



POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

EXECUTIVE SUMMARY OF THE THESIS

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

LAUREA MAGISTRALE IN COMPUTER SCIENCE AND ENGINEERING - INGEGNERIA INFORMATICA

Author: PIETRO MORCAVALLO

Advisor: PROF. ALESSANDRO CAMPI

Co-advisor: ING. MICHELE DI DALMAZI

Academic year: 2025-26

1. Introduction and Motivation

1.1. Motivation and Context

Spoken interaction is now the primary interface in a growing range of high-stakes services, where inferring not only *what* users say but *how* they say it is a critical capability. Commercial contact-centre analytics has nonetheless converged on a paradigm that diverges structurally from the research tradition: sentiment is inferred from speech-recognition transcripts rather than from the acoustic signal, and emotional state is reduced to a coarse polarity label or an opaque scalar rather than a calibrated distribution. On the modelling side the picture is the opposite: self-supervised representations, in particular Whisper [3], substantially outperform hand-crafted pipelines, though categorical models generalise poorly outside their training distribution, which has motivated dimensional representations such as the valence-arousal circumplex [4]. A gap therefore separates benchmarks from deployment: most systems are built for offline, segment-level processing, few address streaming at low latency, and collapsing a posterior to its argmax discards what is needed to capture ambiguous states. A recent survey finds

that roughly 17% of SER studies cite call-centre applications as their motivation yet evaluate on acted, controlled-environment corpora that do not reflect real contact-centre audio [5], a mismatch this thesis is designed to close. Since neural classifiers are systematically overconfident [1], that discarded signal is precisely what an operator needs in order to calibrate trust.

1.2. Objective and Contributions

The thesis addresses four requirements in combination. The *infrastructure* requirement calls for an end-to-end streaming pipeline that is a deployable platform rather than a research artefact. *Posterior preservation* requires retaining the full distribution at each inference step, since the secondary emotions and confidence gradients otherwise discarded are the operationally informative signal. *Preservation across time* requires combining frame-level predictions into stable utterance-level outputs without the distributional erasure of majority voting, so that the affective trajectory of an interaction stays legible instead of being replaced by a snapshot of the current state. And *backend heterogeneity* must be abstracted behind a common interface, since categorical classifiers produce distributions

over incompatible label sets while dimensional models produce continuous coordinates incommensurable with them by construction.

Five contributions develop the resulting principle, the preservation of the full posterior end-to-end: **(C1)** an open, modular real-time SER platform integrating ingestion, denoising, VAD-gated sliding-window inference, speech recognition, and temporal aggregation, with backends behind a common port interface and selected at session time; **(C2)** a probabilistic temporal aggregation algorithm fusing window outputs by overlap-weighted averaging; **(C3)** a unified probabilistic representation for both model families via a geometric *distances-to-probabilities* inversion in the circumplex plane; **(C4)** documented engineering decisions, each stated with the trade-off it resolves; and **(C5)** an interpretability-first visualisation design in which no visual element reduces the distribution to a dominant label.

2. Background and Related Work

Two representational frameworks coexist in affective computing. *Categorical* models describe emotion as membership in a finite set of classes and map directly to actionable states; *dimensional* models embed it in the continuous valence–arousal plane of Russell’s circumplex [4], more faithful to the graded nature of affect. Arousal is markedly easier to predict than valence, which depends on subtle spectral qualities and linguistic content. This work builds on Whisper [3], whose encoder, though trained for speech recognition on roughly 680,000 hours of multilingual audio, acquires broad acoustic representations that transfer to paralinguistic tasks while mitigating the domain mismatch that affects models trained on narrow corpora. Two checkpoints fine-tuned on MSP-Podcast are used, one categorical and one regression-based. Jing *et al.* [2] formalised the *forward* direction between the two frameworks, assigning each category a canonical polar angle; this thesis requires the complementary *inverse* and adopts those anchor positions as its geometric basis.

Aleatoric uncertainty is structurally high in SER: annotators frequently disagree on the dominant emotion, to the point that a classifier at roughly 60% raw accuracy on a four-class spon-

taneous task already matches the average human annotator, and preserving that disagreement as a soft-label distribution improves downstream robustness. Since neural classifiers are more-over overconfident [1], the system preserves the maximum softmax probability on every event and reports Shannon entropy as an uncertainty proxy. These difficulties compound in deployment, where audio is spontaneous and cross-corpus degradation is systematic and worse for valence, motivating both the Whisper backbone and the realistic, conversational-length evaluation material.

3. System Design and Architecture

The system follows the hexagonal architecture and the dependency rule of Clean Architecture: all source-level imports travel inward and never in reverse. The **domain layer** holds pure Python types and functions with no deep-learning or web-framