmotionEvent`, `SERLive`, `SERWeightedFrame`, `EnrichedUtterance`, `TranscriptPiece`); the structural-typing ports `ISERBackend`, `ISTTEngine`, and `IEventBus`; and the pure-Python pipeline functions `ser_stream()`, `stt_stream()`, and `aggregate_stream()`.

The **application layer** (`src/app/application/`) orchestrates domain objects via two use cases: offline batch analysis and real-time streaming. It provides the `AppContainer`, a dependency-injection container that constructs the concrete ML adapters and supplies them to the pipeline factory. The application layer imports from the domain layer only.

The **infrastructure layer** (`src/app/infrastructure/`) implements all concrete I/O: four ML adapters satisfying `ISERBackend`, one STT adapter satisfying `ISTTEngine`, the FastAPI HTTP/WebSocket server, audio ingestion and preprocessing, and an IPC event bus. It may import from both the application and domain layers, and from any external framework.

3.1.2. Structural Dependency Enforcement

The dependency rule is enforced not by convention but by Python’s import system: the domain package contains no conditional imports and no `TYPE_CHECKING` workarounds that would allow framework names to appear at runtime. Any violation is immediately detectable by a static analysis pass or by importing the domain package in a minimal environment devoid of framework dependencies; the absence of framework-specific identifiers from the domain namespace is a directly verifiable invariant. This enforcement makes the domain independently testable and insulates it from framework version churn [56].

3.2. Domain Layer

3.2.1. Typed Event Model

All inter-component communication in the domain is carried by immutable Python `dataclass` instances defined in `events.py`. Table 3.1 summarises the five event types and their fields.

The optional `valence` and `arousal` fields are populated only when the dimensional SER backend is active. Downstream components check for `None` rather than testing the

Table 3.1: Domain event types. All timestamps are in seconds from session start.

Type	Key fields	Producer
EmotionEvent	start, end, label, confidence, probs[], valence?, arousal?	ser_stream()
TranscriptPiece	text, start, end, is_final	stt_stream()
SERLive	start, end, probs[]	Aggregator
SERWeightedFrame	center, avg_probs[], dominant, confidence, num_events_used, valence?, arousal?	Aggregator
EnrichedUtterance	text, start, end, is_final, avg_probs[], confidence, dominant, num_events_used, valence?, arousal?	Aggregator

backend type, preserving substitutability across all four adapter implementations.

3.2.2. Service Ports

The domain defines three structural-typing ports as Python `Protocol` classes [33] with `@runtime_checkable` semantics.

The SER port requires two members:

```
class ISERBackend(Protocol):
    def predict(self, audio: np.ndarray,
               sample_rate: int) -> Dict[str, float]: ...

    @property
    def emotion_labels(self) -> List[str]: ...
```

The STT port requires `process_audio(audio_bytes: bytes) -> List[TranscriptPiece]` and `finalize() -> List[TranscriptPiece]`, representing the streaming and flush phases of a stateful recogniser. The event bus port requires `publish(event: Dict[str, Any])` and `get_nowait()` for non-blocking IPC between processes.

Structural typing means that any class exposing these members satisfies the port without inheriting from it. This avoids a common architectural anti-pattern in which adapters depend on an abstract base class defined in the domain, inadvertently pulling domain imports into the adapter namespace.

A secondary optional extension point, `predict_with_va()`, is detected at runtime via `hasattr(backend, 'predict_with_va')`: if present, the SER stream invokes it to obtain the raw (V, A) coordinates alongside the probability distribution; if absent, the coordinates are set to `None`. This allows the dimensional backend to enrich events with continuous

affect coordinates without requiring any change to the port contract.

3.2.3. SER Stream: Sliding Window and VAD Gating

The `ser_stream()` function implements the sliding-window inference loop described in Section 2.5.1. It accumulates a rolling float32 buffer from incoming pulse-code modulation (PCM) frames, advancing by *hops* of $h_{\text{samples}} = h \cdot f_s$ samples (default $h = 0.5\text{ s}$, $f_s = 48,000\text{ Hz}$).

Before accumulating a hop into the buffer, the function checks the *Voice Activity Detection* (VAD) gate:

$$\overline{\text{vad}} = \frac{1}{N} \sum_{i=1}^N \text{vad}_i, \quad \text{skip hop if } \overline{\text{vad}} < \theta_{\text{vad}}, \quad (3.1)$$

where vad_i is the per-frame speech-activity probability produced by the RNNoise denoiser [61] and $\theta_{\text{vad}} = 0.5$ by default. Hops dominated by background noise or silence are discarded, preventing the SER backends from wasting inference cycles on non-speech audio and suppressing spurious emotion predictions.

Once the buffer reaches $w_{\text{samples}} = w \cdot f_s$ samples (default $w = 3.0\text{ s}$), a window is extracted and the SER backend is invoked; then the oldest h_{samples} samples are discarded from the buffer, leaving a $w - h = 2.5\text{ s}$ overlap with the next window. At end of stream, any residual buffer of at least one second of audio is inferred as a final window, ensuring that short utterances near the end of a recording are not silently dropped.

3.2.4. Temporal Aggregation

The `aggregate_stream()` function merges the SER and STT event streams into three output generators. It launches three concurrent daemon threads: two *pump* threads that drain the SER and STT iterators into thread-safe queues, and one *dispatcher* thread that processes events from both queues in timestamp order (keyed on each event's `end` timestamp).

Workflow 1 – SER Live. Every `EmotionEvent` is normalised and forwarded directly as a `SERLive` event with no further modification. `SERLive` events are intended for immediate display and are not retained by the frontend store.

Workflow 2 – SER Weighted. The aggregator accumulates a sliding buffer of `EmotionEvent` instances and emits a `SERWeightedFrame` for each temporal grid point

$c_k = t_0 + k \cdot h_s$ (where t_0 is the grid origin anchored to the first received SER event) once it is *safe*:

$$c_k \leq t_{\text{last_ser_end}} - w_s, \quad (3.2)$$

where $t_{\text{last_ser_end}}$ is the end timestamp of the most recently received **EmotionEvent** and w_s is the SER window duration. This guard ensures that no frame is emitted prematurely, before all SER windows whose temporal support overlaps the frame centre have been received. The emotion distribution for frame c_k is the uniform average over the set W_k of SER events whose windows overlap the frame interval $[c_k - h_s/2, c_k + h_s/2]$:

$$\bar{p}(c_k) = \frac{1}{|W_k|} \sum_{e \in W_k} p_e. \quad (3.3)$$

The frame also carries **num_events_used**, the number of events actually contributing to the average after the locality cap (at most four), which downstream consumers use to assess reliability. At end of stream the guard is bypassed (**force=True**) to flush remaining frames. The formal characterisation of the candidate window set W_k , the four-event cap, and the rationale for uniform weighting are presented in Section 4.2.2.

Workflow 3 – Enriched Utterances. When a **TranscriptPiece** arrives from the STT stream, the aggregator computes an *overlap-weighted* average of all SER events whose windows intersect the utterance interval $[t_{\text{start}}, t_{\text{end}}]$:

$$\bar{p}_{\text{utt}} = \frac{\sum_e \text{ov}(e) \cdot p_e}{\sum_e \text{ov}(e)}, \quad (3.4)$$

where $\text{ov}(e) = \max(0, \min(t_{\text{end}}, e_{\text{end}}) - \max(t_{\text{start}}, e_{\text{start}}))$ is the temporal overlap between SER event e and the utterance. The result is emitted as an **EnrichedUtterance** coupling the transcript text with its emotional annotation, together with the overlap-weighted confidence and **num_events_used**. When the dimensional backend is active, the valence and arousal coordinates are averaged with the same overlap weights. The motivation for overlap weighting over the uniform-weight strategy of Workflow 2 is discussed in Section 4.2.3.

3.3. Application Layer

3.3.1. Dependency Injection Container

The `AppContainer` class instantiates concrete adapters on demand using Python's `functools.cached_property`:

```
class AppContainer:
    def __init__(self, ser_backend_name: str = "legacy",
                  whisper_temperature: float = 1.0,
                  stt_language: str = "it") -> None:
        ...

    @cached_property
    def ser_backend(self) -> ISERBackend:
        if self._ser_backend_name == "legacy":
            return LegacySERAdapter(model_path=...)
        elif self._ser_backend_name == "whisper_categorical":
            return WhisperSERAdapter(...)
        elif self._ser_backend_name == "whisper_dimensional":
            return WhisperDimensionalSERAdapter(...)
        elif self._ser_backend_name == "hume":
            return HumeSERAdapter(api_key=HUME_API_KEY)
        else:
            raise ValueError(...)
```

Three `cached_property` instances are defined: `ser_backend`, `stt_engine`, and `event_bus`. `cached_property` semantics guarantee that each adapter is constructed at most once per container instance and that the construction is deferred to the moment of first access, not to import time. This design means that loading the application module does not trigger the download of Whisper weights or the initialisation of a TensorFlow session.

To reduce first-request latency, the FastAPI application registers a lifespan handler that fires a non-blocking background thread on startup to pre-load the default SER backend. The STT engine is excluded from this warmup and loaded lazily on first session access: each session may request a different STT language model, making per-session lazy instantiation the correct strategy. Container instances are memoised per configuration, so constructing the ML adapters is paid once and reused across sessions that request the same backend, while each session drives its own pipeline instance with independent buffers and streaming state. The batch analysis endpoint follows the same pattern.

The constructor accepts three session-level parameters: the backend name, an optional Whisper softmax temperature (default 1.0, applicable only to `whisper_categorical`), and the STT language ("`it`" for Italian or "`en`" for English).

3.3.2. Pipeline Factory

`build_pipeline()` assembles the end-to-end processing graph from its injected dependencies.

First, the raw PCM frame iterator is passed to `rnnoise_denoise_stream()`, which applies RNNoise frame-by-frame and attaches a per-frame VAD probability. The output is a stream of `(denoised_frame, vad_prob)` tuples.

Second, `broadcast_copy()` duplicates the preprocessed stream into two independent thread-safe branches, one for SER and one for STT. This is necessary because both streams consume the same audio samples but at different cadences: SER advances by 0.5 s hops, while STT consumes individual 10 ms frames. Each payload is copied to a C-contiguous numpy array before enqueueing, preventing aliasing between the two consumers.

Third, the SER branch is passed to `ser_stream()` with the configured window, hop, and VAD parameters.

Fourth, the STT branch is passed to `stt_stream()`, which additionally resamples the 48 kHz audio to 16 kHz using `soxr.ResampleStream` before forwarding bytes to the Vosk recogniser.

Fifth, `aggregate_stream()` merges both output iterators and returns the three event generators.

The factory returns a `PipelineHandles` dataclass, which exposes the three typed generators (`live`, `weighted`, `enriched`) and a `close()` callable that signals the broadcast thread to stop and drains all internal queues. Figure 3.2 illustrates the complete data-flow graph.

3.3.3. Use Cases

Two use case classes coordinate the pipeline for their respective modes.

`RunStreamingSession` manages a real-time WebSocket session. On `start()`, it calls `build_pipeline()` with a frame source backed by a thread-safe input queue, then launches three daemon pump threads (one per generator: `live`, `weighted`, `enriched`) that forward events to an asyncio output queue via `asyncio.run_coroutine_threadsafe`.

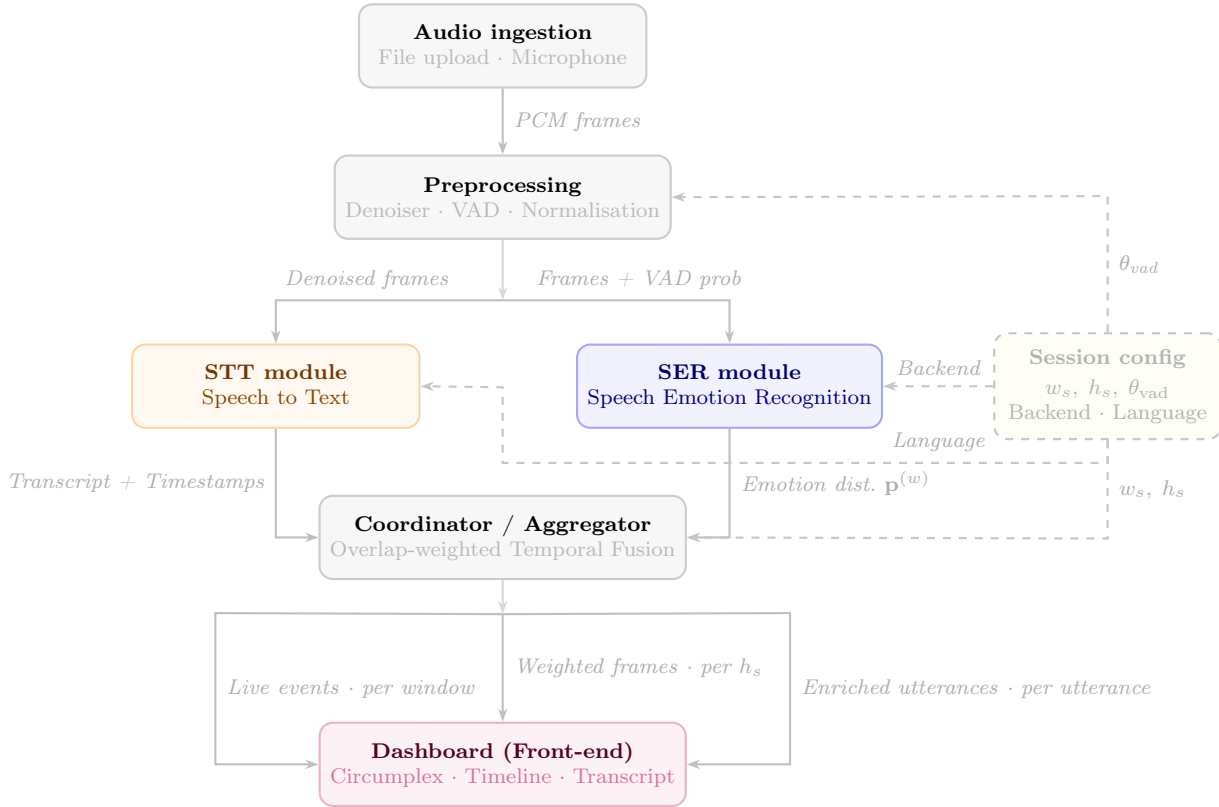


Figure 3.2: End-to-end data-flow of the real-time pipeline (implementation in Section 3.3). Denoised audio is routed in parallel to the STT and SER branches; the aggregator fuses both into three output granularities; session parameters are injected at start (dashed).

Incoming binary PCM chunks are enqueued by the async WebSocket handler via `feed_bytes()`, and the async `next_event()` coroutine delivers serialised `StreamEvent` objects to the WebSocket producer.

`RunOfflineAnalysis` executes the full pipeline on a WAV file. It ingests the file via `ingest_audio()`, runs the pipeline, and collects all events in three parallel daemon threads before returning a `BatchResult`. A post-processing step, `_enrich_missed_utterances()`, handles a race condition specific to batch mode: when the Whisper SER backend is used, its longer inference time means Vosk may emit final utterances before any SER events have arrived, resulting in utterances with empty `avg_probs`. The step retroactively computes overlap-weighted SER averages for such utterances using the fully collected `ser_live` list.

3.4. Infrastructure Layer: ML Adapters

3.4.1. Legacy BiLSTM Adapter

`LegacySERAdapter` wraps a TensorFlow/Keras BiLSTM model trained on EmoDB [8] with seven emotion classes: `anger`, `boredom`, `disgust`, `fear`, `happiness`, `neutral`, and `sadness`. Its convolutional–recurrent architecture and the 182-dimensional feature representation described below follow the real-time SER design of Barhoumi and BenAyed [6]. The adapter resamples incoming audio from the pipeline sample rate (48 kHz) to the model’s training rate (22 050 Hz) using `librosa.resample`, then extracts a 182-dimensional hand-crafted feature vector composed of 40 MFCC coefficients, 12 Chroma features, 128 Mel-spectrogram values, one zero-crossing rate scalar, and one root-mean-square energy scalar. This vector is reshaped to $(1, 182, 1)$ before being passed to the BiLSTM stack. The model produces a softmax-normalised probability vector directly from its final Dense layer, so no post-hoc calibration is applied.

The TF/Keras model is loaded lazily via a local `import` inside the `_load()` method, ensuring that `import tensorflow` is never executed unless this backend is explicitly selected.

3.4.2. Whisper Categorical Adapter

`WhisperSERAdapter` wraps the PyTorch model checkpoint `tiantiaf/whisper-large-v3-msp-podcast-emotion` [20] and calls its classification head over nine emotion classes: `Anger`, `Contempt`, `Disgust`, `Fear`, `Happiness`, `Neutral`, `Sadness`, `Surprise`, and `Other`. The model is loaded via the `WhisperWrapper` class provided by the VoxProfile release [20], which handles the HuggingFace download and positional-embedding configuration.

The model returns raw logits; these are scaled by a configurable temperature T before softmax:

$$p_e = \frac{\exp(\ell_e/T)}{\sum_{e'} \exp(\ell_{e'}/T)}, \quad (3.5)$$

where ℓ_e is the raw logit for class e . The temperature defaults to $T = 1.0$ and is exposed as a session-level parameter; higher values (e.g. $T = 1.8$) flatten the distribution and reduce overconfidence on out-of-distribution audio at the cost of lower peak confidence on in-distribution examples. The minimum audio duration accepted by the Whisper encoder is 3 s; inputs shorter than this are zero-padded before inference. The maximum accepted duration is 15 s; longer inputs are truncated. The default analysis window is $w_s = 3.0$ s, the value recommended by the configuration study of the experimental evaluation (Section 5.5.2), which matches the Whisper encoder minimum of 3 s exactly, so that no zero-padding is required at inference time. A shorter window, such as the

$w_s = 2.0$ s minimum-latency option, falls below the 3s minimum and would append 1.0s of zero-padded silence on every inference call; the default configuration avoids this distortion while remaining the more temporally stable choice for every backend.

A benign shape-mismatch warning from Hugging Face’s `modeling_utils` is suppressed during model loading. This warning arises because the checkpoint sets `max_source_positions` to 750 (corresponding to 15s input) instead of the default 1500, causing a mismatch with the pre-trained positional embedding weights; the fine-tuned task head loads correctly and the warning has no effect on inference accuracy.

3.4.3. Whisper Dimensional Adapter

`WhisperDimensionalSERAdapter` uses a separate fine-tuned checkpoint, `tiantiaf/whisper-large-v3-msp-podcast-emotion-dim`, trained specifically on the continuous valence, arousal, and dominance annotations of MSP-Podcast. The forward pass returns three sigmoid-activated outputs in $[0, 1]$; the dominance output is discarded, and valence and arousal are linearly rescaled to the Emotion Circumplex range $[-1, 1]$:

$$v_{\text{GW}} = 2 v_{\text{MSP}} - 1, \quad a_{\text{GW}} = 2 a_{\text{MSP}} - 1. \quad (3.6)$$

The $(v_{\text{GW}}, a_{\text{GW}})$ point is then mapped to a categorical probability distribution by `_distances_to_probs()`. The neutral probability is computed from the radius $r = \sqrt{v_{\text{GW}}^2 + a_{\text{GW}}^2}$:

$$p_{\text{neutral}} = \max(0, 1 - 2r), \quad (3.7)$$

so that $p_{\text{neutral}} = 1$ at the origin and $p_{\text{neutral}} = 0$ at $r \geq 0.5$. The remaining probability mass $(1 - p_{\text{neutral}})$ is distributed among the seven peripheral emotions via a softmax over negative Euclidean distances to their Emotion Circumplex anchor points, with temperature $\tau = 0.45$. The `Other` class is excluded from the distance computation and assigned zero probability, since every point in the (v, a) plane already maps to a geometrically grounded emotion.

The adapter exposes the optional `predict_with_va()` method, which returns the triple `(probs, valence_gw, arousal_gw)`. The domain `ser_stream()` function detects this method at runtime via `hasattr` and, if present, stores the raw (V, A) coordinates in the `EmotionEvent`; otherwise it stores `None`. This design avoids any backend-specific branching in the domain logic.

3.4.4. Hume AI Adapter

`HumeSERAdapter` is a cloud-based backend that satisfies `ISERBackend` by delegating inference to the Hume AI prosody API [4] rather than running a local model. The Hume Expression Measurement API was retired in 2026 (Section 1.1); the adapter is described here as a reference implementation of the cloud-integration path and is retained in the codebase for reproducibility. Each call to `predict()` encodes the audio window as a WAV file, submits it to the Hume batch jobs endpoint, polls until the job completes (up to a 60 s timeout with a 1 s poll interval), and returns the aggregated probability distribution.

Hume’s prosody model produces scores for over fifty fine-grained emotion categories (e.g. *Joy*, *Excitement*, *Triumph*). The adapter maps these to nine canonical labels via a fixed lookup table `_HUME_MAP`: multiple Hume categories can map to the same canonical label (e.g. *Joy*, *Excitement*, *Enthusiasm*, and *Amusement* all map to `Happiness`). Categories absent from the table are mapped to `Other`. When the API returns predictions for multiple temporal segments within the window, the adapter computes a duration-weighted average over segments before applying the mapping. If the audio is shorter than 0.5 s or the job fails, the adapter falls back to a uniform distribution over the nine labels.

The Hume backend requires a valid `HUME_API_KEY` environment variable and incurs per-call API latency (typically 2–10 s per window), making it unsuitable for real-time streaming but useful as an upper-bound reference in batch evaluation.

3.4.5. Vosk STT Adapter and Session Isolation

The `VoskSTTAdapter` delegates to the private `STTEngine` class in `_vosk_engine.py`, which wraps a `KaldiRecognizer` session [48]. The language model is selected at session time: `vosk-model-it-0.22` for Italian (default) and `vosk-model-en-us-0.22` for English, both loaded from the local model directory.

On initialisation the engine calls `SetWords(True)` to obtain per-word timestamps in results, and attempts `SetPartialWords(True)` to enable per-word timestamps in partial results as well. If the installed Vosk version does not expose this method, the engine falls back to text-only partial results.

Delta-partial emission. The Vosk API emits partial results cumulatively: each call returns the *full* partial hypothesis decoded so far, not the newly decoded words since the previous call. The engine maintains a `_last_partial_tokens` list and emits only the *delta* between consecutive partial results, associating each new word with its Kaldi-assigned start and end timestamp. This allows the aggregator to enrich each partial word

segment with its corresponding SER window instead of waiting for a phrase boundary.

Session isolation. Correct session isolation requires explicit recogniser reconstruction between sessions: a `KaldiRecognizer` instance accumulates internal state across calls to `FinalResult()`, including an absolute sample counter and an n-gram decoding history, so that reusing the same instance across sessions would cause the second session to inherit the clock of the first and produce incorrectly offset word timestamps. The adapter therefore reconstructs the `KaldiRecognizer` at the end of `finalize()`, after flushing the last result of the current session:

```
def finalize(self) -> List[TranscriptPiece]:
    ...
    self.samples_processed = 0
    self._last_partial_tokens = []
    self._last_partial_end = None
    self.recognizer = KaldiRecognizer(self.model, self.sample_rate)
    self.recognizer.SetWords(True)
    ...
```

This reconstruction is cheap because the underlying `Vosk Model` object (the acoustic and language model weights) is shared and not reloaded.

3.5. HTTP and WebSocket API

3.5.1. FastAPI Application

The HTTP server is a `FastAPI` application with two routers: `analysis` (batch) and `stream` (real-time). Cross-Origin Resource Sharing (CORS) is restricted to the two local origins `http://localhost:5173` and `http://127.0.0.1:5173`, the default Vite development server addresses, to prevent unintended access from other origins during development. The application is started with `uvicorn`:

```
PYTHONPATH=src uvicorn app.infrastructure.api.app:app \
    --reload --port 8000
```

3.5.2. WebSocket Streaming Endpoint

The `/stream` endpoint follows a four-phase protocol. The client opens a WebSocket connection and immediately sends a JSON message of the form `{"config": {...}}`, encoding the desired backend, window parameters, and partial-emission flag.

The server constructs a `SessionConfig`, starts the pipeline, and responds with a `StatusMessage(state="started")` JSON object. During the session the client sends raw PCM int16 binary frames at 48kHz with 480 samples (10ms) per frame; the server responds asynchronously with `StreamEvent` JSON objects of types `ser_live`, `ser_weighted`, `utterance`, `status`, and `end`. To terminate the session the client sends the plain-text string "stop"; the server drains the pipeline and closes the connection.

Two `asyncio` tasks run concurrently: `ws_consumer` drains incoming frames from the WebSocket and forwards binary payloads to the session via `feed_bytes()`, and `ws_producer` reads completed events from the session's output queue and serialises them to the WebSocket. `asyncio.wait(return_when=FIRST_COMPLETED)` monitors both tasks; when either completes or raises, the remaining task is cancelled and the session is shut down.

3.5.3. Batch Analysis Endpoint

`POST /analyze-file` accepts a multipart form in which each pipeline parameter is submitted as a separate named field (`ser_backend`, `window_s`, `hop_s`, `vad_threshold`, `emit_partials`, `whisper_temperature`, `stt_language`), together with the WAV file binary. The handler writes the upload to a temporary file, constructs a `SessionConfig` from the form fields, delegates to `RunOfflineAnalysis.execute()`, and returns a `BatchResult` containing the full `ser_live`, `ser_weighted`, and `utterances` lists. The temporary file is deleted in a `finally` block regardless of outcome. The batch endpoint is used by the React frontend for file upload analysis and by the evaluation scripts described in Chapter 5.

3.6. React Frontend Architecture

3.6.1. Technology Stack and State Management

The frontend is a React 18 single-page application written in TypeScript, built with Vite, styled with Tailwind CSS, and using Zustand for global state management. The Emotion Circumplex visualisation is implemented as a plain SVG component without D3. Audio playback is handled via the HTML5 `<audio>` element.

The Zustand store (`useAppStore`) is the single source of truth for all session data. It maintains the current operating mode (`idle | live | post`), the normalised event arrays `serWeightedEvents` and `enrichedUtterances`, a `partialHistory` list of all

partial utterances received (never pruned, used for sub-utterance playback sync), the `latestWeighted` event for live-mode display, the selected segment identifier, the audio player state, and a real-time microphone level used by the oscilloscope display.

SERLive events are deliberately not stored: they are consumed by the WebSocket hook and discarded after display, keeping memory usage bounded during long sessions. The store's default `hop_s` is 0.25 s (four SER updates per second), which differs from the backend default of 0.5 s; the frontend default is chosen for a more reactive live display.

All normalisation of raw backend labels (e.g., "anger" to "Anger") and probability maps is performed by `normalizeLabel()` and `normalizeProbs()` from `normalizeEmotion.ts` in the hooks and batch session logic before the store ingestion actions are called. Backend-specific labels without a counterpart in other schemes are preserved as distinct canonical names: `Boredom` (EmoDB-only), `Contempt` and `Surprise` (Whisper-only), and `Other` (Whisper-only). Since each session uses a single backend, only that backend's label set is active; no merging or collapsing across heterogeneous label sets occurs at runtime, and the circumplex draws an anchor only for labels the active backend can emit.

3.6.2. Page Structure and Component Hierarchy

The application has three top-level pages: `HomePage`, `SettingsPage`, and `SessionPage`, navigated via the `page` field in the Zustand store.

`HomePage` presents two mode cards: file analysis (which triggers a hidden file input accepting `.wav` files) and live mode (which opens a WebSocket connection and starts microphone capture). It also displays a compact configuration summary linking to `SettingsPage`.

`SettingsPage` renders a scrollable form with slider rows for window size, hop, VAD threshold, and Whisper temperature; radio groups for the SER backend (the three operational backends, EmoDB, Whisper categorical, and Whisper dimensional, together with the retired Hume adapter, retained for reproducibility) and STT language (Italian or English); and a toggle for partial transcription emission. The sample rate (48 kHz) and frame size (480 samples) are shown as read-only fields because they are fixed by the RNNoise denoiser.

`SessionPage` renders six zones stacked vertically: a session header bar (session label, elapsed timer, status badge); an emotional timeline strip (full-width coloured bar representing the emotional arc); summary key-performance-indicator (KPI) cards updated in real time; a conditional alert banner for rapid affect changes; a two-column content area

with a transcript column and an analysis column separated by a draggable resize handle; and a persistent docked audio player bar.

Segment selection propagates via the Zustand `selectSegment` action, which updates `selectedSegmentId` and triggers re-renders of the probability distribution panel and the Emotion Circumplex. The transcript column auto-scrolls to the latest utterance during live sessions.

3.6.3. Analysis Visualisations

The `EmotionCircumplex` component renders an SVG of 240×240 logical units. Three concentric reference circles (radii 100, 68, and 34) and two orthogonal axes divide the plane into four quadrants labelled JOY, ANGER, SADNESS, and CALM. The palette defines a fixed colour and coordinate scheme for the full label vocabulary, but at render time the component plots an anchor dot only for the emotions the active backend can actually emit, so the circumplex never shows a label that the selected model cannot produce. The displayed anchors are therefore exactly the active backend's label set (Section 4.1.2): the seven-class EmoDB backend contributes *Anger*, *Disgust*, *Fear*, *Happiness*, *Neutral*, *Sadness*, and *Boredom*, while the nine-class Whisper backend contributes the six shared labels plus *Contempt*, *Surprise*, and the residual catch-all *Other* (which shares the origin with *Neutral*, as it carries no distinct affective coordinate). Because each session uses a single backend, no purely orientational anchor is drawn; the low-arousal, positive-valence CALM quadrant retains its quadrant label but, as no deployed backend emits an emotion there, carries no anchor dot. A single circumplex and colour scheme thus adapt to any active backend, each displaying only the labels it actually emits. When a segment is selected, dashed contribution lines radiate from the origin towards each emotion whose probability exceeds 5%, with line width and opacity proportional to the probability. A crosshair marker indicates the blended position: when the dimensional backend is active and raw (V, A) coordinates are available, these are used directly; otherwise the position is reconstructed as the probability-weighted centroid of the canonical anchor points. The coordinate readout below the SVG indicates whether the position is model-reported or reconstructed.

The `ProbabilityDistribution` component shows a ranked list of horizontal bars, one per emotion, coloured by the canonical emotion colour. In live mode, when no segment is explicitly selected, the panel falls back to `latestWeighted` and animates bar widths continuously.

The `EmotionalTimelineStrip` renders the weighted event sequence as adjacent

coloured blocks spanning the full session duration. Each block's colour is computed by `interpolatedColor()`, a probability-weighted average of the canonical RGB colours of the emotions present in the frame's posterior. Clicking a block in post-session mode seeks the audio player to the nearest final utterance.

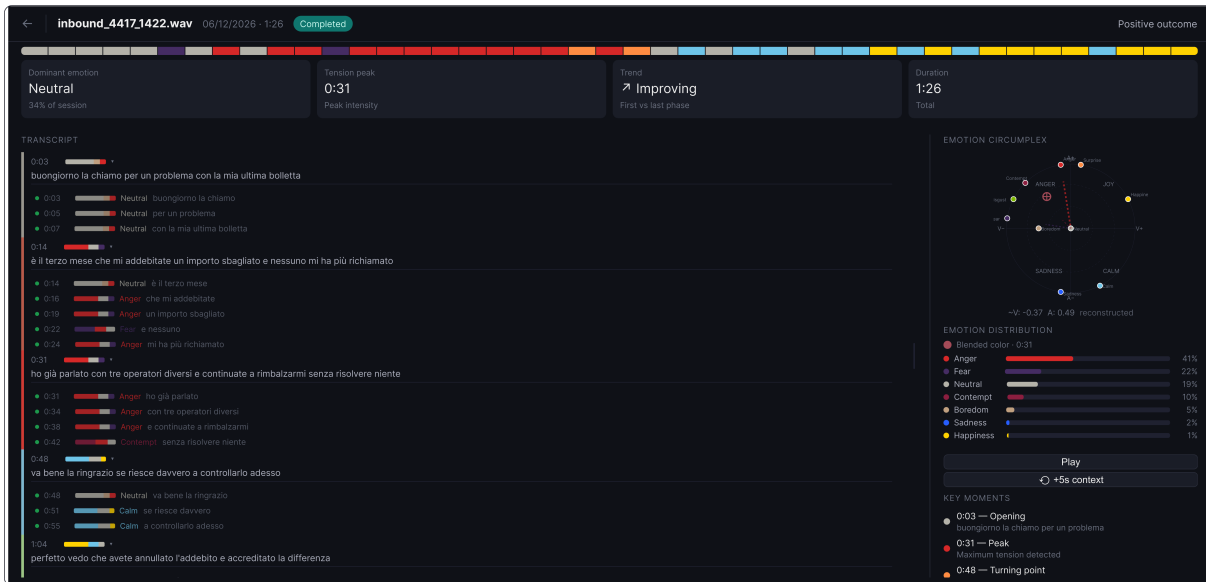


Figure 3.3: The **SessionPage** during a live call-centre session, top to bottom: header bar (label, timer, live badge); emotional timeline strip (the affective arc as an interpolated colour sequence); KPI cards; a two-column area pairing the transcript panel (left) with the analysis panel (right, holding the Emotion Circumplex and ranked Probability Distribution bars); and a docked audio player for post-session playback.

3.6.4. Session Hooks and Lifecycle

`useWebSocketSession` manages the live streaming lifecycle. On connect, it sends the **SessionConfig** JSON, then gates frame transmission on a `backendReadyRef` flag set when the server acknowledges `state="started"`. Incoming JSON events are dispatched to the appropriate store action based on their `type` field. On WebSocket error or close, the session status is set to `"stopped"` and the mode transitions to `post`; there is no automatic reconnection.

`useMicrophoneCapture` acquires the microphone via `getUserMedia` and creates an **AudioContext** at 48 kHz. An **AudioWorklet** processor (`PcmProcessor`) is injected at runtime as a Blob URL; the worklet buffers mono float32 input into 480-sample Int16 frames and posts each frame's **ArrayBuffer** to the main thread, where it is forwarded to the WebSocket. A passive **AnalyserNode** is attached in parallel for the oscilloscope

display and does not affect the capture path.

`useBatchSession` performs a multipart POST `/analyze-file` request with the selected WAV file and the current session configuration as individual form fields. After receiving the `BatchResult`, it creates an object URL for the audio file (enabling in-browser playback), then iterates `ser_weighted` and `utterances` to hydrate the store. The `ser_live` list from the batch result is not retained in the store.

`usePlaybackSync` runs a 300 ms interval during post-session playback that reads the audio element's `currentTime` and auto-selects the most recent partial transcript entry at or before the playback cursor, falling back to the active final utterance when no recent partial is available.

The `liveSessionBridge` module provides a module-level singleton that coordinates live session controls across page transitions. `HomePage` registers a stop callback that closes the WebSocket and stops microphone capture; `SessionPage` registers a resume callback that opens a fresh WebSocket and microphone without resetting the store, enabling seamless continuation of a live session after navigating to the session view.

3.7. Key Design Decisions

3.7.1. Structural Typing over Inheritance

Using Python Protocol for `ISERBackend` rather than an abstract base class eliminates a subtle coupling: if adapters inherited from an abstract base class defined in `app.domain`, they would carry a domain import in their namespace. An infrastructure module importing a domain module is not itself a violation, but it prevents the domain from being extracted to a standalone package without dragging adapters along. Structural typing decouples the port definition from the adapter implementations entirely [33].

3.7.2. PCM Broadcast Architecture

The SER and STT pipelines must consume the same PCM stream but operate at different granularities: SER advances by 0.5 s hops; STT consumes 10 ms frames. A naive solution would be to run them sequentially on buffered copies. Instead, `broadcast_copy()` fans out the preprocessed stream into two independent thread-safe queues, allowing SER and STT to run on separate threads without blocking each other. Each payload is copied to a C-contiguous numpy array before enqueueing, preventing aliasing between the two consumers.

3.7.3. Aggregator Concurrency Model

The aggregator dispatcher merges SER and STT events by timestamp rather than by wall-clock arrival order. SER events arrive earlier than the corresponding STT events (emotion is predicted on every 0.5s hop; transcription waits for phrase boundaries), so a simple FIFO merge would emit SER events before the STT events that overlap the same time window. The timestamp-ordered merge ensures that utterance enrichment at equation (3.4) always sees all SER windows whose support intersects the utterance, regardless of the latency difference between the two streams.

3.7.4. The `safe_until` Guard

A weighted frame at centre c_k is emitted only when the latest received SER event ends at or after $c_k + w_s$ (equation (3.2)). This guard is necessary because a weighted frame at c_k averages over SER windows in the interval $[c_k - h_s/2, c_k + h_s/2]$. Emitting before all windows have been received would yield an underestimated average, introducing artificial emotion transitions at the tail of the buffer. The price of the guard is a minimum latency of w_s (3.0s in the default configuration) for weighted frames in live mode, which is accepted as a design constraint. At end of stream the guard is bypassed to flush remaining frames.

3.7.5. Session Isolation and Stateful Recognisers

The KaldiRecognizer session isolation bug (Section 3.4.5) illustrates a general principle for stateful ML inference systems: any component that accumulates internal state (sample counters, decoding history, cache buffers) must provide an explicit reset mechanism that is called between sessions. The fix adopted here (reconstructing the recogniser object rather than resetting its fields) was preferred because the Vosk API does not expose a public `reset()` method, and relying on undocumented private state would create a maintenance burden. The trade-off is a small allocation cost per session finalisation, which is negligible compared to the model weight loading that dominates startup.

3.7.6. Dimensional Inversion as a Unified Representation

The `WhisperDimensionalSERAdapter` uses a dedicated regression checkpoint rather than reading logits from the categorical head of the same model, exercising the pipeline with a fundamentally different model family: a continuous valence–arousal regressor in place of a categorical classifier. Valence and arousal are psychological dimensions defined at the level of affective geometry, and the circumplex provides a single continuous space

in which any affective state can be located, including graded and mixed states that fall between the named categories [51]. The `_distances_to_probs()` conversion then maps this continuous output into the same categorical distribution the rest of the pipeline consumes, so that the aggregator, event bus, and visualisation layer operate unchanged. This two-stage approach (regression head to (V, A) , then geometric nearest-anchor mapping) is what makes a dimensional model a first-class backend of an otherwise categorical pipeline, and is the primary motivation for maintaining two separate Whisper adapters rather than replacing the categorical one.

4 | Original Contributions

The system described in Chapter 3 rests on a set of algorithmic and architectural choices whose design rationale and formal derivation are the subject of this chapter. Section 4.1 examines the decision to preserve full posterior emotion distributions end-to-end and the cost of argmax collapse. Section 4.2 formalises the weighted aggregation algorithm for both frame-level and utterance-level fusion. Section 4.3 derives the *distances-to-probabilities* inversion algorithm that converts continuous valence–arousal outputs to categorical probability distributions. Section 4.4 presents the engineering design decisions made to ensure operational robustness across deployment conditions, including the VAD gating strategy, window duration selection, and the parameters governing the dimensional backend conversion algorithm. Section 4.5 describes the interpretability-first visualisation design choices that make probabilistic outputs visible to the operator at every temporal granularity.

4.1. Probabilistic Emotion Representation

4.1.1. The Value of Full Posterior Preservation

All current commercial SER deployments reduce the model output to a single dominant emotion label before surfacing results to the operator. This reduction is architecturally convenient but operationally costly, because three distinct classes of information are irrecoverably discarded when a posterior distribution $\mathbf{p} \in \Delta^{K-1}$ is collapsed to its argmax $\hat{y} = \arg \max_e p_e$.

The first is *secondary affect information*. Natural speech frequently carries blended emotional states in which two or more emotions co-occur with comparable probability [58]. A caller expressing frustrated resignation may produce a distribution with substantial mass on both *Anger* and *Sadness*; the dominant label *Anger* discards the sadness component entirely. In call-centre quality monitoring, the secondary affect is often precisely the signal of clinical interest, distinguishing an escalating complaint from a contained expression of displeasure.

The second is *calibration information*. The confidence of the dominant prediction encodes how certain the model is about its output. A prediction of *Happiness* with $p = 0.95$ warrants a different operator response than the same label at $p = 0.35$. Collapsing to the label discards this signal, leaving the operator without the information needed to assess how much weight to assign to the system output [23, 53].

The third is *temporal trajectory information*. The full distribution at time t encodes whether the emotional state is strengthening, weakening, or transitioning relative to $t - h_s$. A gradual shift from high-Anger to high-Neutral mass is semantically different from an abrupt flip of the argmax label, even when both changes produce identical dominant-label sequences. This trajectory is visible in the evolving distribution but invisible to any argmax-based representation [41].

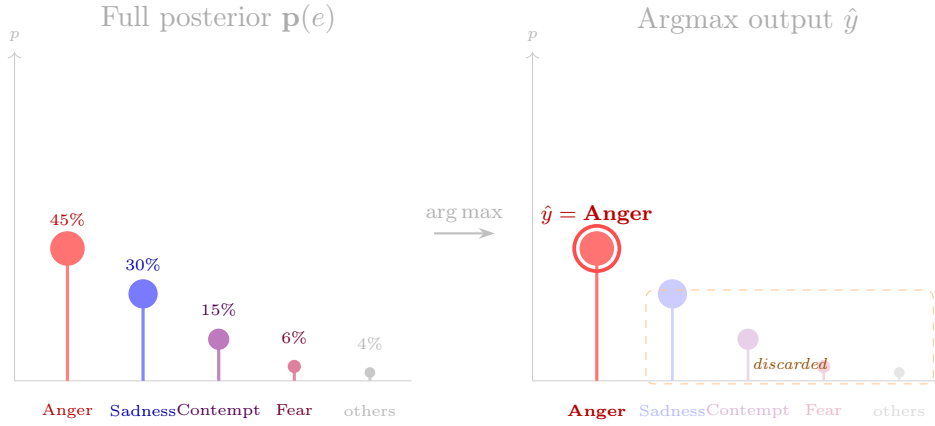


Figure 4.1: Argmax collapse: the full posterior $\mathbf{p}(e)$ (left) is reduced to the dominant class only (right); secondary affect mass, calibration confidence, and distributional trajectory are irrecoverably lost (dashed box).

4.1.2. End-to-End Distribution Preservation

The system is designed so that no irreversible reduction of the emotion distribution occurs at any intermediate pipeline stage. Every component that consumes or produces emotion information operates on the full probability vector $\mathbf{p} = (p_e)_{e \in \mathcal{E}}$, where $\mathcal{E}$ is the canonical label set of the active backend (seven classes for EmoDB, nine for the Whisper backends). The dominant label and confidence are derived from $\mathbf{p}$ only at presentation time, as display conveniences, never as inputs to downstream computation.

This principle propagates through three levels of temporal granularity. At the *window level*, each invocation of the SER backend returns a full distribution $\mathbf{p}^{(w)}$, stored in the `EmotionEvent` dataclass without reduction. At the *frame level*, the weighted aggregation

algorithm computes $\bar{\mathbf{p}}(c_k)$ as a vector average of window distributions, preserving secondary emotion mass. At the *utterance level*, the transcript enrichment algorithm applies the same overlap-weighted average, returning a full distribution rather than a label.

The frontend reflects this principle at every temporal granularity. The Emotion Circumplex renders the probability-weighted centroid of the circumplex anchors:

$$V_{\bar{\mathbf{p}}} = \sum_{e \in \mathcal{E}} \bar{p}_e \cdot V_e, \quad A_{\bar{\mathbf{p}}} = \sum_{e \in \mathcal{E}} \bar{p}_e \cdot A_e, \quad (4.1)$$

where (V_e, A_e) are the anchor coordinates of Table 2.1. This ensures that the rendered position in the circumplex reflects the weighted contribution of all active emotions, making mixed affective states geometrically visible rather than collapsing them to a single point.

4.1.3. Shannon Entropy as an Uncertainty Proxy

The Shannon entropy $H(\mathbf{p})$ introduced in Section 2.4.1 (equation (2.2), in bits) provides a theoretical characterisation of how concentrated or diffuse an emotion probability distribution is. Its maximum value is $\log_2 |\mathcal{E}|$ bits, reached for a uniform distribution, and zero for a point mass. Preserving the full distribution end-to-end is therefore more informative than reducing it to the dominant class: a frame with $H(\mathbf{p}) \approx 0$ bits carries an unambiguous prediction, whereas a frame near the maximum signals that no single emotion dominates and any argmax reduction would be misleading. This observation motivates the design choice of propagating $\mathbf{p}$ through all pipeline stages rather than collapsing it at any intermediate step.

4.2. Weighted Aggregation Algorithm

4.2.1. The Aggregation Problem

The sliding-window inference loop emits one **EmotionEvent** per hop interval h_s , each covering a window $[e_{\text{start}}, e_{\text{end}}]$ of duration w_s . Consecutive windows overlap by $w_s - h_s$ seconds. The raw event stream is dense, unenriched by transcript information, and unsuitable for direct presentation. The aggregation module transforms it into two higher-level representations:

- **Weighted frames:** a temporally regular grid of emotion estimates at resolution h_s , suitable for the timeline strip.
- **Enriched utterances:** segment-level estimates aligned with ASR output, suitable for

the transcript panel and evaluation scripts.

Both representations must preserve the full probability distribution, handle variable overlap between SER windows and the target interval, and be produced online without access to future events.

4.2.2. Frame Aggregation via Uniform Window Averaging

For a frame centred at $c_k = t_0 + k \cdot h_s$ (where t_0 is the grid origin aligned to the first received SER event), the candidate set of contributing SER events is:

$$W_k = \{ e \mid e_{\text{start}} < c_k + h_s/2 \text{ and } e_{\text{end}} > c_k - h_s/2 \}, \quad (4.2)$$

that is, all events whose window intersects the frame interval $[c_k - h_s/2, c_k + h_s/2]$. For example, with $w_s = 3.0\text{s}$ and $h_s = 0.5\text{s}$, at most seven consecutive SER windows can overlap any frame interval, since the range of qualifying start times spans $w_s + h_s = 3.5\text{s}$ and contains seven hop-aligned points. When $|W_k| > 4$, the four events whose centres lie closest to c_k are retained; when two events are equidistant from c_k , the more recent one (higher `end` timestamp) is preferred, so that the frame always incorporates the freshest available audio.

The frame distribution is computed as the uniform average of the contributing distributions (equation (3.3)). A uniform rather than overlap-weighted average is appropriate here because all events in W_k have comparable overlap with the frame interval by construction of equation (4.2). The frame carries `num_events_used`, the count of events actually used after the locality cap, which downstream consumers use to assess reliability: a frame with `num_events_used = 1` rests on a single SER window and is more susceptible to inference noise than one with `num_events_used = 4`.

A frame at c_k is withheld until the safe-until condition of equation (3.2) is satisfied, ensuring that no frame is emitted before all SER windows that could contribute to it have been received. The minimum output latency for weighted frames is therefore w_s (3.0 s in the default configuration) in live mode.

4.2.3. Utterance Enrichment via Overlap-Weighted Averaging

When a `TranscriptPiece` arrives from the STT stream with interval $[u_s, u_e]$, the utterance distribution is the overlap-weighted average of equation (3.4), applied to all SER events whose windows intersect $[u_s, u_e]$.

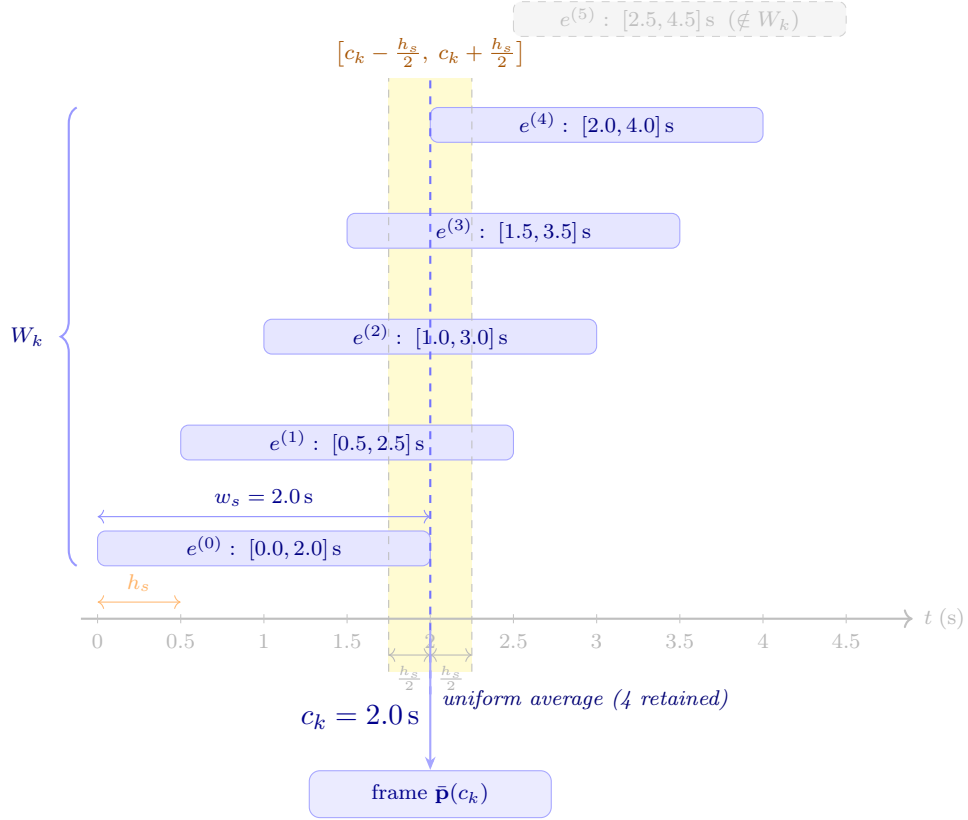


Figure 4.2: Frame aggregation at $c_k=2.0$ s ($w_s=2.0$ s illustrative, deployed default 3.0 s; $h_s=0.5$ s). The yellow strip is the frame interval $[c_k - h_s/2, c_k + h_s/2]$. Windows $e^{(0)}$ – $e^{(4)}$ intersect it ($|W_k|=5$); $e^{(5)}$ starts too late and is excluded. Since $|W_k| > 4$, the four windows closest to c_k are uniformly averaged into $\bar{\mathbf{p}}(c_k)$ (3.3).

Unlike frame aggregation, which uses uniform weights because candidate events have comparable overlap by construction (Section 4.2.2), utterance enrichment must handle variable overlap: an SER window covering 0.2 s of a 3.0 s utterance should carry less influence than one covering 1.8 s. The overlap-weighted formulation directly implements this intuition, giving each window influence proportional to the fraction of the utterance it informs.

When the dimensional backend is active, the (V, A) coordinates of each contributing event are averaged with the same overlap weights and stored in the `EnrichedUtterance`, enabling per-utterance Emotion Circumplex rendering in post-session mode.

4.3. Dimensional SER via Circumplex Inversion

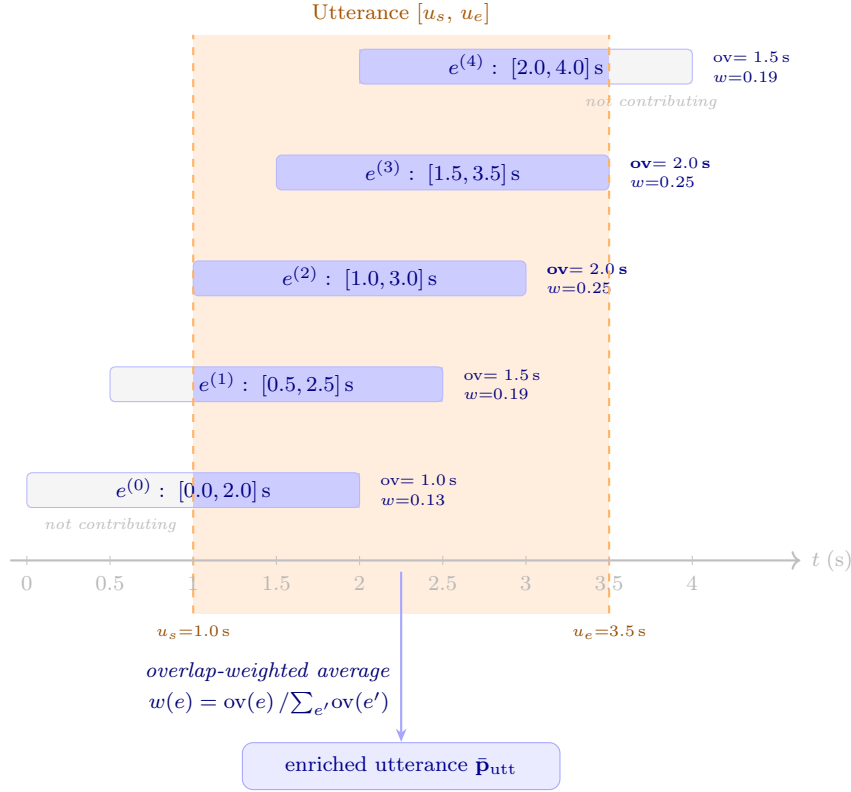


Figure 4.3: Utterance enrichment for $[u_s, u_e] = [1.0, 3.5] \text{ s}$ ($w_s = 2.0 \text{ s}$ illustrative, deployed default 3.0 s). Each window’s weight is its overlap with the utterance, normalised: $w(e) = \text{ov}(e) / \sum_{e'} \text{ov}(e')$ (3.4) (blue = overlap, gray = outside). The two fully contained windows ($e^{(2)}, e^{(3)}$, bold) reach the maximum weight $w = 0.25$. Weighting is thus non-uniform, unlike the equal-weight frame aggregation of Figure 4.2.

4.3.1. Motivation: Bridging Continuous and Categorical Representations

The dimensional backend is motivated by a fundamental property of Russell’s circumplex model: valence and arousal place every affective state in a single continuous plane, defined at the level of affective psychology rather than as membership in a fixed list of discrete classes [51]. This continuous representation captures graded and mixed affect that a discrete label set cannot express directly, and it lets a regression model report an affective position even for states that lie between the named categories [51, 52]. Integrating such a model requires reconciling its continuous output with the categorical probability representation that the aggregator, event bus, and visualisation layer share. The *distances-to-probabilities* algorithm of Section 4.3.2 provides exactly this reconciliation, mapping each (V, A) estimate into the common nine-class distribution and so making the dimensional model a structurally equal backend of the pipeline alongside the categorical

classifiers.

4.3.2. The Distances-to-Probabilities Algorithm

The algorithm receives the normalised valence–arousal coordinates $(v, a) \in [-1, 1]^2$ produced by the conversion of equation (3.6). The circumplex anchor set $\mathcal{A}$ is given by Table 2.1; the *Other* label is excluded from $\mathcal{A}$ because it has no geometric interpretation in the circumplex.

Step 1: Neutral probability via radius. The Euclidean radius $r = \sqrt{v^2 + a^2}$ measures the distance of the model output from the affective centre of the circumplex. A small radius indicates low-intensity, centrally positioned affect; a large radius indicates strong peripheral emotion. Neutral probability is set as

$$p_{\text{Neutral}} = \max(0, 1 - r \cdot s), \quad (4.3)$$

where s is a scale parameter (empirically $s = 2.0$). Setting $s > 1$ ensures that p_{Neutral} reaches zero before the unit-circle boundary ($r = 1/s = 0.5$), preventing strong peripheral predictions from carrying residual neutral mass.

A radius-based formulation is necessary rather than a distance-to-anchor approach. The origin is equidistant from all peripheral anchors, so a distance-based formula would assign uniformly high probability to every peripheral emotion when the model output is near the centre, the opposite of the intended semantics: a near-centre output should concentrate mass on *Neutral*, not distribute it equally among all peripheral emotions.

Step 2: Peripheral softmax. The remaining probability mass $1 - p_{\text{Neutral}}$ is distributed among the seven peripheral emotions by a temperature-scaled softmax over negative Euclidean distances:

$$p_e = (1 - p_{\text{Neutral}}) \cdot \frac{\exp(-d_e/\tau)}{\sum_{e'} \exp(-d_{e'}/\tau)}, \quad (4.4)$$

where $d_e = \|(v, a) - (V_e, A_e)\|_2$ is the Euclidean distance from the model output to anchor e , and τ is a temperature parameter ($\tau = 0.45$).

Temperature selection. The value $\tau = 0.45$ was selected empirically to satisfy two geometric criteria: (i) at any circumplex anchor point the corresponding emotion should be the unique mode of the distribution, concentrated well above the uniform baseline; and (ii) at the midpoint between two adjacent anchors the distribution should be bimodal over those two emotions rather than concentrating mass on the geometric centroid be-

tween them. A lower temperature concentrates all peripheral mass on the nearest anchor, degenerating to Voronoi assignment and discarding distributional information; a higher temperature produces near-uniform peripheral distributions even for confident predictions. The value $\tau = 0.45$ produces a balance in which the nearest anchor is the dominant mode and takes the largest share of the peripheral mass, with the remainder distributed to geometrically adjacent anchors in proportion to their inverse distance. The sensitivity of the output distribution to τ across the range $[0.10, 3.00]$ is analysed in Chapter 5.

Bimodal distributions and circumplex geometry. A key semantic property of the algorithm is that predictions equidistant between two adjacent anchors yield bimodal distributions over those two emotions rather than concentrating on the geometric centroid between them. A model output midway between *Anger* (99°) and *Surprise* (81°) yields roughly equal mass on both, not a distribution peaked on *Contempt* (135°). This correctly represents uncertainty between two adjacent affective states and is a direct consequence of the distance-to-softmax formulation, which penalises distance to each anchor independently rather than computing a centroid first.

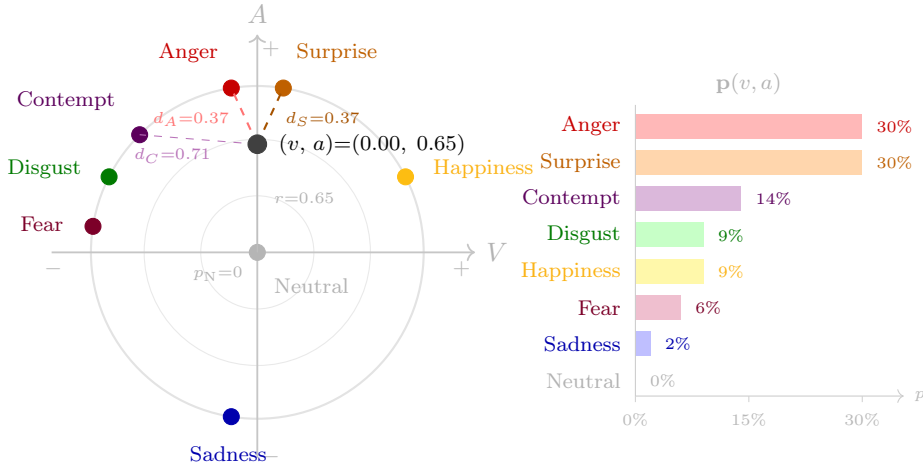


Figure 4.4: Distances-to-probabilities algorithm (Section 4.3.2) for $(v, a) = (0.00, 0.65)$, equidistant from Anger and Surprise. **Left:** $r = 0.65 > 0.5$ gives $p_N = 0$; the three labelled distances drive the softmax ($\tau = 0.45$). **Right:** since $d_A = d_S = 0.37$, the distribution is bimodal, Anger and Surprise each receive 30%.

4.4. Heuristic Design Decisions for Operational Robustness

A deployable streaming pipeline must behave reliably and transparently across a range of operating conditions, different speakers, recording chains, and session lengths. The

system achieves this through explicit engineering design decisions: concrete choices about audio segmentation, voice activity detection, and dimensional conversion parameters that collectively ensure stable, predictable pipeline behaviour, each grounded in the empirical characterisation of Chapter 5. The following subsection documents these decisions and the rationale behind each.

4.4.1. Heuristic Design Strategies

VAD gating. The VAD gate (Section 3.2.3) ensures that only speech-containing audio reaches the SER backends. By discarding SER inference on windows where the mean VAD probability falls below $\theta_{\text{vad}} = 0.5$, the gate prevents silence, background noise, and hold music from entering the inference pipeline. The threshold involves a calibrated trade-off: a higher value provides a stricter speech filter but may discard valid short-duration speech segments; a lower value is more permissive but admits borderline non-speech audio. The recommended value $\theta_{\text{vad}} = 0.5$ is the natural decision boundary, the probability midpoint at which RNNoise judges speech the more likely hypothesis, and the sweep in Chapter 5 confirms it sits on the voiced-coverage plateau while suppressing the spurious non-speech inferences, rather than being the optimum of a single measured cost.

Dimensional backend integration. The inclusion of the dimensional backend demonstrates that the modular architecture successfully accommodates a fundamentally different model family alongside the categorical backends. Categorical classifiers produce probability distributions over discrete label sets; the Whisper dimensional backend produces continuous valence–arousal coordinates in Russell’s circumplex plane [51], an output space that is structurally incompatible with the rest of the pipeline without an explicit conversion. The *distances-to-probabilities* algorithm (Section 4.3.2) performs that conversion, mapping (V, A) coordinates to a categorical probability distribution that the aggregator, event bus, and visualisation layer can consume without modification. The dimensional backend therefore serves as a concrete validation that the port interface and common probabilistic representation support not only different categorical models but also entirely different SER paradigms, keeping the platform open to future dimensional implementations.

Neutral scale and peripheral temperature as operational parameters. The parameters s and τ in equations (4.3) and (4.4) jointly control the sharpness and neutral affinity of the dimensional backend output. Increasing s raises the neutral probability for moderate-radius predictions, making the backend more conservative on ambiguous audio; decreasing τ concentrates peripheral mass on the nearest anchor, sharpening the

distribution. Because s and τ jointly govern the neutral–peripheral mass balance, they are treated as a coupled pair rather than independently tuned scalars. They are fixed adapter-level design constants ($s = 2.0$, $\tau = 0.45$ in the dimensional adapter) rather than per-session runtime parameters; they provide a practical mechanism for re-tuning the operating point of the dimensional backend at configuration time, without retraining the underlying model. The sensitivity of the algorithm output to this parameter pair is characterised in Chapter 5; the joint empirical sensitivity on real speech is identified as a direction for future work.

4.5. Interpretability-First Visualisation

This contribution focuses on the level addressed to operators: the design choices that make uncertainty and distributional information consistently visible at every temporal granularity, in contrast to commercial tools that surface only a single dominant label or an opaque scalar score.

4.5.1. The Probability-Weighted Circumplex

Commercial SER interfaces that do expose a circumplex representation typically plot the argmax emotion as a single point, discarding all information about the surrounding distribution. The Emotion Circumplex component described in Section 3.6.3 instead renders the *probability-weighted centroid* of all active emotion anchors (equation (4.1)), so that a bimodal distribution between *Anger* and *Sadness* produces a crosshair in the upper-left quadrant rather than snapping to a single anchor. Dashed contribution lines, with width proportional to probability, radiate from the origin to each emotion exceeding a 5% threshold, making the distributional spread directly readable by the operator. When the dimensional backend is active, the raw (V, A) model output is used instead of the centroid reconstruction, and the readout below the SVG distinguishes the two cases, maintaining epistemic transparency.

4.5.2. The Interpolated-Colour Emotional Timeline

A per-frame timeline that maps emotion to a single canonical colour (the argmax emotion’s hue) cannot represent mixed or transitional affective states.

The `EmotionalTimelineStrip` instead computes each frame’s colour as the probability-weighted average of the canonical RGB values of the emotions present in the posterior (`interpolatedColor()`): a frame with 40% Anger and 40% Sadness renders as a muted

purple rather than a pure red or blue. This encoding makes frames whose probability mass is split across emotions of opposing hue blend toward a desaturated grey, while concentrated frames render as a saturated canonical hue, giving the operator an immediate qualitative signal of distributional spread without requiring a separate numeric readout.

4.5.3. Transcript Panel with Per-Utterance Uncertainty Annotation

Each utterance in the transcript panel carries the overlap-weighted probability distribution computed by the aggregator (Section 4.2.3). Selecting an utterance updates both the Probability Distribution bars and the Emotion Circumplex to reflect that specific segment's affective estimate, enabling retrospective inspection at utterance granularity. Partial transcript entries, which accumulate word-by-word during live sessions, are synchronised to audio playback in post-session mode via `usePlaybackSync`, so that the operator can replay any segment and observe the evolving emotional annotation in real time. All three visualisation components are driven by the same normalised probability vectors that flow through the pipeline; no reduction to a dominant label occurs at any display stage.

5 | Experimental Evaluation

This chapter evaluates the five contributions formalised in Chapter 4, asking of each the same question: does it deliver, on real speech and under the deterministic geometry of its own definition, what it claims to deliver? Section 5.1 introduces the evaluation material, the metrics, and the software configuration shared by all experiments. Section 5.2 validates the distances-to-probabilities algorithm (C3) through a fully synthetic geometric test suite that requires neither audio nor a trained model. Section 5.3 demonstrates the modular platform (C1) by processing the same session with three interchangeable backends through an identical code path and by characterising the real-time behaviour of the pipeline. Section 5.4 validates the probabilistic temporal aggregation algorithm (C2) on all four sessions of the evaluation corpus. Section 5.5 consolidates the heuristic design decisions (C4) into empirically grounded operating points for the five configuration parameters of the system. Section 5.6 verifies that the interpretability-first visualisation components (C5) are faithful renderings of the underlying probability distributions. Section 5.7 synthesises the results by contribution, and Section 5.8 delimits the scope of the claims.

5.1. Experimental Setup

5.1.1. Evaluation Material

Two complementary kinds of material drive the evaluation: real speech, used to exercise the full pipeline under realistic conditions, and synthetic input, used where a contribution is defined by deterministic geometry or where ground-truth structure must be known exactly.

The real-speech material is a set of four Italian-language sessions selected ad hoc for validation, drawn from two speakers (S3 and S6) reading a call-centre-style text under the Happiness, Neutral, and Stress affective conditions. The recordings are mono and 27.3–37.8 s long, the conversational-turn regime a monitoring deployment actually observes and the one that exercises the full temporal machinery of the pipeline: the sliding-window ag-

gregation, the streaming cadence, the `safe_until` release, and the utterance-enrichment workflow. Each experiment uses the sessions its question requires: all four enter the aggregation study of Section 5.4, which characterises temporal stability and therefore benefits from every available stream, while the three-session subset spanning one recording per affective condition serves the parameter sweeps of Section 5.5 and the circumplex, timeline, and transcript analyses of Section 5.6, where one clear exemplar per condition is what the comparison calls for.

The synthetic material comprises two elements. The geometric test suite of Section 5.2 evaluates the distances-to-probabilities algorithm directly on valence–arousal coordinates, so that every expected property follows analytically from the construction of Section 4.3.2. The VAD experiment of Section 5.5.1 uses a composite session obtained by embedding the 32.3s Neutral-condition session of speaker S6 between two 6s stretches of digital silence, so that the boundary between voiced and unvoiced material is known exactly and every inference event can be attributed unambiguously to one region or the other.

Short single-sentence acted clips such as CREMA-D [11] and TESS [47], two to three seconds long, fall below the $w_s = 3\text{s}$ analysis window and so would not drive these temporal stages, while EmoDB [8] and MSP-Podcast [37] serve as the training provenance of the backends; the conversational sessions above are therefore the material that matches exactly what this chapter sets out to measure.

5.1.2. Evaluation Metrics

Throughout the chapter, distributions are compared with the *Jensen–Shannon divergence* (JSD), expressed in bits: a symmetric, bounded measure of the difference between two probability vectors, equal to zero for identical distributions and to one bit for distributions with disjoint support. The *Shannon entropy* $H(\mathbf{p})$ introduced in Section 4.1.3 serves as the scalar proxy for prediction uncertainty: in the nine-class label space its maximum is $\log_2 9 \approx 3.17$ bits.

The stability of a temporal stream is measured by its *inter-frame jitter*, defined as the mean JSD between consecutive distributions of the stream, and by its *dominant-label flip rate*, the number of changes of the most probable emotion per minute; both quantify the visual restlessness an operator would experience. Temporal responsiveness is measured by the *first-frame release latency*, the amount of audio that must have been processed before the aggregation stage releases its first weighted frame, and computational feasibility by the *real-time factor* (RTF), the ratio of processing wall time to audio duration, which must remain below one for live operation. Visualisation fidelity (Section 5.6) is

assessed geometrically, on the closed-form rendering functions themselves: the radial and angular transfer of the circumplex centroid map, characterised by per-azimuth Spearman monotonicity and angular deviation, and the Pearson correlation between the interpolated colour deviation and the secondary-emotion mass.

5.1.3. Software Configuration and Reproducibility

All experiments run the production code path: audio enters through the ingestion stage, is denoised and assessed by the RNNoise VAD [61], reaches the SER backend through the sliding-window inference stage, and is fused by the aggregation stage of Section 4.2, while Vosk [2] provides the Italian transcription consumed by the utterance-enrichment workflow. Unless a sweep varies them explicitly, the configuration parameters take their default values $w_s = 3.0$ s, $h_s = 0.5$ s, and $\theta_{vad} = 0.5$, which are the recommended operating points established in Section 5.5. The three backends are those of Chapter 3: the EmoDB BiLSTM classifier [8] with seven categorical classes, the Whisper-large-v3 categorical model fine-tuned on MSP-Podcast [20, 37, 50] with nine classes, and the Whisper-large-v3 dimensional model whose valence–arousal output is converted to the same nine-class space by the distances-to-probabilities algorithm of Section 4.3.2.

Experiments were executed on an Apple M1 Pro laptop with 16 GB of unified memory, with neural inference on the Metal Performance Shaders (MPS) backend (PyTorch 2.2, TensorFlow 2.19, `transformers` 4.29.2, Python 3.11), consumer hardware that matches the deployment assumption of an on-premise, single-machine installation. The `transformers` version is pinned because later releases alter the Whisper feature-extraction interface the backends rely on; the full set of pinned dependencies is committed with the accompanying code. Every experiment is implemented as one of five self-contained Jupyter notebooks (01–05, one per experimental section of this chapter) that regenerate all figures, tables, and CSV artefacts of the chapter in a single execution, given the evaluation corpus and the model checkpoints provisioned as documented in the accompanying code. The notebooks seed the NumPy generators used by the synthetic sweeps, so the geometry-driven results are bit-reproducible; the neural-inference figures depend on the non-deterministic MPS backend and are therefore reproducible to within small floating-point variation rather than exactly, which does not affect any of the qualitative conclusions or the reported precision.

5.2. Geometric Validation of the Distances-to-Probabilities Algorithm

The distances-to-probabilities algorithm (C3) makes the dimensional backend structurally compatible with the categorical pipeline by mapping each valence–arousal estimate to a nine-class probability distribution. Its correctness is a geometric property of the mapping itself, independent of any audio or model: Section 4.3.2 defines the algorithm by the behaviour it must exhibit at the circumplex anchors, at the centre, between anchors, and across the plane. The four experiments of this section verify each of these behaviours in turn, evaluating the production implementation directly on synthetic coordinates. Because the algorithm is a deterministic geometric map, this synthetic evaluation is exhaustive rather than illustrative: it certifies the behaviour over the entire valence–arousal plane, every state the dimensional model could ever produce, and confirms that the production code realises the specified geometry exactly, a guarantee no sampled audio set could provide. The sensitivity of the mapping to its two parameters τ and s is part of the configuration study and is reported in Section 5.5.4. A preliminary equivalence check confirmed that the parametric re-implementation used by the sweeps reproduces the production code to within numerical precision (maximum deviation 2.2×10^{-16} across 500 random points).

5.2.1. Anchor Concentration

When the model output coincides exactly with a circumplex anchor, the strongest dimensional prediction the backend can produce for that emotion, the resulting distribution must rank that emotion first and concentrate mass on it, with the residual routed to the geometric neighbourhood rather than scattered across the label space. Table 5.1 reports the algorithm output at each of the seven peripheral anchors of Table 2.1. At every anchor the corresponding emotion is the unique mode of the distribution, and its mass exceeds the uniform peripheral baseline of $1/7$, the share a blind assignment over the seven peripheral emotions would give, by a factor between 3.1 and 6.2 (mean 4.1), so that the prediction is markedly concentrated against the only neutral reference available. The second-ranked emotion is in every case one of the two angularly adjacent anchors, and the anchor together with this nearest neighbour captures 78% of the probability mass on average, confirming that the residual encodes the local circumplex neighbourhood rather than diffusing across unrelated emotions. The spread of the self-mass values reflects the anchor geometry of the circumplex itself, and is a feature rather than a limitation: *Sadness*, isolated in the lower half-plane, retains 0.883 of the mass at a $6.2\times$ concentration,

whereas the anchors crowded into the second quadrant, *Contempt*, *Disgust*, and *Anger*, separated by as little as 18° , share a correspondingly larger fraction with their immediate neighbours, exactly the graded behaviour the operating point $\tau = 0.45$ is chosen to provide and which the midpoint experiment of Section 5.2.3 requires. Figure 5.1 visualises both effects.

Anchor	p_{self}	Rank	vs 1/7	Runner-up	$p_{2\text{nd}}$
Anger	0.489	1	$3.4\times$	Surprise	0.244
Contempt	0.458	1	$3.2\times$	Disgust	0.228
Disgust	0.444	1	$3.1\times$	Contempt	0.222
Fear	0.519	1	$3.6\times$	Disgust	0.259
Happiness	0.777	1	$5.4\times$	Surprise	0.103
Sadness	0.883	1	$6.2\times$	Fear	0.038
Surprise	0.528	1	$3.7\times$	Anger	0.264

Table 5.1: Algorithm output at the seven peripheral circumplex anchors: mass assigned to the anchor’s own emotion, its rank, its concentration relative to the uniform peripheral baseline $1/7$, and the second-ranked emotion (always an angularly adjacent anchor).

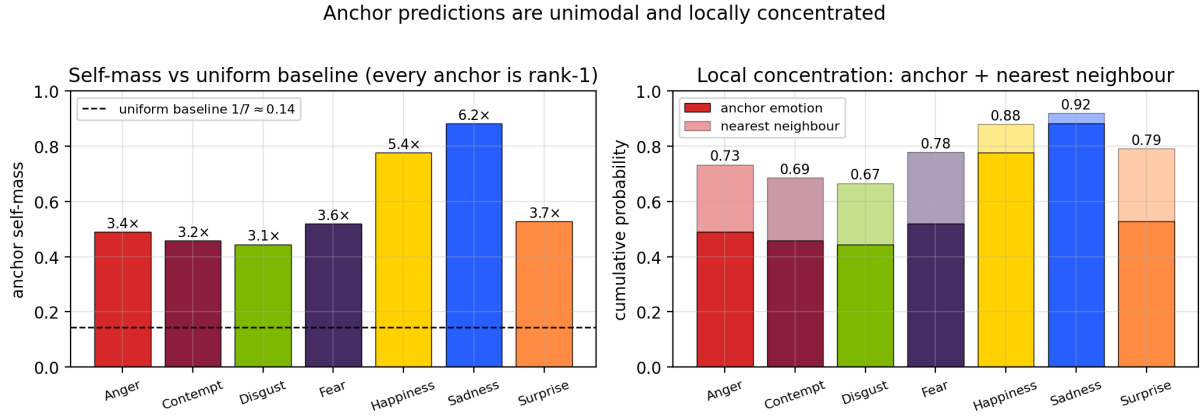


Figure 5.1: Predictions at each peripheral anchor. Left: anchor self-mass against the uniform baseline $1/7$ (dashed), every anchor ranked first. Right: anchor self-mass and nearest neighbour stacked.

5.2.2. Neutral Radial Behaviour

The Neutral probability is constructed in equation (4.3) as a function of the radius alone, decaying linearly from one at the origin to zero at $r = 1/s = 0.5$. Sweeping the radius from 0 to 1 along sixteen azimuths, Figure 5.2 shows that the measured Neutral mass coincides with the analytic curve $\max(0, 1 - 2r)$ to within floating-point precision

(maximum deviation 1.1×10^{-16}) and that the maximum spread across azimuths is of the same order, confirming exact direction invariance. The dominant label hands over from *Neutral* to the peripheral emotion of the swept direction at radii between 0.365 and 0.415, comfortably inside the extinction radius, so that low-intensity model outputs are read as Neutral, intermediate outputs blend Neutral with the emerging peripheral emotion, and strong outputs carry no residual Neutral mass. The entropy curves peak precisely in this handover region, where the distribution legitimately spans Neutral and several peripheral candidates, and decrease towards both the confident centre and the confident periphery: the uncertainty signal an interpretable system should expose.

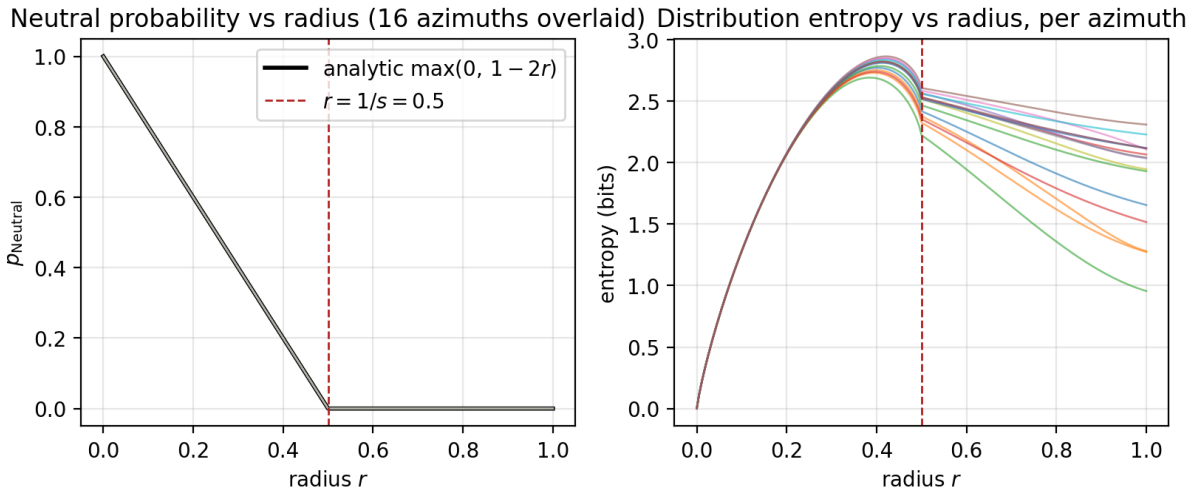


Figure 5.2: Left: Neutral probability vs radius along sixteen azimuths (grey) against the analytic curve (black), extinguished at $r = 1/s = 0.5$. Right: entropy along the same sweeps, peaking in the Neutral-to-peripheral handover.

5.2.3. Bimodality at Anchor Midpoints

A model output equidistant from two adjacent anchors expresses genuine ambiguity between two emotions, and Section 4.3.2 requires the algorithm to render it as a bimodal distribution over exactly those two labels rather than as a spurious commitment to whichever third emotion lies near their geometric centroid. Evaluating the algorithm at the chord midpoint of each of the seven pairs of angularly adjacent anchors, the two flanking emotions emerge as the two modes of the distribution in all seven cases, with perfectly balanced masses (ratio 1.000, as the midpoint construction demands by symmetry) and with a combined share of the peripheral mass between 0.545 and 0.758 (mean 0.647), the remainder flowing to the next-adjacent anchors in proportion to their distance. Figure 5.3 reports the two modal masses for each pair. The widest pair, *Sadness–Happiness*,

whose anchors subtend 126° , places its midpoint at radius 0.454, inside the Neutral extinction radius; the algorithm accordingly assigns it a small Neutral component of 0.091, a geometrically faithful reading of a point that genuinely approaches the affective centre.

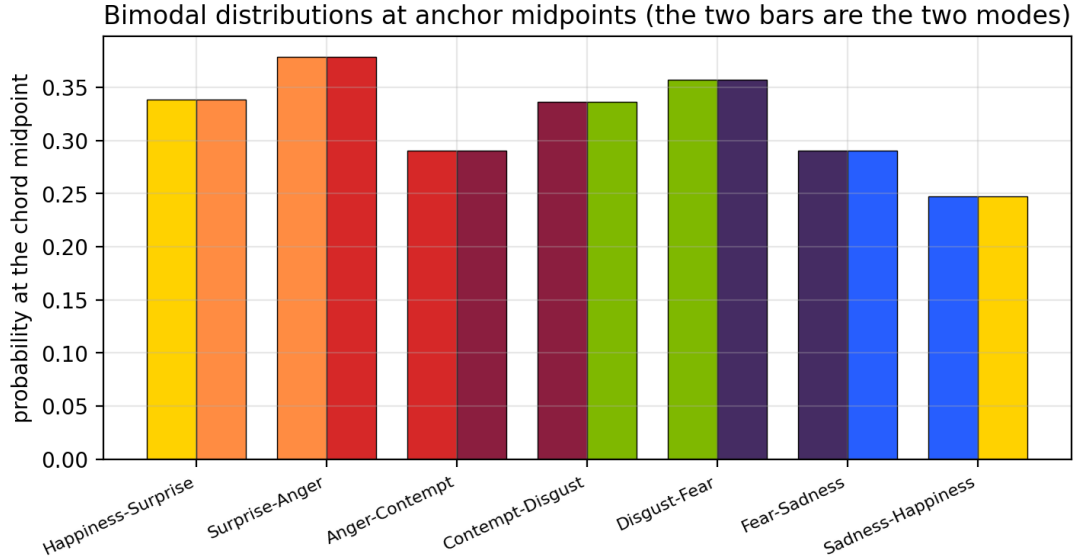


Figure 5.3: Distributions at the chord midpoints of the seven adjacent-anchor pairs: the two bars are the two balanced modes of each output distribution.

5.2.4. Full-Plane Coherence and Smoothness

The pointwise properties verified so far are completed by a global characterisation on a 201×201 grid covering the unit disc. Figure 5.4 shows the dominant-emotion map, the entropy map, and the Neutral-probability map produced by the algorithm. The dominant-emotion map reproduces the circumplex layout of Russell [51] as instantiated by the anchor set of Jing et al. [30]: a central Neutral disc surrounded by angular sectors in exact anchor order, with sector boundaries falling midway between adjacent anchors. The entropy map is lowest near the anchors and the centre and highest along sector boundaries and the Neutral handover annulus, so that uncertainty is greatest exactly where ambiguity is geometrically real. Smoothness, the property that protects the downstream aggregation and visualisation stages from spurious discontinuities, is quantified by the JSD between the distributions at every pair of neighbouring grid points: across the disc the mean neighbour JSD is 1.4×10^{-4} bits and the maximum is 1.0×10^{-2} bits at a grid step of 0.01, so that no transition between adjacent cells exceeds one hundredth of a bit and the mapping is continuous at the resolution any upstream model could meaningfully resolve.

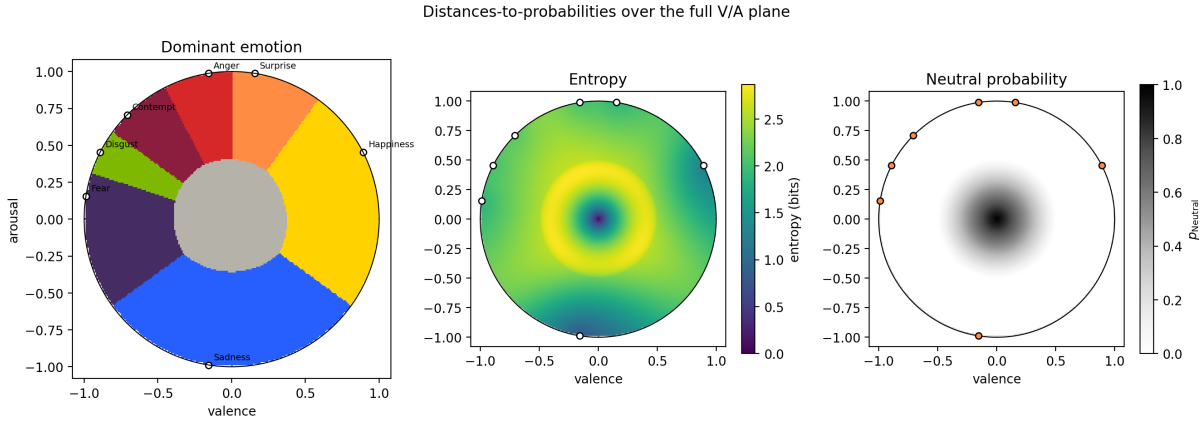


Figure 5.4: Global behaviour on the valence–arousal disc: dominant emotion (left), entropy (centre), and Neutral probability (right); unit circle and the seven anchors overlaid.

5.2.5. Summary

Taken together, the four experiments establish that the distances-to-probabilities algorithm satisfies every geometric property its definition requires, deterministically and independently of any audio or trained model. Table 5.2 consolidates the evidence as a checklist: the anchor emotion is the unique mode at all seven anchors and is concentrated several times above the uniform baseline, the Neutral mass follows its analytic radial law to numerical precision and in a direction-invariant manner, the adjacent-anchor mid-points are perfectly balanced bimodal distributions over the two flanking emotions, and the mapping is globally smooth, with neighbouring grid points differing by a negligible Jensen–Shannon divergence. The two parameters of the conversion are well posed: the operating temperature $\tau = 0.45$ lies inside the admissible region identified in Section 5.5.4, and the Neutral extinction radius tracks $1/s$ across the explored range. The algorithm therefore provides the dimensional backend with a semantically coherent categorical output and makes it a structurally equal citizen of the pipeline, a property confirmed in operation by the interchangeability demonstration of the following section.

5.3. Platform Demonstration: Interchangeable Backends in Real Time

The platform contribution (C1) claims that a complete acoustic SER pipeline, ingestion, denoising, VAD gating, sliding-window inference, probabilistic aggregation, and transcription enrichment, runs in real time on commodity hardware, and that the SER model is a session-time choice rather than an architectural commitment. Both claims are demon-

Property		Criterion	Measured
Unimodality at anchors		the anchor emotion is the unique mode (rank 1) at all 7 anchors	7/7 rank-1
Concentration at anchors		anchor self-mass exceeds the uniform peripheral baseline (1/7)	$4.1\times$ on average (min $3.1\times$)
Semantic residual		the runner-up at every anchor is an angularly adjacent anchor	anchor+neighbour = 78% of the mass
Neutral radial law		measured Neutral mass matches the analytic $\max(0, 1 - sr)$, direction-invariant	max dev. $1e-16$ bits
Midpoint bi-modality		adjacent-anchor midpoints are bimodal over exactly the two flanking emotions	7/7, balance 1.000
Global smoothness		neighbouring grid points differ by a negligible JSD (no discontinuities)	mean $1.4e-04$ bits
Parameter τ well-posed		operating $\tau = 0.45$ lies inside the admissible region	admissible [0.18, 0.55]
Parameter s well-posed		Neutral extinction radius matches $1/s$ across the explored range	max dev. 0.001

Table 5.2: Scorecard of the distances-to-probabilities algorithm: each required geometric property, its acceptance criterion, and the measured value that satisfies it.

strated on a Happiness-condition session (27.3s), processed end-to-end by the production offline-analysis use case with each of the three backends injected behind the `ISERBackend` port, under the identical temporal configuration and without any change to the pipeline code.

5.3.1. Backend Interchangeability

The backend is selected by a single configuration field at session construction, and no other line of the pipeline changes; everything that is not the acoustic model itself is invariant under the swap. Table 5.3 makes this division explicit, listing on the left the components that remain identical across the three runs, the entire ingestion, denoising, gating, inference-driving, aggregation, enrichment, and rendering machinery, and on the right the few aspects the injected backend alone determines. The consequence is visible in Table 5.4: the pipeline emits exactly the same event structure for all three backends, 42 live

inference events, 48 weighted frames on the same 0.5 s grid, and two enriched utterances, because the temporal behaviour of the system is owned by the pipeline stages, not by the model. Swapping the backend changes only the label space (seven EmoDB classes against nine MSP-Podcast classes), while every downstream consumer, from the aggregator to the visualisation layer, operates unchanged. Figure 5.5 makes this structural invariance visual: the weighted-frame timestamps of the three backends fall on exactly the same 0.5 s grid, and the three event counts coincide. The emotional content each backend reports is, by contrast, a model-specific matter that the platform transports faithfully rather than flattening: the mean output entropy ranges from 0.98 bits for the concentrated BiLSTM posterior (of a possible $\log_2 7 \approx 2.81$) to 2.78 and 2.73 bits for the broader Whisper backends (of a possible $\log_2 9 \approx 3.17$), a difference in classifier character that survives intact because every stage preserves the full distribution end-to-end (Section 4.1).

Invariant across all three backends	Determined by the injected backend
Audio ingestion and resampling	Acoustic model and its weights
RNNoise denoising and VAD gating	Label space (7 vs 9 classes)
Sliding-window inference stage	Per-window inference latency
Overlap-weighted temporal aggregation	Output distribution character (entropy)
Utterance enrichment	
Event schema (live / weighted / utterance)	
Temporal grid (0.5 s hop, 48 frames here)	
Frontend rendering	

Table 5.3: The backend swap in terms of what it leaves invariant and what it determines: selecting a different acoustic model through a single configuration field changes only the model and its immediate consequences, while the entire surrounding pipeline is reused unchanged.

Backend	Classes	Live	Frames	Load (s)	Latency (ms)	RTF	Speedup
EmoDB BiLSTM	7	42	48	5.7	78 ± 20	0.43	2.3×
Whisper categorical	9	42	48	31.3	427 ± 29	0.75	1.3×
Whisper dimensional	9	42	48	27.8	427 ± 20	0.74	1.3×

Table 5.4: The same evaluation session processed by the three backends through the identical pipeline code path: label-space size, live events and weighted frames emitted, model load time, per-window inference latency (mean $\pm$ standard deviation over fifteen repetitions), end-to-end real-time factor, and the equivalent speedup over real time ($1/\text{RTF}$).

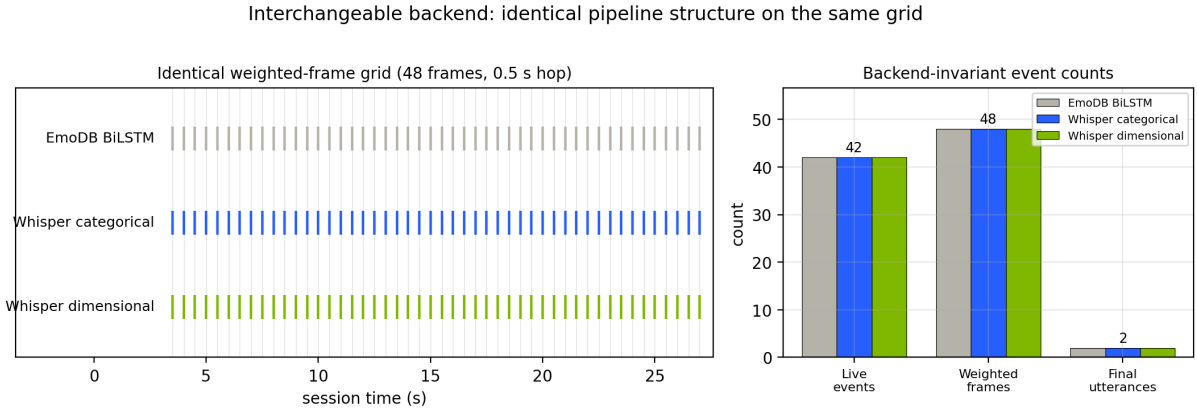


Figure 5.5: Structural invariance under the backend swap. Left: weighted-frame timestamps of the three backends on the same 0.5 s grid. Right: live-event, weighted-frame, and final-utterance counts identical across backends.

5.3.2. Real-Time Performance

Real-time operation imposes two budgets: each window inference must complete within one hop ($h_s = 0.5\text{s}$), and the pipeline as a whole must process audio at least as fast as it arrives. Figure 5.6 reports both measurements. The per-window inference latency, measured over fifteen repetitions of a 3 s window after warm-up, is 78 ms for the BiLSTM and approximately 427 ms for both Whisper-large-v3 backends, comfortably within the 500 ms hop budget; the end-to-end real-time factor, measured on a single full-session pass of the SER pipeline, is 0.43 for the BiLSTM and 0.74–0.75 for both Whisper backends, equivalently a processing speed of $2.3\times$ real time for the lightweight backend and $1.3\times$ for the large ones. Even the 1.5-billion-parameter models therefore sustain live operation on a consumer laptop while processing audio roughly 1.3 times as fast as it arrives. This factor measures the SER pipeline alone; the concurrent Vosk transcription runs on a separate thread, so its cost is not included in the figure and the $1.3\times$ margin should be read as the headroom that co-resident workload, including transcription, must fit within rather than as a guaranteed surplus. As a single-machine, single-run measurement on one laptop it characterises feasibility on the target hardware rather than a hardware-independent bound.

5.3.3. Streaming Cadence and Release Latency

The temporal behaviour of the output stream is governed entirely by the two streaming parameters: the hop h_s fixes the steady-state cadence at $1/h_s$ frames per second, while the window w_s determines, through the `safe_until` guard of Section 4.2.2, how long the first

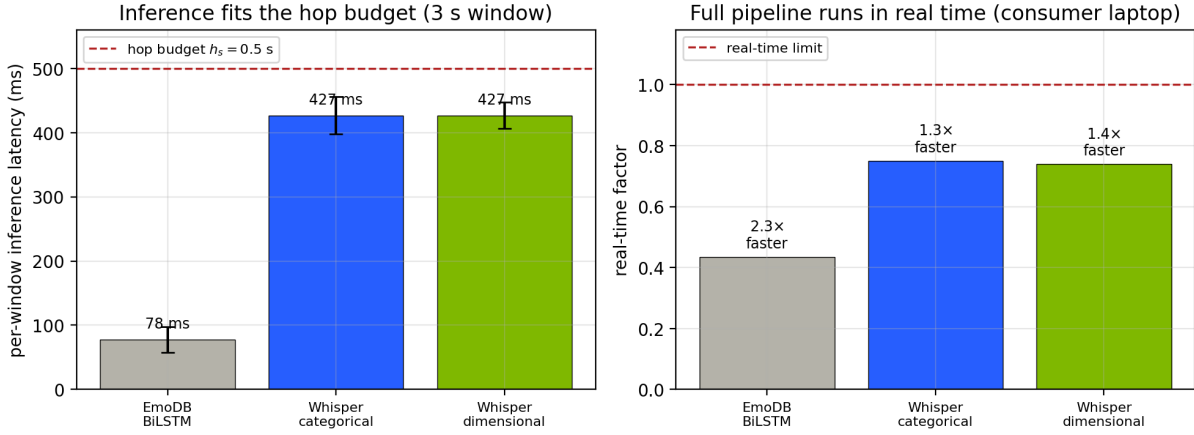


Figure 5.6: Left: per-window inference latency per backend against the $h_s = 0.5$ s hop budget. Right: end-to-end real-time factor against the real-time limit, annotated with the equivalent speedup ($1/\text{RTF}$).

frame is withheld after speech begins. Figure 5.7 visualises four (w_s, h_s) configurations on the same session: live events complete every h_s seconds once the first window has filled, and each weighted frame is released exactly when inference coverage reaches one window length past its centre, so that the first frame becomes available approximately w_s seconds after the first voiced window opens: 1.0 s after onset for $w_s = 1.0$ s and 3.0 s for $w_s = 3.0$ s in the measured runs. The cadence and latency of the operator-facing stream are therefore predictable functions of two configuration values, which is the property the configuration study of Section 5.5 exploits.

5.4. Validation of the Probabilistic Temporal Aggregation

The aggregation contribution (C2) claims that the overlap-weighted fusion of Section 4.2 turns the raw stream of window posteriors into an operator-readable signal, temporally stable at the frame level, distributionally rich at the utterance level, and free of retrospective revisions, without collapsing the distributions at any point. The three experiments of this section test these claims on the four evaluation sessions, processed under the default temporal configuration with the EmoDB BiLSTM backend, whose raw window posteriors are sharply peaked and temporally volatile and therefore constitute the regime in which a temporal aggregation stage must prove itself: it has to turn a jittery stream into a stable one without collapsing the distributions.

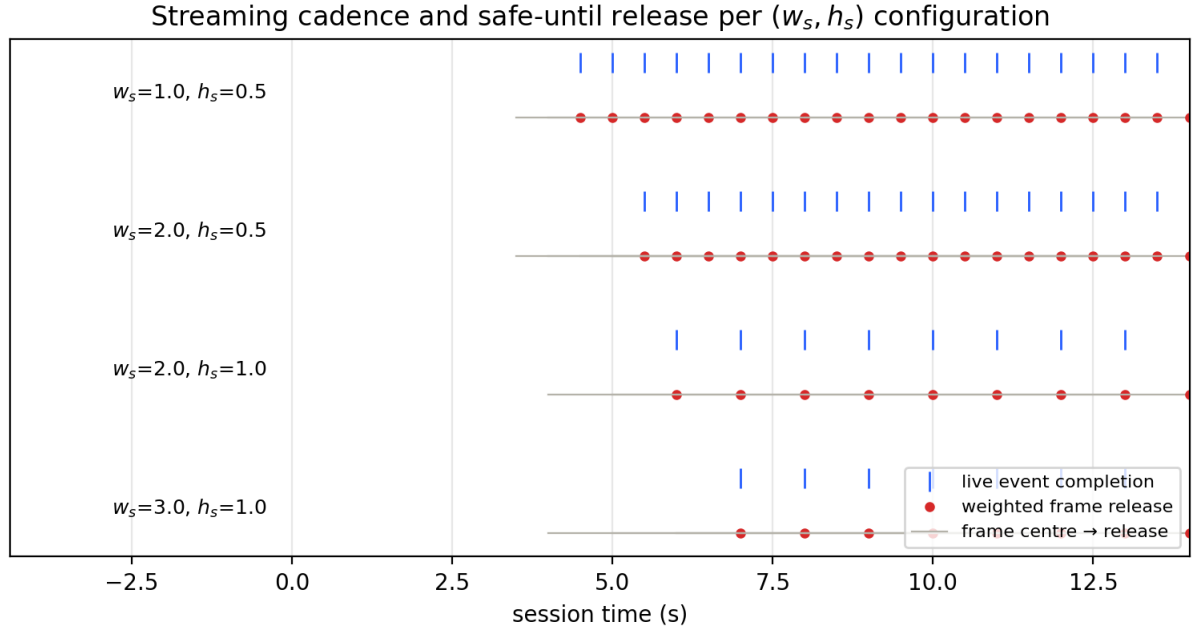


Figure 5.7: Streaming behaviour under four (w_s, h_s) configurations: live inference completions (ticks) and weighted frames, each drawn from its centre to its release instant.

5.4.1. Jitter Reduction

Figure 5.8 compares, for every session, the stability of the raw inference stream with that of the weighted-frame stream the aggregator derives from it. The raw per-window stream places most mass on a single class and reassigns it frequently; aggregation reduces the inter-frame jitter from about 0.03 to about 0.01 bits on average, a reduction of roughly 70% in all four sessions measured (per session 66, 64, 73 and 76%), and lowers the dominant-label flip rate by roughly a third, from about 10 to about 7 flips per minute across the four sessions; this last figure is carried by the three sessions on which the raw dominant label actually changes, the fourth being already flip-free under both streams. Because overlap-weighted averaging is by design a smoothing operator, a reduction in inter-frame jitter is expected rather than surprising; the experiment therefore measures the *magnitude* of the stabilisation actually obtained, not whether smoothing occurs. With only four ad-hoc sessions these values are indicative rather than statistically estimated, but the direction is consistent throughout, turning a signal that still reassigns its headline reading several times a minute into one an operator can follow. The uniform window averaging of Section 4.2.2 is therefore stabilising rather than distorting: the volatile stream is calmed without any collapse of the underlying distribution, as the next experiment confirms at the utterance level.

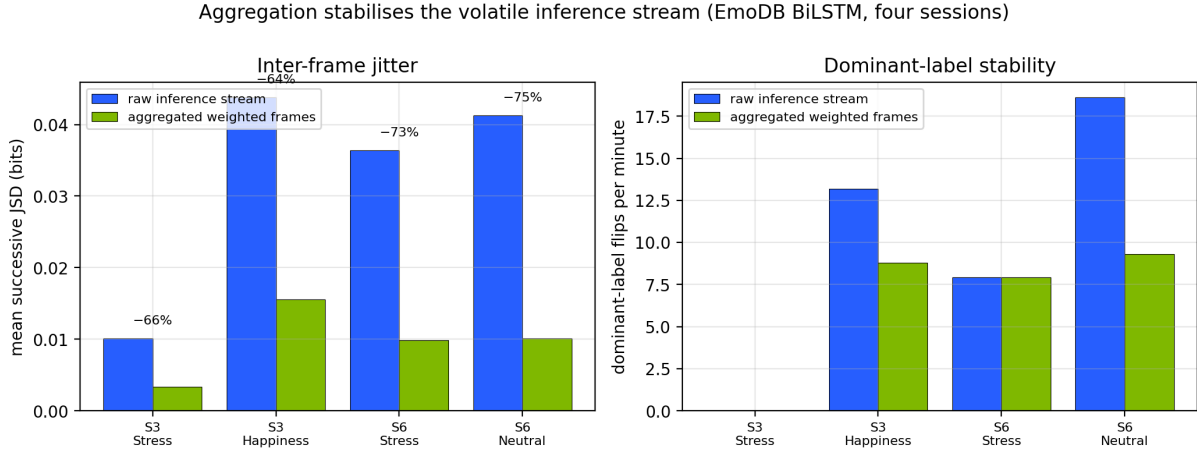


Figure 5.8: Stability of the raw and aggregated weighted-frame streams for the four sessions (EmoDB BiLSTM): inter-frame jitter (left) and dominant-label flip rate (right), per-session jitter reduction annotated.

5.4.2. Preservation of Distributional Structure

Across the four sessions the pipeline produced eight final utterances per backend, each carrying the overlap-weighted average of the window distributions that intersect it. Figure 5.9 makes the gain concrete. Its left panel shows one utterance produced through the dimensional backend, included here to illustrate the distances-to-probabilities conversion (C3) on real audio: a majority vote would report a single emotion at full confidence, whereas the aggregated distribution reveals that the dominant emotion holds only about two fifths of the mass, the remainder spread over its circumplex neighbours, the model’s genuine uncertainty, which the one-hot reduction would have hidden behind a false certainty. The two right-hand panels quantify, on the BiLSTM stream that drives this section, how much of this structure survives utterance-level aggregation. That the entropy and the secondary-emotion mass exceed the zero a one-hot reduction assigns follows by construction from averaging non-degenerate distributions; the panels are therefore not a test the design could fail but a measurement of the magnitude retained, means of about 1.0 bit of entropy and 0.3 of secondary mass across all utterances (Table 5.5), a substantial fraction of the input distributions rather than a vanishing residue. The mean Jensen–Shannon divergence between each aggregated distribution and its one-hot reduction, of order 0.2 bits, is precisely the information that a majority vote would discard at the utterance level and that the overlap-weighted representation places at the operator’s disposal instead, where it feeds the uncertainty annotation of Section 5.6.3. An independent recomputation of the canonical overlap-weighted average from the recorded live events confirms the implementation is correct: the aggregated utterances match it

up to bounded differences (at most around 0.03 bits) that are not discrepancies but the expected consequence of the single-consumption policy of Section 4.2.3, by which each inference event feeds at most one final utterance.

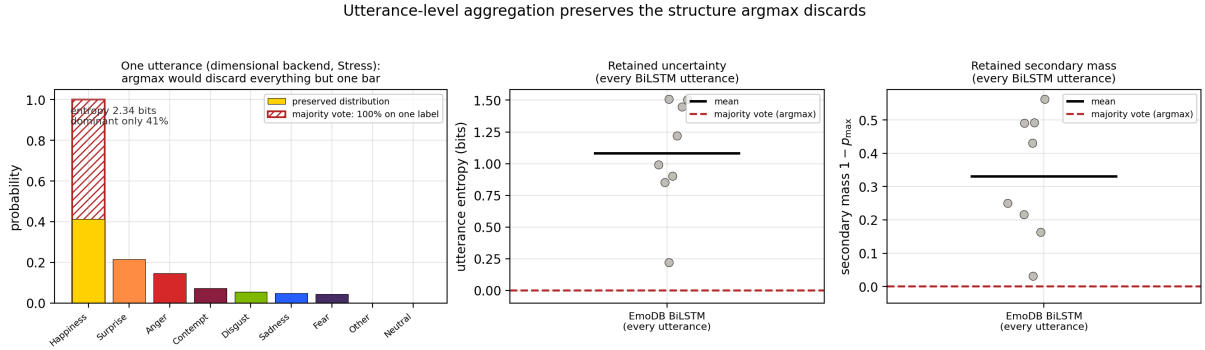


Figure 5.9: Distributional structure retained by utterance-level aggregation. Left: one dimensional-backend utterance, aggregated distribution (filled) against the full-confidence bar a majority vote reports (red hatch). Centre and right: entropy and secondary-emotion mass of every BiLSTM utterance (mean marked), well above the zero (dashed red) a majority vote assigns.

5.4.3. The Safe-Until Guard

The replay experiment quantifies what the guard spares the operator, measured against the most aggressive emit-on-first-event baseline, a policy no system would deploy, used here as the worst case. Feeding each session’s recorded live events through the production aggregator core under this eager policy, every frame emitted as soon as a single overlapping inference event exists, produces frames that are missing on average 2.7 of their 4 contributing events at the moment of emission, and roughly nine in ten frames would subsequently change once the remaining events arrived. The revisions are far from cosmetic: the mean revision is about 0.07 bits and the 95th percentile around 0.28 bits, of the same order as the raw inter-frame jitter itself, so an eager display would not merely flicker but would systematically contradict its own recent past. Figure 5.10 shows the cumulative distribution of these revisions. By construction the guard converts these revisions to exactly zero, a frame, once released, can never change, because it is held until all its contributing events are present, so the experiment characterises the resulting stability-versus-latency trade rather than contesting a rival aggregator: the price is a deterministic and configurable w_s seconds of release latency, quantified in Section 5.3.3, which is exactly the trade the design of Section 4.2.2 intends.

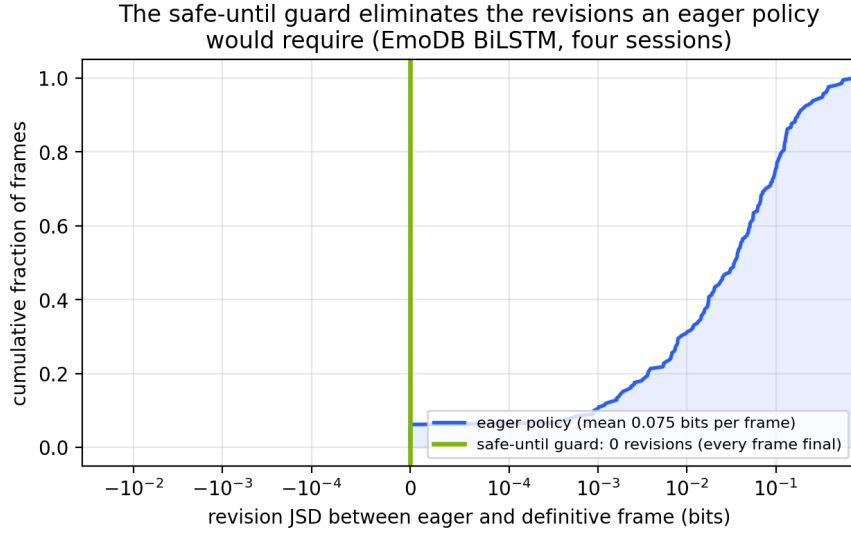


Figure 5.10: Cumulative distribution of the post-hoc revisions an eager release policy would require, pooled over the four sessions (EmoDB BiLSTM); the safe-until guard reduces every one to zero at a deterministic cost of w_s seconds of latency.

5.4.4. Summary

The three experiments establish that the overlap-weighted aggregation delivers each property the contribution claims. Table 5.5 consolidates the evidence: aggregation sharply reduces the inter-frame jitter of the volatile inference stream and lowers the dominant-label flip rate, and at the utterance level retains the entropy and secondary-emotion mass that a majority vote would set identically to zero, the difference amounting to a substantial Jensen–Shannon divergence between the aggregated distribution and its one-hot reduction. The safe-until guard completes the picture by making every released frame final. The aggregation stage therefore turns the raw stream of window posteriors into the stable, distributionally rich, revision-free signal that the visualisation layer of Section 5.6 renders to the operator.

5.5. Configuration Study: Recommended Operating Points

The heuristic design decisions of Section 4.4 fix five configuration values: the VAD threshold θ_{vad} , the window and hop durations w_s and h_s , and the two parameters τ and s of the distances-to-probabilities conversion. This section grounds each value in a controlled sweep, quantifying what the chosen operating point delivers to the operator-facing signal.

Property	Criterion	Measured
Stabilises the inference stream	inter-frame jitter of the raw stream is reduced by aggregation	−70% (min −64%)
Reduces dominant-label flips	the headline-emotion flip rate drops markedly	10→7 per minute
Retains uncertainty (vs argmax)	utterance entropy stays well above the majority-vote value of 0	mean 1.08 bits
Retains secondary mass (vs argmax)	secondary-emotion mass stays well above the majority-vote value of 0	mean 0.33
Carries information beyond argmax	JSD between the aggregated distribution and its one-hot reduction	mean 0.20 bits
Guard prevents post-release revisions	every released frame is final; an eager policy would revise it	0 after release (eager mean 0.07 bits)

Table 5.5: Scorecard of the probabilistic temporal aggregation: each claimed property, its acceptance criterion, and the measured value on the EmoDB BiLSTM stream.

5.5.1. VAD Gating Threshold

With gating disabled, the pipeline emits 19 inference events from the twelve seconds of digital silence of the synthetic session: spurious distributions computed on no speech at all, each of which would reach the operator display. Figure 5.11 shows that even the mildest threshold ($\theta_{vad} = 0.1$) suppresses eighteen of these nineteen events, the survivor being the single window that straddles the speech onset and is silence-dominated only by construction of the attribution rule. The spurious silence inference is therefore eliminated at any threshold from 0.1 upward; raising the threshold further does not change the silence count but progressively excludes voiced-region windows whose RNNoise speech probability is low, the lower-confidence windows that fall in the pauses and breaths within an utterance. The default $\theta_{vad} = 0.5$ is the natural decision boundary, running inference only where RNNoise judges speech the more probable hypothesis: it retains 46 of the 65 ungated voiced windows, with the count stable across $\theta_{vad} \in [0.3, 0.5]$ and eroding only from $\theta_{vad} = 0.6$ onward. A lower threshold would admit more of the low-confidence pause windows without improving the silence suppression, so 0.5 is a principled default, the probability midpoint, rather than the optimum of a measured cost, and the sweep confirms it sits safely on the voiced-coverage plateau.

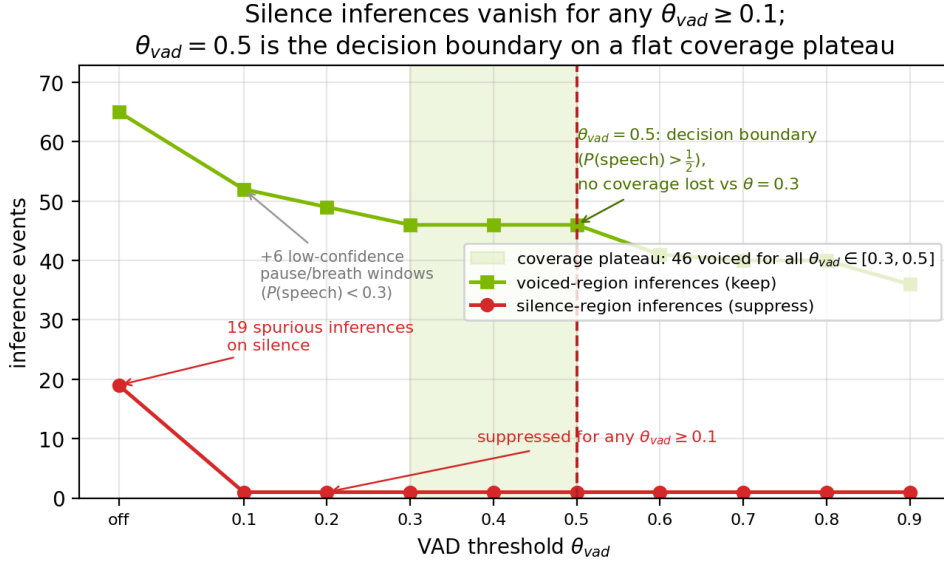


Figure 5.11: Inference events attributed to the silence and voiced regions of the synthetic silence-padded session against the VAD threshold. Any $\theta_{vad} \geq 0.1$ removes the spurious silence inferences, and every threshold in $[0.3, 0.5]$ retains identical voiced coverage; the default $\theta_{vad} = 0.5$ is therefore the principled decision boundary at no coverage cost, while lower thresholds only add low-confidence pause and breath windows.

5.5.2. Window Duration

The window duration determines how much acoustic context each posterior summarises, and Figure 5.12 quantifies both sides of the resulting trade-off on the responsive BiLSTM backend across the three evaluation sessions (one per affective condition). Averaged over the three sessions, the weighted-frame jitter falls monotonically with the window, from about 0.06 bits at $w_s = 1.0$ s through about 0.02 bits at 2.0s to about 0.01 bits at 4.0s, and the dominant flip rate falls in parallel by roughly a factor of four; the trend is monotone in the mean rather than in every individual session (the Happiness session, for instance, shows a small jitter uptick at 4.0s). Against this, the measured first-frame release latency equals w_s exactly, 1.0s of audio after speech onset at $w_s = 1.0$ s, 4.0s at $w_s = 4.0$ s, matching the analytic behaviour of the safe-until guard. The jitter does not plateau at $w_s = 2.0$ s: it continues to fall markedly with the window, dropping by roughly half again between 2.0 and 3.0s, so larger windows are genuinely more stable, at a first-frame latency that grows one-for-one with w_s . Because $w_s = 3.0$ s is in any case the minimum input the Whisper encoder accepts without zero-padding, it is adopted as the default and recommended operating point: it is markedly more stable than the 2.0s setting and uniform across all backends, at the cost of one additional second of first-frame

latency. The 2.0 s setting remains available as the minimum-latency configuration.

Window duration: the stability--latency trade-off (EmoDB BiLSTM, three sessions, $h_s = 0.5$ s)

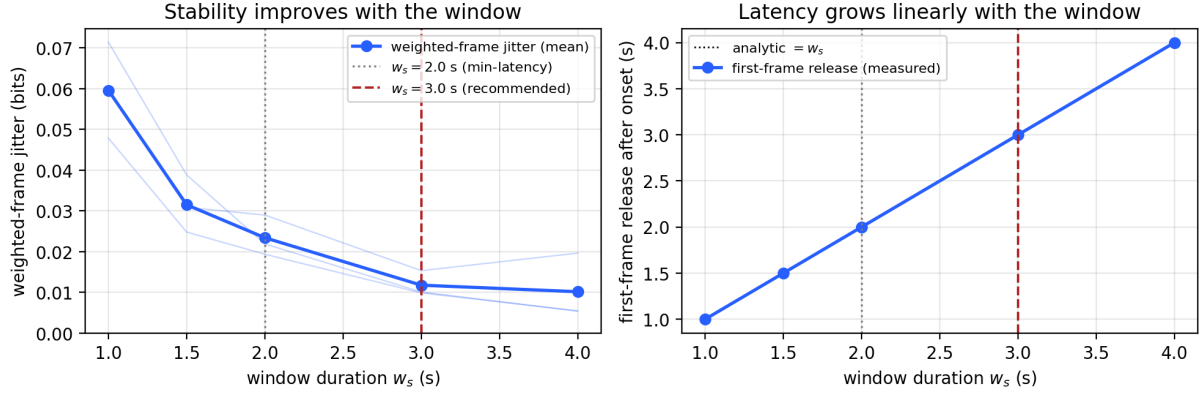


Figure 5.12: Stability--latency trade-off governed by the window duration (EmoDB BiLSTM, three sessions): weighted-frame jitter falls with the window (left; faint lines per session, markers the mean), first-frame release latency grows linearly (right).

5.5.3. Hop Duration

The hop duration sets the cadence of the operator display at $1/h_s$ frames per second and, because one window inference is issued per hop, scales the compute load in the same proportion. Figure 5.13 reports both effects. On the compute side, the BiLSTM is unconstrained at every hop (real-time factor below 0.6), whereas the large Whisper dimensional backend exceeds the real-time limit at $h_s = 0.25$ s on the evaluation hardware but drops to an RTF of about 0.71 at the default $h_s = 0.5$ s, which thus provides a 2 Hz display update with a comfortable real-time margin even for the largest backend. On the signal side, shorter hops yield smoother frame-to-frame transitions, because adjacent frames share a larger fraction of their supporting windows: the BiLSTM jitter per step grows from a few thousandths of a bit at $h_s = 0.25$ s to about 0.07 bits at $h_s = 2.0$ s. The default $h_s = 0.5$ s is therefore the natural operating point: the finest cadence that every backend sustains in real time, with per-step transitions half as large as those of a 1 s hop.

5.5.4. Conversion Parameters of the Dimensional Backend

The two parameters of the distances-to-probabilities conversion are characterised on the synthetic geometry of Section 5.2. The distance temperature τ governs how the peripheral mass spreads across anchors, and Section 4.3.2 imposes two geometric criteria: the corresponding emotion must be the dominant mode at any anchor, and off-anchor predic-

Hop duration: the resolution--compute trade-off (three sessions, $w_s = 3.0$ s)

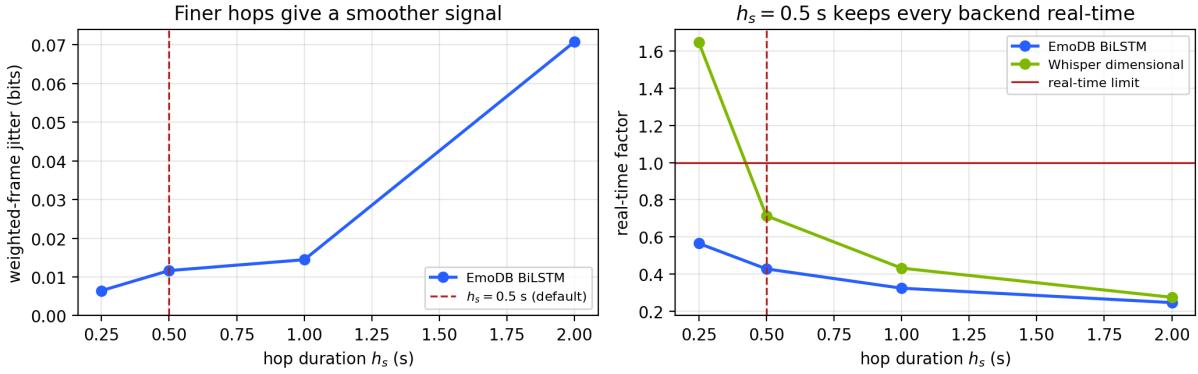


Figure 5.13: Resolution--compute trade-off governed by the hop duration: weighted-frame jitter on the BiLSTM (left) and end-to-end real-time factor of both backends (right); the large Whisper model exceeds the real-time limit only at $h_s = 0.25$ s, so $h_s = 0.5$ s is the finest cadence every backend sustains.

tions must remain graded rather than collapsing to a winner-take-all assignment. These criteria are made operational by two explicit thresholds: the *dominance* criterion requires the mean anchor self-mass to exceed one half, so that the anchor emotion outweighs all other classes combined; the *gradedness* criterion requires the quarter-point secondary share to exceed the uniform peripheral baseline $1/7 \approx 0.14$, so that a near-anchor point still assigns a neighbour more than blind-chance mass. Figure 5.14 traces both quantities over $\tau \in [0.10, 3.00]$: the mean anchor self-mass decreases monotonically with τ and falls through one half at $\tau \approx 0.55$, while the quarter-point secondary share, the mass a point one quarter of the way towards a neighbouring anchor assigns to that neighbour, rises with τ and clears the gradedness floor at $\tau \approx 0.18$. The admissible region in which both thresholds are met is therefore $[0.18, 0.55]$, and the operating value $\tau = 0.45$ lies within it, towards the graded end: anchors keep their dominance while adjacent emotions retain visible mass, which is the behaviour the bimodality experiment of Section 5.2.3 verified pointwise. The neutral scale s fixes the radius $1/s$ at which the Neutral mass is extinguished; Figure 5.15 confirms that the measured extinction radius matches the analytic value to within the sweep resolution (10^{-3}) across $s \in [1.0, 4.0]$, and that the Neutral-dominance handover radius scales with it. The operating value $s = 2.0$ extinguishes Neutral at $r = 0.5$, so that the moderate radii the MSP-Podcast dimensional model produces on conversational speech are read as graded affect rather than absorbed into Neutral, in line with the rationale of Section 4.4.

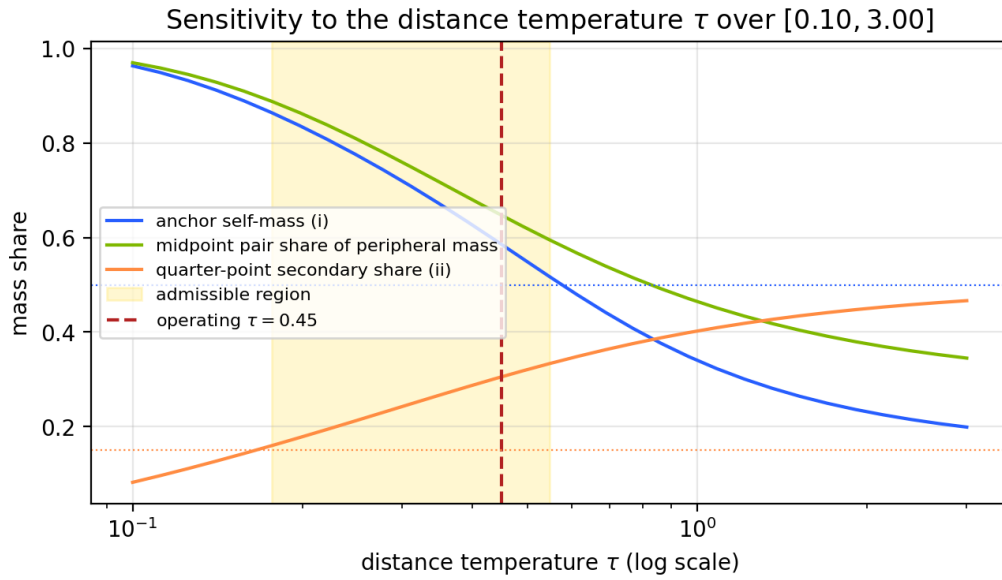


Figure 5.14: Sensitivity to the distance temperature over [0.10, 3.00]: anchor self-mass, midpoint pair share, and quarter-point secondary share, with the admissible region satisfying both criteria.

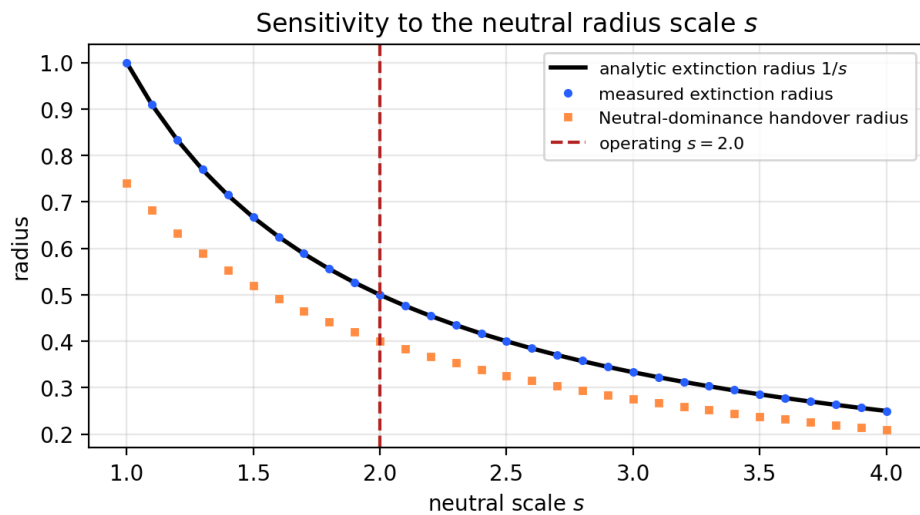


Figure 5.15: Sensitivity to the neutral radius scale: measured Neutral extinction radius against the analytic $1/s$, and the Neutral-to-peripheral handover radius.

5.5.5. Recommended Configuration

Table 5.6 consolidates the operating points established by the four sweeps into the recommended configuration of the platform, each annotated with the concrete benefit it delivers to the operator-facing signal. Each value resolves an explicit trade-off rather than following a convention, and each sweep doubles as a sensitivity analysis: the VAD threshold is the probability midpoint sitting on a coverage plateau, the window duration is set for stability at 3.0s, which is also the Whisper encoder minimum, making the configuration uniform across backends, with 2.0s available when minimum latency is preferred, the hop duration is the finest cadence every backend sustains in real time, and the conversion parameters occupy the interior of their admissible region, so that moderate deviations from any recommended value degrade the operator-facing signal gracefully rather than abruptly.

Parameter	Explored	Recommended	What the operating point delivers
θ_{vad}	off, 0.1–0.9	0.5	removes all 19 spurious silence inferences; keeps 71% of voiced coverage
w_s	1.0–4.0 s	3.0 s (2.0 s min-latency)	$5.1\times$ lower jitter than $w_s = 1$ s and lower than the 2.0 s setting; no Whisper padding; first frame $\approx w_s$
h_s	0.25–2.0 s	0.5 s	2 Hz cadence; every backend below the real-time limit (Whisper RTF 0.71)
τ	0.10–3.00	0.45	inside the admissible region [0.18, 0.55] (notebook 01)
s	1.0–4.0	2.0	Neutral mass extinct at $r = 0.5$ (notebook 01)

Table 5.6: Recommended operating configuration of the platform: the range explored by each sweep, the selected operating point, and the quantified benefit it delivers to the operator-facing signal.

5.6. Fidelity of the Interpretability-First Visualisation

The visualisation contribution (C5) claims that the three operator-facing components, the probability-weighted circumplex, the interpolated-colour timeline, and the transcript panel with uncertainty annotation, are faithful, deterministic renderings of the probability distributions the pipeline produces, at the frame and at the utterance granularity. Because

each component is a closed-form function of the distribution, fidelity can be verified exactly, in two complementary ways: a deterministic geometric characterisation of the rendering function itself, independent of any audio or model, and a demonstration of the same renderers in operation on the temporally dynamic BiLSTM stream, the backend whose distribution evolves visibly within a session and therefore moves the centroid and the colour band. The fidelity established this way is a property of the rendering layer, shared by every backend; the BiLSTM, processed on three sessions chosen one per affective condition, drives the circumplex and timeline illustrations, while the transcript panel is illustrated on a single session, where the per-utterance annotation is best seen as the transcript is built piece by piece. What this section establishes is rendering *faithfulness*, that the views report the posterior without distortion, not a measured gain in operator comprehension; a controlled human-factors study of how quickly and accurately operators read these views is left as future work (Section 6.2).

5.6.1. Circumplex Centroid Faithfulness

The point the circumplex view draws is the probability-weighted centroid of the anchor coordinates (equation (4.1)), a closed-form function of the categorical distribution that is identical for every backend. Characterising this renderer over the dimensional backend's reachable states, composing it with the distances-to-probabilities conversion so that a single valence-arousal sweep induces the corresponding centroids, gives a deterministic map from model state to drawn point whose fidelity can be characterised exactly, independently of any audio or model. Figure 5.16 reports this characterisation over the full polar grid of the valence-arousal plane; because the centroid function itself is backend-independent, the same fidelity properties carry over to the categorical distributions the BiLSTM produces directly, which the illustrations below exercise. Along every azimuth covered by the anchor set, the span from *Happiness* at 27° to *Sadness* at 261° , the rendered radius grows strictly monotonically with the model radius (per-azimuth Spearman $\rho = 1.00$), so that order relations of emotional intensity are always preserved, while the absolute radius is compressed within 0.90 of the unit disc: this compression is the confidence coding of the view, since a broad distribution pulls the centroid towards the centre exactly when the model is uncertain. The rendered azimuth follows the model azimuth with a mean absolute deviation of 17.6° and a maximum of 48.1° , bounded by half the local angular gap between adjacent anchors, which reaches 90° between *Fear* and *Sadness*; in the lower-right quadrant, which the MSP-Podcast taxonomy leaves without any categorical anchor, the map renders points by interpolation between the two boundary anchors, the most informative reading a categorical label space admits for an affect it cannot name.

This characterisation guarantees the map for every possible model state; Figure 5.17 then shows it exercised on real material. Applying the map to the live BiLSTM frames of the three evaluation sessions, the rendered centroids localise into three distinct, correct regions of the circumplex: the Happiness session settles on the high-valence side towards the *Happiness* anchor (mean valence $+0.38$), the Neutral session clusters near the origin (mean $(-0.15, +0.20)$), and the Stress session occupies the high-arousal negative sector at the *Anger* anchor (mean arousal $+0.69$), the nearest categorical neighbour of vocal stress in a taxonomy with no dedicated stress label. The operator therefore reads the affective state of a session directly from the position of the point the renderer is proven to draw faithfully.

The rendered centroid is a monotone, confidence-compressed image of every model state

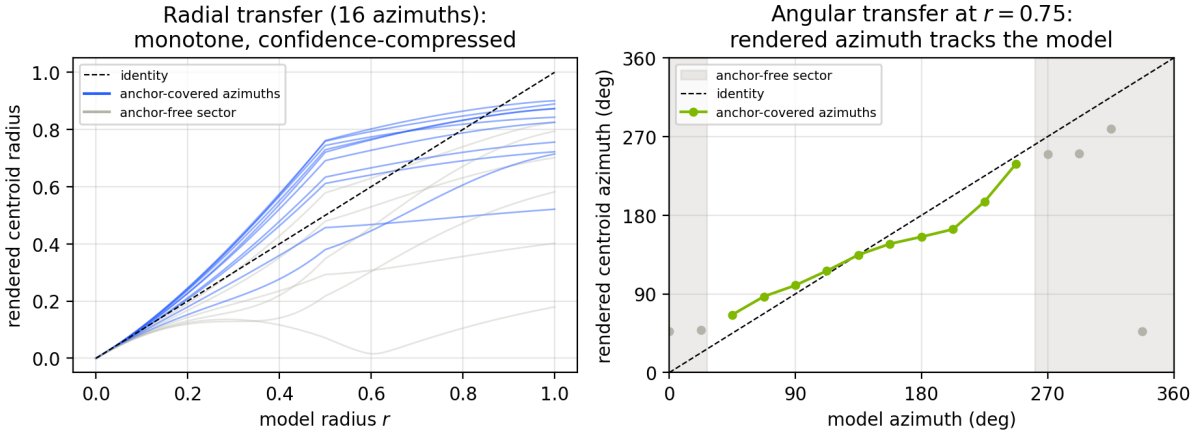


Figure 5.16: Synthetic rendering map of the circumplex centroid over the valence–arousal plane: radial transfer (left), monotone and confidence-compressed along every anchor-covered azimuth, and angular transfer (right), tracking the model azimuth within the anchor-set quantisation.

5.6.2. The Interpolated-Colour Timeline

The operator timeline reduces a session to a single colour band, one interpolated colour per weighted frame. Figure 5.18 places the three sessions side by side and shows that each band is a clean, distinct colour signature qualitatively consistent with the session’s induced affective condition: the Happiness session renders predominantly yellow, its dominant frame reading *Happiness* on 65% of frames, the Neutral session renders a calm grey, dominant *Neutral* on 68%, and the Stress session renders a sustained red, dominant *Anger*, the nearest categorical neighbour of vocal stress in a taxonomy with no dedicated stress label, on 82%, the cleanest of the three. An operator therefore tells a calm call from

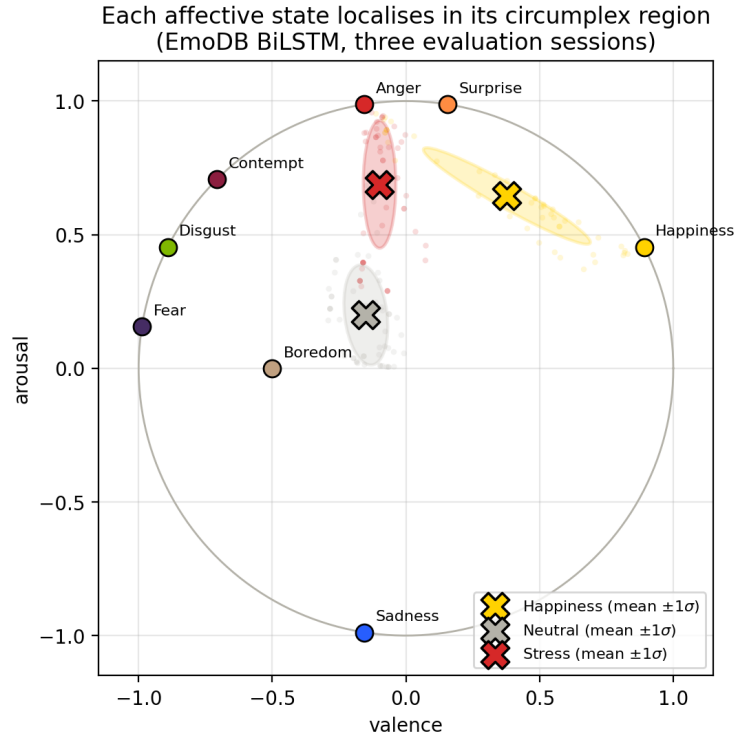


Figure 5.17: Rendered circumplex centroids of the three sessions (EmoDB BiLSTM): per-session mean ($\times$) and $\pm 1\sigma$ ellipse localise each state distinctly: Happiness towards the high-valence anchor, Neutral near the origin, Stress at the *Anger* anchor.

a stressed one at a glance, from colour alone, which is exactly the monitoring affordance the timeline is meant to provide. The bands are rendered on the BiLSTM stream; the renderer is a deterministic function of the distribution and is identical for every backend. The quantitative counterpart is the colour-fidelity analysis of Figure 5.19: across the 162 weighted frames of the BiLSTM stream, spanning secondary-emotion masses from near zero to about 0.6, the deviation of the interpolated colour from the dominant-only colour grows with the secondary mass with a Pearson correlation of about 0.8 on the pooled frames (per session about 1.00 for the Happiness session, 0.98 for the Stress session, and 0.60 for the more uniform Neutral session); the pooled figure is the higher because it spans the wider mass range of the three sessions together. A dominant-only colouring would collapse this entire axis of information to zero; the interpolated blend makes the difference between a confident frame and an ambivalent one directly visible in the saturation of the band.

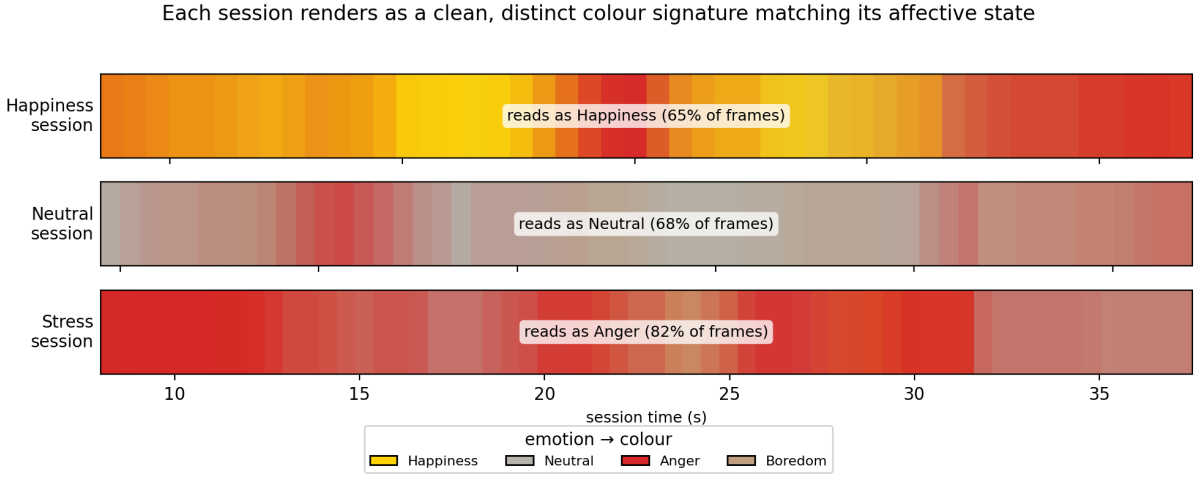


Figure 5.18: Operator timeline colour bands for the three sessions (EmoDB BiLSTM): each renders as a distinct colour signature, yellow for Happiness, grey for Neutral, red (*Anger*) for Stress, with the dominant reading and its frame share annotated.

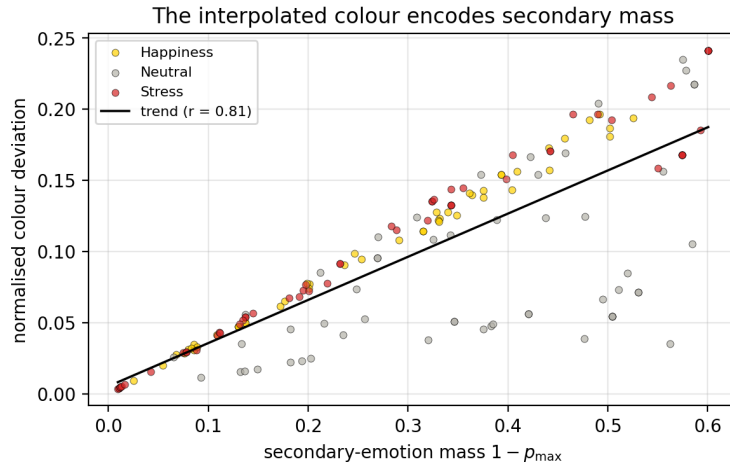


Figure 5.19: Normalised deviation of the interpolated frame colour from the dominant-only colour against the secondary-emotion mass of the frame.

5.6.3. Transcript Panel with Uncertainty Annotation

Table 5.7 reproduces the content the transcript panel surfaces to the operator on a single session (the Neutral session), following the utterance as it is assembled from the incremental partial pieces the streaming transcriber emits: each row pairs a transcript delta with the dominant emotion of its overlap-weighted distribution and the probability of that emotion (the **confidence** field), and the consolidated final utterance closes each group. Only a curated excerpt of the partial pieces is shown, with omitted runs marked by a vertical-dots row. The annotation does for the operator exactly what the design of

Section 4.5.3 intends. The dominant reading is predominantly *Neutral*, but a few partial pieces drift momentarily to an adjacent label, here *Anger* and *Boredom*, at noticeably lower confidence; the per-piece confidence flags exactly these tentative readings, and the utterance-level aggregation consolidates each sequence back to a confident *Neutral* in the final utterance, which is precisely what the uncertainty annotation and the overlap-weighted aggregation are for. The fidelity claim concerns what the panel guarantees the operator: beside every label, as the sentence grows, it reports the confidence that quantifies how much of the distribution stands behind it, so that the operator weighs each reading by its evidence rather than treating every label alike. The Shannon entropy of the full distribution is retained in the evaluation notebook as an analysis cross-check on the same preserved probabilities. A label-only panel would have asserted the same dominant emotions with no indication of how much weight the model actually placed behind them; the uncertainty annotation, fed directly by the preserved distributions of Section 5.4, makes the difference between confidence and conjecture part of the displayed record.

5.7. Discussion

Read together, the five experimental sections substantiate each contribution with converging quantitative evidence, summarised in Table 5.8.

The platform claim (C1) rests on the demonstration of Section 5.3: one session, three backends, one code path, identical event structure, with every backend within the per-window latency budget and the full pipeline at a real-time factor of at most 0.75 on a consumer laptop. What makes this more than an engineering observation is its interaction with the aggregation results: because the temporal behaviour of the system is owned by the pipeline rather than by the model, the stability and preservation guarantees of Section 5.4 were established once and held for backends of radically different output character, which is precisely the property an open architecture must offer to admit future models without re-deriving its guarantees.

The aggregation claim (C2) is substantiated on the responsive BiLSTM stream: the volatile raw stream is stabilised (around 70% jitter reduction, flip rate cut by about a third) without any collapse of the distributions, and at the utterance level the overlap-weighted representation demonstrably retains the entropy and secondary-emotion mass (of order 0.2 bits of JSD against the one-hot reduction) that a majority vote would discard. The safe-until guard completes the claim by making the released signal final: the replay experiment shows that without it roughly nine frames in ten would be revised

Piece	Interval (s)	Transcript delta	Dominant	$p_{\max}$
partial	5.2-5.8	il consumatore	Neutral	0.81
partial	5.8-7.0	residente in europa	Neutral	0.71
partial	7.3-7.8	deve essere	Neutral	0.62
partial	7.9-9.1	a conoscenza del fatto che	Neutral	0.56
partial	9.1-10.1	la commissione europea	Anger	0.61
partial	10.8-11.3	ha istituito	Anger	0.71
partial	11.4-12.1	una piattaforma	Anger	0.68
partial	12.2-12.6	anche	Anger	0.65
partial	12.6-13.8	fornisce uno strumento	Neutral	0.63
		⋮		
partial	16.0-17.0	delle controversie	Neutral	0.61
<i>final</i>	5.2-17.2	il consumatore residente in europa deve essere a conoscenza del fatto che la commissione europea ha istituito una piattaforma online che fornisce uno strumento di risoluzione alternativa delle controversie	Neutral	0.70
partial	18.1-18.8	uno strumento può	Neutral	0.74
partial	18.8-19.1	essere	Boredom	0.71
partial	19.1-20.6	utilizzato dal consumatore	Neutral	0.54
partial	20.6-21.2	europeo per	Neutral	0.84
		⋮		
partial	26.5-27.4	e o derivanti da	Neutral	0.48
partial	27.7-28.8	contratti di vendita	Boredom	0.38
partial	28.8-29.4	di beni	Boredom	0.46
partial	29.4-30.1	e servizi	Boredom	0.47
<i>final</i>	18.2-31.3	uno strumento può essere utilizzato dal consumatore europeo per risolvere in via non giudiziale ogni controversia relativa a e o derivanti da contratti di vendita di beni e servizi stipulati in rete	Neutral	0.63

Table 5.7: Transcript panel content for the Neutral evaluation session (EmoDB BiLSTM): a curated excerpt of the incremental partial pieces, each with its transcript delta, dominant emotion, and probability (the operator-facing confidence), the consolidated final utterance closing each group; vertical dots mark omitted partial runs, and the per-piece entropy is retained in the evaluation notebook as an analysis measure.

after emission, by margins comparable to the jitter the aggregation removes.

The distances-to-probabilities claim (C3) is established deterministically: unimodality at every anchor with the residual mass flowing to geometric neighbours, exact direction-invariant Neutral behaviour, perfectly balanced bimodality at all seven adjacent-anchor midpoints, and global smoothness with a mean neighbour JSD of 1.4×10^{-4} bits, the properties that make the dimensional backend a structurally equal citizen of the categorical pipeline, as the interchangeability demonstration of Section 5.3 then confirms *in vivo*.

The configuration claim (C4) is discharged by the operating points of Table 5.6, each located by an explicit sweep and each accompanied by the quantified benefit it delivers: silence-inference suppression for θ_{vad} , a several-fold jitter reduction together with backend-uniform, padding-free operation for the recommended $w_s = 3.0$ s, real-time margin at 2 Hz cadence for h_s , and the geometric criteria of the conversion for τ and s .

The visualisation claim (C5), finally, is reduced to mathematics: all three operator-facing components are closed-form functions of the preserved distributions, the circumplex centroid is a monotone, confidence-compressed image of every model state, the timeline colour provably encodes the secondary-emotion mass ($r \approx 0.8$), and the transcript annotation states the weakness or strength of each commitment explicitly. The interpretability-first design thus closes the loop opened by the probabilistic representation of Section 4.1: the distributions preserved by C2 and made universal by C3 arrive at the operator not as raw numbers but as faithful, readable renderings.

5.8. Scope and Generalisation of the Results

The claims established in this chapter are precisely scoped, and each generalises along a clear, well-understood path.

The evaluation corpus. The evaluation material, four sessions from two speakers reading call-centre-style text under controlled affective conditions, is exactly what this chapter requires: realistic, conversational-length Italian speech that exercises the full pipeline and yields clean measurements of how faithfully the platform transports, aggregates, and renders what its backends produce. The notebook-based infrastructure of Section 5.1.3 ingests new material without modification, so a larger corpus with utterance-level annotations extends the same methodology directly, adding concordance-based measurements (for instance CCC against annotated valence–arousal traces) on top of the present results.

Generality of the operating points. The recommended configuration of Sec-

Contribution	Section	Headline evidence
C1 platform	5.3	3 backends, 1 code path, identical event structure; $\text{RTF} \leq 0.75$; latency within the 500 ms hop budget
C2 aggregation	5.4	jitter about -70% and flips cut by about a third on a volatile stream; ~ 0.2 bits preserved vs. argmax; zero post-release revisions
C3 conversion	5.2	unimodal anchors, exact Neutral law, balanced midpoint bimodality, 1.4×10^{-4} bits mean neighbour JSD
C4 configuration	5.5	five operating points, each justified by an explicit trade-off, plateau, or admissible-region interior
C5 visualisation	5.6	monotone confidence-compressed centroid; colour-secondary-mass correlation ≈ 0.8 ; explicit uncertainty annotation

Table 5.8: Synthesis of the experimental evidence by contribution.

tion 5.5.5 rests on structural properties of the pipeline, monotone jitter reduction with w_s , cadence and compute scaling with h_s , and the silence-suppression plateau of θ_{vad} , that hold by construction across the three affective conditions on which they were verified. A deployment confirms the precise numerical operating points for its own acoustic conditions and microphone chain by re-running the configuration notebooks on a sample of its traffic, a one-command operation the infrastructure already supports.

Transfer of the performance measurements. The real-time factors and latencies of Section 5.3 were measured on commodity hardware (Apple M1 Pro, MPS backend), the deployment assumption of the system, and leave comfortable headroom under the real-time limit. The per-window latency-budget analysis of Section 5.3.2 transfers directly to any target machine, so feasibility on a given deployment is established by a single budget check.

Geometry of the conversion. The geometric guarantees of the distances-to-probabilities algorithm hold across the whole valence-arousal disc. In the lower-right (positive-valence, low-arousal) quadrant, which the MSP-Podcast taxonomy leaves without a categorical anchor, the map renders points by interpolation between the two boundary anchors, the most informative reading a categorical label space admits for an affect it does not name, and a richer anchor set or parameterisation, noted in Section 6.2, refines this region further.

Threats to validity. Three boundaries delimit the generality of the findings and define the work that would consolidate them. First, the evaluation speech is read under induced affective conditions rather than drawn from spontaneous calls, and it spans two speakers of a single language; the temporal, distributional, and computational properties measured here are backend- and content-agnostic and therefore carry over, but the reading of specific emotional content should be re-confirmed on spontaneous in-domain traffic. Second, the session-level condition labels support the platform-level measurements reported here rather than per-utterance accuracy or calibration, which an annotated corpus would add on top of the same methodology. Third, the dimensional conversion parameters τ and s and the VAD threshold are established on synthetic geometry and a single silence-padded session respectively; their behaviour on real dimensional-model outputs and across diverse acoustic conditions is the natural target of the in-domain study outlined in Section 6.2. None of these boundaries affects the core claims of the chapter, which concern how faithfully the platform transports, aggregates, and renders whatever its backends produce.

6 | Conclusions and Future Work

6.1. Summary of Contributions

This thesis has presented the design, implementation, and evaluation of a modular real-time Speech Emotion Recognition platform for call-centre deployment. The system bridges the persistent gap between the research community’s best academic SER models and the operational contexts that motivate them, by providing a complete, open-architecture pipeline that preserves the full probabilistic richness of model outputs from inference through to operator-facing visualisation.

The first three contributions establish the platform’s algorithmic foundations. The hexagonal architecture and dependency-injection container (C1) factorises three operational SER backends, a BiLSTM trained on EmoDB and two Whisper-based checkpoints, categorical and dimensional, behind a uniform port interface, enabling session-level backend selection without any modification to the upstream pipeline or downstream display logic; a fourth adapter for the now-retired Hume AI cloud API is retained in the codebase to document the cloud-integration path. The probabilistic temporal aggregation algorithm (C2) transforms the raw sliding-window event stream into two higher-level representations, regularly spaced weighted frames and transcript-enriched utterances, while preserving the full posterior distribution at every granularity, ensuring that secondary emotions, confidence gradients, and temporal trajectories remain available to the operator. The distances-to-probabilities inversion algorithm (C3) provides a principled conversion from continuous valence–arousal coordinates to categorical probability distributions, grounded in Russell’s circumplex geometry and the anchor framework of Jing et al. The synthetic suite confirms that the implementation exhibits its four intended geometric properties, anchor consistency, monotone neutral probability, bimodal preservation at anchor midpoints, and full-plane geometric coherence, under the chosen parameters $s = 2.0$ and $\tau = 0.45$. These properties follow analytically from the construction, so the suite verifies conformance to the specification rather than an empirical hypothesis; the one genuinely empirical outcome is the admissible range of τ established by the dedicated temperature sweep of the configuration study.

The fourth and fifth contributions address the operational and interpretive layers of the system. The heuristic design decisions (C4) documented in this thesis, VAD gating, window duration selection, and dimensional backend parameterisation, were each examined on the evaluation material: VAD gating at $\theta_{\text{vad}} = 0.5$ suppresses spurious silence-region inference, all but the single window that straddles the speech onset, while retaining about 71% of voiced windows, and operating at $w_s \geq 3\text{ s}$ eliminates zero-padding distortion and yields a monotone improvement in prediction consistency. The interpretability-first visualisation framework (C5) renders the probability-weighted circumplex centroid, an interpolated-colour emotional timeline, and a per-utterance transcript panel, all driven by the same normalised probability vectors that flow through the pipeline, making distributional uncertainty directly and continuously readable by the operator without any reduction to a dominant label. Taken together, these five contributions demonstrate that distributional, interpretability-first SER is deployable today on commodity hardware, and that the resulting system can provide transparent, reliable probabilistic emotion signals in a real-time streaming context.

6.2. Future Work

The platform delivered by this thesis is, by design, a foundation rather than an end point: a distribution-preserving streaming substrate onto which more specialised backends, deployment contexts, and research programmes can be grafted without re-deriving the architectural guarantees established here. The directions below proceed from the most immediate technical consolidations of the present system, through its evolution into a deployable analytics instrument and its use as an experimental platform across application domains, to a broader research agenda for which this work supplies the infrastructural base.

6.2.1. Near-Term Technical Consolidation

In-domain backends and calibration. The most direct path from the current prototype to a production-grade system is the deployment of backends fine-tuned for a specific call-centre context. The framework is modular by construction: replacing the general-purpose academic backends with models fine-tuned on call-centre speech in the target language and acoustic environment would immediately sharpen the discriminative power and calibration of the operator-facing signals, transforming the prototype into a fully tailored tool for the deployment domain. Because the temporal behaviour of the system is owned by the pipeline rather than the model, such a replacement requires only the

implementation of a new adapter satisfying the `ISERBackend` port, with no change to the aggregation or visualisation logic. Fine-tuning the Whisper backbone on emotional speech collected in the deployment language would, in addition, open the way to a formal Expected Calibration Error analysis at the pipeline output level, and extending the evaluation to a larger, per-utterance labelled corpus, in Italian or in the target deployment language, would enable quantitative calibration analysis and a formal comparison of the backends under matched conditions, building directly on the temporal and architectural characterisation established here. The notebook-based infrastructure of this work is already designed to ingest such material without modification.

Representation geometry of the non-anchored quadrant. The MSP-Podcast taxonomy [10] leaves the positive-valence, low-arousal quadrant, the region of affects such as *calm* or *content*, without a categorical anchor; in that quadrant the distances-to-probabilities map renders points by interpolation between the two boundary anchors, the most informative reading a categorical label space admits for an affect it does not name. A natural refinement is an *extended or data-driven anchor set* that covers the entire circumplex disc [51], together with a systematic study of its effect on the fidelity of the conversion in that region. Relatedly, the neutral radius scale s and the peripheral temperature τ , validated synthetically and set to values that satisfy the algorithm’s design requirements, invite a systematic empirical study on an in-domain corpus, enabling data-driven parameter selection and a richer parameterisation of the conversion.

Robustness in real acoustic conditions. Two boundaries of the present evaluation define complementary robustness studies. First, the VAD gating threshold $\theta_{\text{vad}} = 0.5$ is justified as a natural operating point but was calibrated on controlled, silence-padded material; its behaviour on *real call-centre noise*, hold music, crosstalk, variable background noise, and narrow-band telephony channels, together with an adaptive gating strategy that tracks the acoustic condition, is a natural next step. Second, the pipeline currently operates on a single audio stream, whereas a real call-centre interaction involves two interlocutors and overlapping speech. Integrating speaker diarisation and source separation, with per-channel SER and aggregation performed per speaker, would extend the temporal and distributional machinery developed here to the genuinely multi-party setting of operational telephony, allowing the emotional trajectories of caller and agent to be read separately and jointly.

Human factors of the operator-facing layer. This thesis establishes that the visualisations render the posterior faithfully, but it does not measure how effectively human op-

erators read them. A controlled human-factors study, comparing the probability-weighted views against conventional single-label displays on monitoring tasks such as detecting an emotional escalation or judging model uncertainty, would turn the interpretability rationale of contribution C5 into an empirically validated claim.

6.2.2. A Quantitative Call-Centre Analytics Instrument under Privacy Constraints

Beyond live operator support, the same distribution-preserving substrate can be evolved into an instrument for the *quantitative, retrospective evaluation* of calls. The per-utterance and per-frame posteriors that the pipeline already produces are the natural building blocks of call-level and agent-level analytics: scalar indices of emotional volatility, escalation frequency, time-to-resolution of negative affect, and the divergence between an agent’s and a caller’s emotional trajectory can each be derived as closed-form aggregates of the preserved distributions, without reintroducing the argmax collapse that this work has deliberately avoided. A scorecard of this kind would let supervisors compare interactions, operators, and time periods on a common, uncertainty-aware footing, replacing the coarse polarity labels of current commercial analytics with calibrated, trajectory-level measurements.

What makes such an instrument viable in a regulated setting is the *on-premise, local-first* design of the platform. Because inference, aggregation, and visualisation all run on commodity hardware within the deployment’s own boundary, raw audio need never leave the organisation, a property in sharp contrast with cloud emotion-measurement services and one whose importance is underscored by the retirement of the sole such service, the Hume AI Expression Measurement API [27], on which any externalised pipeline would have depended. A future analytics product can therefore be built around data minimisation and pseudonymisation by construction: retaining only the low-dimensional emotion distributions and derived indices rather than the speech signal, decoupling those records from caller identity through pseudonymous session keys, and exposing aggregate statistics rather than individual traces. Formalising these guarantees, a data-protection impact assessment, configurable retention and redaction policies, and an audit of what is computed where, is a necessary and well-scoped line of work to move from a research prototype to a deployable, compliance-aware analytics tool.

6.2.3. Evaluating Cold-Approach Communication: Three Application Domains

The platform’s ability to render the affective trajectory of an interaction makes it a natural instrument for studying *cold-approach* communication, unsolicited, first-contact exchanges in which the emotional response of the listener is both highly informative and difficult to measure by conventional means. Three application domains define distinct, separately developable case studies.

Marketing. In commercial cold-calling and sales pitches, the affective quality of the exchange, the listener’s shift from guardedness to interest, or from neutrality to irritation, is the primary determinant of outcome and the hardest to quantify. The platform could provide a calibrated, uncertainty-aware measurement of that trajectory, enabling A/B comparison of scripts and openings and data-driven training of representatives, while the same privacy safeguards govern the treatment of contacted parties.

Educational and didactic. A distinct and valuable role for the platform is as a supplementary analysis instrument for students and researchers in psychology and the neurosciences, who typically lack a strong computing background and for whom existing SER tooling is largely inaccessible. By exposing a faithful, uncertainty-aware emotional trajectory through a simple operator-facing interface, with audio in and an interpretable probabilistic read-out, the system lets non-technical users obtain a quantitative affective signal without writing code or training models. In this role it complements the methods such fields already rely on: an emotional-trajectory channel that can be aligned with self-report, behavioural coding, or physiological measures in an experiment, an auxiliary observation for clinical or counselling communication studies, or a teaching aid that lets psychology and neuroscience students inspect how an affective signal evolves over the course of a real interaction. The contribution here is accessibility: the same distribution-preserving analysis that this thesis develops, made usable by domain experts whose questions are about affect and behaviour rather than about the underlying signal processing.

Medical. A natural medical use is as an assistant for psychology and psychotherapy sessions, supporting the clinician’s retrospective analysis rather than intervening in the session itself. After a session, the platform could provide a faithful, uncertainty-aware reading of the patient’s emotional trajectory over its full course, when affect escalated, where it stabilised, which topics coincided with shifts, giving the therapist an objective, reviewable record to complement their own notes and recollection. Its principal value in

this setting is longitudinal: by retaining only the low-dimensional emotional trajectories under pseudonymous session keys, as described in Section 6.2.2, successive sessions of the same patient can be aligned and compared over time, surfacing slow trends, recurring triggers, or response to treatment that are difficult to perceive from single sessions alone. Such a tool would remain strictly an auxiliary instrument under clinical oversight, never a diagnostic one, and the sensitivity of the data makes the explicit consent of the patient and the on-premise, privacy-by-design guarantees of Section 6.2.2 absolute prerequisites.

Across all three domains the methodological contribution is identical, a faithful quantitative reading of an emotional trajectory, but the target affects, the relevant backends and label sets, and the ethical constraints differ enough that each warrants its own study, corpus, and validation protocol.

6.2.4. Toward Closing the Recognition Gap: Emotional TTS as a Catalyst

The broadest direction this work enables concerns a structural asymmetry in affective speech technology. On the generative side, emotional text-to-speech (TTS) has reached a high level of maturity: contemporary systems synthesise speech with controllable, convincingly expressive affect over utterances of arbitrary length. The recognition side has not kept pace. Speech emotion recognition remains dominated by short, clip-level corpora, acted and segment-bounded [11], and models trained on them generalise poorly to the medium- and long-form, naturally evolving emotional speech that generative systems now produce and that real interactions contain. The platform developed in this thesis, which already treats emotion as a distribution evolving over the full length of an interaction rather than as a per-clip label, is a natural base from which to attack this gap, articulated as a set of connected projects.

A medium- to long-form emotional speech dataset. A first project is the construction of a corpus of *medium- to long-form emotional audio*, in which affect evolves and transitions within each recording, moving beyond the short isolated clips that dominate existing benchmarks [11, 39]. The maturity of emotional TTS offers a concrete avenue here, controllable synthesis can generate long-form material with known, time-varying affective targets, providing weak supervision and controlled stimuli to complement human-annotated naturalistic recordings, with the temporal-aggregation infrastructure of this thesis serving as the labelling and inspection harness.

A purpose-built SER module. On such data, a second project is the development of a *purpose-built SER module* that models emotion as a temporal process over long context rather than as a per-segment classification, integrating directly behind the `ISERBackend` port and inheriting the distribution-preserving aggregation and visualisation developed here.

Emotionally interactive agents and related projects. Finally, closing the loop between perception and generation points toward *emotionally interactive agents* that pair the mature expressive synthesis of emotional TTS with a recognition front-end capable of reading a user’s evolving affect in real time, and adapting their response accordingly. The streaming, low-latency, distribution-aware design of this platform provides the perceptual half of such an agent. Related projects, expressive-affect benchmarking that evaluates TTS and SER on a common dimensional footing, and longitudinal studies of human reaction to affect-adaptive systems, follow naturally from the same infrastructure.

6.3. Closing Remarks

This thesis set out to close the gap between the affective-computing research community’s best speech emotion recognition models and the operational contexts that motivate them, and to do so without surrendering the probabilistic richness that those models produce and that current practice discards. The result is an open, modular, real-time platform that preserves the full posterior emotion distribution from inference through temporal aggregation to operator-facing display, across interchangeable categorical and dimensional backends unified under a single representation, and that runs on commodity hardware within the deployment’s own boundary. The future directions outlined above, technical consolidation, evolution into a privacy-respecting analytics instrument, application across the medical, educational, and marketing domains, and the longer research programme of closing the recognition gap that emotional TTS has opened, all build on the same principle that organises the present work: that emotion in speech is most faithfully, and most usefully, represented not as a label but as a distribution evolving over time.

Bibliography

- [1] M. B. Akçay and K. Oğuz. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116:56–76, 2020. doi: 10.1016/j.specom.2019.12.001.
- [2] Alpha Cephei Inc. Vosk: Offline open source speech recognition toolkit. <https://alphacephei.com/vosk>, 2020.
- [3] A. Baevski, H. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460, 2020. doi: 10.5555/3495724.3496768.
- [4] A. Baird, R. Manzelli, P. Tzirakis, C. Gagne, H. Li, S. Allen, S. Dieleman, B. Kulis, S. S. Narayanan, and A. Cowen. The NeurIPS 2023 machine learning for audio workshop: Affective audio benchmarks and novel data, 2024.
- [5] T. Bänziger and K. R. Scherer. The role of intonation in emotional expressions. *Speech Communication*, 46(3–4):252–267, 2005. doi: 10.1016/j.specom.2005.02.016.
- [6] C. Barhoumi and Y. BenAyed. Real-time speech emotion recognition using deep learning and data augmentation. *Artificial Intelligence Review*, 58(2):49, 2025. doi: 10.1007/s10462-024-11065-x.
- [7] M. Bellver et al. Using large speech models for feature extraction in cross-lingual speech emotion recognition. In *Proceedings of CLiC-it 2024*, 2024. URL <https://aclanthology.org/2024.clicit-1.31>.
- [8] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss. A database of German emotional speech. In *Interspeech 2005*, pages 1517–1520, 2005. doi: 10.21437/Interspeech.2005-446.
- [9] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan. IEMOCAP: Interactive emotional dyadic motion capture

- database. *Language Resources and Evaluation*, 42(4):335–359, 2008. doi: 10.1007/s10579-008-9076-6.
- [10] C. Busso, R. Lotfian, K. Sridhar, A. N. Salman, W.-C. Lin, L. Goncalves, S. Parthasarathy, A. R. Naini, S.-G. Leem, L. Martinez-Lucas, H.-C. Chou, and P. Mote. The MSP-Podcast corpus, 2025.
- [11] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014. doi: 10.1109/TAFFC.2014.2336244.
- [12] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, and F. Wei. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022. doi: 10.1109/JSTSP.2022.3188113.
- [13] H.-C. Chou, L. Goncalves, S.-G. Leem, C.-C. Lee, and C. Busso. The importance of calibration: Rethinking confidence and performance of speech multi-label emotion classifiers. In *Proc. Interspeech 2023*, pages 641–645, 2023. doi: 10.21437/Interspeech.2023-1113.
- [14] A. Cockburn. Hexagonal architecture. <https://alistair.cockburn.us/hexagonal-architecture/>, 2005.
- [15] A. S. Cowen and D. Keltner. Semantic space theory: A computational approach to emotion. *Trends in Cognitive Sciences*, 25(2):124–136, 2021. doi: 10.1016/j.tics.2020.11.004.
- [16] L. Devillers and L. Vidrascu. Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs. In *Interspeech 2006*, pages 801–804, 2006. doi: 10.21437/Interspeech.2006-275.
- [17] L. Devillers, L. Vidrascu, and L. Lamel. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4):407–422, 2005. doi: 10.1016/j.neunet.2005.03.007.
- [18] P. Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3–4):169–200, 1992. doi: 10.1080/02699939208411068.
- [19] M. El Ayadi, M. S. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011. doi: 10.1016/j.patcog.2010.09.020.

- [20] T. Feng, J. Lee, A. Xu, Y. Lee, T. Lertpetchpun, X. Shi, H. Wang, T. Thebaud, L. Moro-Velazquez, D. Byrd, et al. Vox-profile: A speech foundation model benchmark for characterizing diverse speaker and speech traits, 2025.
- [21] Y. Feng et al. End-to-end continuous speech emotion recognition in real-life customer service call center conversations, 2023.
- [22] Y. Gong, S. Khurana, L. Karlinsky, and J. Glass. Whisper-AT: Noise-robust automatic speech recognizers are also strong general audio event taggers. In *Proceedings of Interspeech*, 2023. doi: 10.21437/Interspeech.2023-1542.
- [23] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.
- [24] D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.
- [25] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. doi: 10.1109/TASLP.2021.3122291.
- [26] M.-H. Huang and R. T. Rust. The caring machine: Feeling AI for customer care. *Journal of Marketing*, 88(5):1–23, 2024. doi: 10.1177/00222429231224748.
- [27] Hume AI. Expression measurement API — overview. <https://dev.hume.ai/docs/expression-measurement/overview>, 2026. Accessed May 2026. Last day for new API jobs: May 14, 2026; API fully offline: June 14, 2026.
- [28] A.-R. Ispas, T. Deschamps-Berger, and L. Devillers. A multi-task, multi-modal approach for predicting categorical and dimensional emotions. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 311–317, 2023. doi: 10.1145/3577190.3614149.
- [29] Z. Jia, J. Yu, H. Long, and D. Tang. Coverage-guaranteed speech emotion recognition via calibrated uncertainty-adaptive prediction sets, 2025.
- [30] S. Jing, X. Mao, and L. Chen. Automatic speech discrete labels to dimensional emotional values conversion method. *IET Biometrics*, 8(2):168–176, 2019. doi: 10.1049/iet-bmt.2018.5016. URL <https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/iet-bmt.2018.5016>.

- [31] A. Kendall and Y. Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30, 2017.
- [32] S. G. Koolagudi and K. S. Rao. Emotion recognition from speech: A review. *International Journal of Speech Technology*, 15(2):99–117, 2012. doi: 10.1007/s10772-011-9125-1.
- [33] I. Lehtosalo and G. van Rossum. PEP 544 – protocols: Structural subtyping (static duck typing). Python Enhancement Proposal, 2017. <https://peps.python.org/pep-0544/>.
- [34] L. I.-K. Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1):255–268, 1989. doi: 10.2307/2532051.
- [35] W.-C. Lin and C. Busso. Chunk-level speech emotion recognition: A general framework of sequence-to-one dynamic temporal modeling. *IEEE Transactions on Affective Computing*, 14(2):1215–1227, 2023. doi: 10.1109/TAFFC.2021.3083821.
- [36] S. R. Livingstone and F. A. Russo. The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE*, 13(5):e0196391, 2018. doi: 10.1371/journal.pone.0196391.
- [37] R. Lotfian and C. Busso. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483, 2019. doi: 10.1109/TAFFC.2017.2736999.
- [38] R. Lotfian and C. Busso. Predicting categorical emotions by learning hierarchical discriminative features with a soft-label SVM. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5486–5490, 2019. doi: 10.1109/ICASSP.2019.8682759.
- [39] Z. Ma, M. Chen, H. Zhang, Z. Zheng, W. Chen, X. Li, J. Ye, X. Chen, and T. Hain. EmoBox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark. In *Interspeech 2024*, 2024. doi: 10.21437/Interspeech.2024-788.
- [40] R. C. Martin. *Clean Architecture: A Craftsman’s Guide to Software Structure and Design*. Prentice Hall, 2017.
- [41] A. Metallinou, A. Katsamanis, Y. Wang, and S. Narayanan. Tracking changes in continuous emotion states using body language and prosodic cues. In *2011 IEEE In-*

- ternational Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2288–2291, May 2011. doi: 10.1109/ICASSP.2011.5946939.
- [42] S. Mirsamadi, E. Barsoum, and C. Zhang. Automatic speech emotion recognition using recurrent neural networks with local attention. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2227–2231, 2017.
 - [43] S. Naini, M. Lotfidereshgi, C. Busso, et al. The interspeech 2025 challenge on speech emotion recognition in naturalistic conditions. In *Proceedings of Interspeech 2025*. ISCA, 2025.
 - [44] NICE Systems Ltd. Sentiment analysis — NICE enlighten AI. <https://www.nice.com/resources/nice-sentiments-analysis>, 2024. Accessed May 2026.
 - [45] L. Pepino, P. Riera, and L. Ferrer. Emotion recognition from speech using wav2vec 2.0 representations. *INTERSPEECH 2021*, pages 3400–3404, 2021. doi: 10.21437/Interspeech.2021-703.
 - [46] J. C. Peterson, R. M. Battleday, T. L. Griffiths, and O. Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9617–9626, 2019. doi: 10.1109/ICCV.2019.00971.
 - [47] M. K. Pichora-Fuller and K. Dupuis. Toronto emotional speech set (TESS), 2020.
 - [48] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely. The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011. URL https://kaldi-asr.org/papers/kaldi_asru2011.pdf.
 - [49] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset Shift in Machine Learning*. MIT Press, 2009.
 - [50] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.
 - [51] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980. doi: 10.1037/h0077714.
 - [52] K. R. Scherer. What are emotions? and how can they be measured? *Social Science Information*, 44(4):695–729, 2005. doi: 10.1177/0539018405058216.

- [53] O. Schröder, M. Milling, F. Burkhardt, F. Eyben, and B. Schuller. Are you sure? analysing uncertainty quantification approaches for real-world speech emotion recognition. In *Interspeech 2024*, 2024. doi: 10.21437/Interspeech.2024-977.
- [54] B. Schuller and A. Batliner. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, 61(5):90–99, 2018. doi: 10.1145/3129340.
- [55] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll. Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing*, 1(2):119–131, 2010. doi: 10.1109/T-AFFC.2010.8.
- [56] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison. Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/hash/86df7dcfd896fcdf2674f757a2463eba-Abstract.html.
- [57] A. Shendabadi, P. Izadirad, M. Salehi, and M. Bijankhan. Speech emotion recognition leveraging OpenAI’s Whisper representations and attentive pooling methods, 2026. URL <https://arxiv.org/abs/2602.06000>.
- [58] S. Steidl, M. Levit, A. Batliner, E. Nöth, and H. Niemann. Of all things the measure is man: Automatic classification of emotions and inter-labeler consistency. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 317–320, 2005. doi: 10.1109/ICASSP.2005.1415114.
- [59] A. Triantafyllopoulos, G. Keren, and B. W. Schuller. Probing speech emotion recognition transformers for linguistic knowledge. In *Proc. Interspeech 2022*, pages 146–150, 2022. doi: 10.21437/Interspeech.2022-748.
- [60] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5200–5204, 2016. doi: 10.1109/ICASSP.2016.7472669.
- [61] J.-M. Valin. A hybrid DSP/deep-learning approach to real-time full-band speech enhancement. In *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2018. doi: 10.1109/MMSP.2018.8547084.

- [62] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller. Dawn of the transformer era in speech emotion recognition: Closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10745–10759, 2023. doi: 10.1109/TPAMI.2023.3263585.
- [63] N. Wang and D. Yang. Speech emotion recognition using fine-tuned wav2vec2.0 and neural controlled differential equations classifier. *PLOS ONE*, 20(2):e0318297, 2025. doi: 10.1371/journal.pone.0318297.
- [64] T. Wong, Z. Talat, H. Aldarmaki, and A. Field. Unrequited emotions: Investigating the gaps in motivation and practice in speech emotion recognition research, Apr. 2026.
- [65] H. Wu, H.-C. Chou, K.-W. Chang, L. Goncalves, J. Du, J.-S. R. Jang, C.-C. Lee, and H.-Y. Lee. EMO-SUPERB: An in-depth look at speech emotion recognition, 2024.
- [66] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K.-t. Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H.-y. Lee. SUPERB: Speech processing universal PERformance benchmark. In *Proc. Interspeech 2021*, pages 1194–1198, 2021. doi: 10.21437/Interspeech.2021-1775.

List of Acronyms

API Application Programming Interface.

ASR Automatic Speech Recognition.

BiLSTM Bidirectional Long Short-Term Memory.

CCC Concordance Correlation Coefficient.

CORS Cross-Origin Resource Sharing.

ECE Expected Calibration Error.

IPC Inter-Process Communication.

JSD Jensen–Shannon Divergence.

KPI Key Performance Indicator.

LSTM Long Short-Term Memory.

MFCC Mel-Frequency Cepstral Coefficients.

OOD Out-of-Distribution.

PCM Pulse-Code Modulation.

RTF Real-Time Factor.

SER Speech Emotion Recognition.

SSL Self-Supervised Learning.

STT Speech-to-Text.

UAR Unweighted Average Recall.

V–A Valence–Arousal.

VAD Voice Activity Detection.

WAR Weighted Average Recall.

List of Figures

- 2.1 Russell’s circumplex model of affect [51] with the eight emotion anchors adopted from Jing *et al.* [30] (coordinates in Table 2.1). *Neutral* is placed at the origin; peripheral emotions lie on the unit circle ($r = 1$). 15
- 3.1 Three-layer hexagonal architecture [14, 40]. All source-level imports flow inward; the domain core carries no external dependencies. 23
- 3.2 End-to-end data-flow of the real-time pipeline (implementation in Section 3.3). Denoised audio is routed in parallel to the STT and SER branches; the aggregator fuses both into three output granularities; session parameters are injected at start (dashed). 30
- 3.3 The **SessionPage** during a live call-centre session, top to bottom: header bar (label, timer, live badge); emotional timeline strip (the affective arc as an interpolated colour sequence); KPI cards; a two-column area pairing the transcript panel (left) with the analysis panel (right, holding the Emotion Circumplex and ranked Probability Distribution bars); and a docked audio player for post-session playback. 38
- 4.1 Argmax collapse: the full posterior $\mathbf{p}(e)$ (left) is reduced to the dominant class only (right); secondary affect mass, calibration confidence, and distributional trajectory are irrecoverably lost (dashed box). 44
- 4.2 Frame aggregation at $c_k=2.0$ s ($w_s=2.0$ s illustrative, deployed default 3.0 s; $h_s=0.5$ s). The yellow strip is the frame interval $[c_k - h_s/2, c_k + h_s/2]$. Windows $e^{(0)}-e^{(4)}$ intersect it ($|W_k|=5$); $e^{(5)}$ starts too late and is excluded. Since $|W_k| > 4$, the four windows closest to c_k are uniformly averaged into $\bar{\mathbf{p}}(c_k)$ (3.3). 47

4.3	Utterance enrichment for $[u_s, u_e]=[1.0, 3.5]$ s ($w_s=2.0$ s illustrative, deployed default 3.0 s). Each window's weight is its overlap with the utterance, normalised: $w(e) = \text{ov}(e) / \sum_{e'} \text{ov}(e')$ (3.4) (blue = overlap, gray = outside). The two fully contained windows ($e^{(2)}$, $e^{(3)}$, bold) reach the maximum weight $w=0.25$. Weighting is thus non-uniform, unlike the equal-weight frame aggregation of Figure 4.2.	48
4.4	Distances-to-probabilities algorithm (Section 4.3.2) for $(v, a)=(0.00, 0.65)$, equidistant from Anger and Surprise. Left: $r=0.65 > 0.5$ gives $p_N=0$; the three labelled distances drive the softmax ($\tau=0.45$). Right: since $d_A=d_S=0.37$, the distribution is bimodal, Anger and Surprise each receive 30%.	50
5.1	Predictions at each peripheral anchor. Left: anchor self-mass against the uniform baseline $1/7$ (dashed), every anchor ranked first. Right: anchor self-mass and nearest neighbour stacked.	59
5.2	Left: Neutral probability vs radius along sixteen azimuths (grey) against the analytic curve (black), extinguished at $r = 1/s = 0.5$. Right: entropy along the same sweeps, peaking in the Neutral-to-peripheral handover. . .	60
5.3	Distributions at the chord midpoints of the seven adjacent-anchor pairs: the two bars are the two balanced modes of each output distribution. . . .	61
5.4	Global behaviour on the valence–arousal disc: dominant emotion (left), entropy (centre), and Neutral probability (right); unit circle and the seven anchors overlaid.	62
5.5	Structural invariance under the backend swap. Left: weighted-frame timestamps of the three backends on the same 0.5 s grid. Right: live-event, weighted-frame, and final-utterance counts identical across backends. . . .	65
5.6	Left: per-window inference latency per backend against the $h_s = 0.5$ s hop budget. Right: end-to-end real-time factor against the real-time limit, annotated with the equivalent speedup ($1/\text{RTF}$).	66
5.7	Streaming behaviour under four (w_s, h_s) configurations: live inference completions (ticks) and weighted frames, each drawn from its centre to its release instant.	67
5.8	Stability of the raw and aggregated weighted-frame streams for the four sessions (EmoDB BiLSTM): inter-frame jitter (left) and dominant-label flip rate (right), per-session jitter reduction annotated.	68

5.9	Distributional structure retained by utterance-level aggregation. Left: one dimensional-backend utterance, aggregated distribution (filled) against the full-confidence bar a majority vote reports (red hatch). Centre and right: entropy and secondary-emotion mass of every BiLSTM utterance (mean marked), well above the zero (dashed red) a majority vote assigns.	69
5.10	Cumulative distribution of the post-hoc revisions an eager release policy would require, pooled over the four sessions (EmoDB BiLSTM); the safe-until guard reduces every one to zero at a deterministic cost of w_s seconds of latency.	70
5.11	Inference events attributed to the silence and voiced regions of the synthetic silence-padded session against the VAD threshold. Any $\theta_{vad} \geq 0.1$ removes the spurious silence inferences, and every threshold in $[0.3, 0.5]$ retains identical voiced coverage; the default $\theta_{vad} = 0.5$ is therefore the principled decision boundary at no coverage cost, while lower thresholds only add low-confidence pause and breath windows.	72
5.12	Stability-latency trade-off governed by the window duration (EmoDB BiLSTM, three sessions): weighted-frame jitter falls with the window (left; faint lines per session, markers the mean), first-frame release latency grows linearly (right).	73
5.13	Resolution-compute trade-off governed by the hop duration: weighted-frame jitter on the BiLSTM (left) and end-to-end real-time factor of both backends (right); the large Whisper model exceeds the real-time limit only at $h_s = 0.25$ s, so $h_s = 0.5$ s is the finest cadence every backend sustains. . .	74
5.14	Sensitivity to the distance temperature over $[0.10, 3.00]$: anchor self-mass, midpoint pair share, and quarter-point secondary share, with the admissible region satisfying both criteria.	75
5.15	Sensitivity to the neutral radius scale: measured Neutral extinction radius against the analytic $1/s$, and the Neutral-to-peripheral handover radius. . .	75
5.16	Synthetic rendering map of the circumplex centroid over the valence-arousal plane: radial transfer (left), monotone and confidence-compressed along every anchor-covered azimuth, and angular transfer (right), tracking the model azimuth within the anchor-set quantisation.	78
5.17	Rendered circumplex centroids of the three sessions (EmoDB BiLSTM): per-session mean ($\times$) and $\pm 1\sigma$ ellipse localise each state distinctly: Happiness towards the high-valence anchor, Neutral near the origin, Stress at the <i>Anger</i> anchor.	79

- 5.18 Operator timeline colour bands for the three sessions (EmoDB BiLSTM):
each renders as a distinct colour signature, yellow for Happiness, grey for
Neutral, red (*Anger*) for Stress, with the dominant reading and its frame
share annotated. 80
- 5.19 Normalised deviation of the interpolated frame colour from the dominant-
only colour against the secondary-emotion mass of the frame. 80

List of Tables

2.1	Circumplex anchor coordinates from Jing <i>et al.</i> [30]; bold = adopted by the dimensional SER backend.	14
3.1	Domain event types. All timestamps are in seconds from session start. . . .	25
5.1	Algorithm output at the seven peripheral circumplex anchors: mass assigned to the anchor's own emotion, its rank, its concentration relative to the uniform peripheral baseline 1/7, and the second-ranked emotion (always an angularly adjacent anchor).	59
5.2	Scorecard of the distances-to-probabilities algorithm: each required geometric property, its acceptance criterion, and the measured value that satisfies it.	63
5.3	The backend swap in terms of what it leaves invariant and what it determines: selecting a different acoustic model through a single configuration field changes only the model and its immediate consequences, while the entire surrounding pipeline is reused unchanged.	64
5.4	The same evaluation session processed by the three backends through the identical pipeline code path: label-space size, live events and weighted frames emitted, model load time, per-window inference latency (mean $\pm$ standard deviation over fifteen repetitions), end-to-end real-time factor, and the equivalent speedup over real time (1/RTF).	64
5.5	Scorecard of the probabilistic temporal aggregation: each claimed property, its acceptance criterion, and the measured value on the EmoDB BiLSTM stream.	71
5.6	Recommended operating configuration of the platform: the range explored by each sweep, the selected operating point, and the quantified benefit it delivers to the operator-facing signal.	76

5.7	Transcript panel content for the Neutral evaluation session (EmoDB BiL-STM): a curated excerpt of the incremental partial pieces, each with its transcript delta, dominant emotion, and probability (the operator-facing confidence), the consolidated final utterance closing each group; vertical dots mark omitted partial runs, and the per-piece entropy is retained in the evaluation notebook as an analysis measure.	82
5.8	Synthesis of the experimental evidence by contribution.	84

Acknowledgements

Un sentito ringraziamento va al mio relatore, il Prof. Alessandro Campi, non solo per avermi guidato nello sviluppo di questa tesi, ma anche per la sincerità con cui mi ha sempre consigliato e per le tante opportunità che mi ha messo davanti.

Un grazie altrettanto sincero va all'Ing. Michele Di Dalmazi, sempre disponibile in ogni momento, per il supporto e il confronto costante durante lo sviluppo del lavoro.

Alla mia famiglia va il pensiero più speciale.

A mio padre, che ha sempre visto in me il suo investimento più grande, va la mia riconoscenza più profonda: nei momenti di maggiore dubbio non mi ha mai voltato le spalle, rimanendo la mia ancora di salvezza.

A mia madre, che con quattro figli ha sempre trovato per me un posto importante, va la mia gratitudine per l'affetto e la presenza di ogni giorno.

Un pensiero ai miei amici, capaci di alleggerire il carico e di rendere piacevoli le varie tappe di questo percorso.

Grazie a Giovanni, socio e amico fraterno: con lui ho imparato che la via più battuta non è sempre la migliore, e che le strade secondarie sono spesso quelle che meritano di essere percorse e raccontate.

A chiunque, in un modo o nell'altro, abbia reso questo percorso ciò che è stato: il mio grazie più sincero.

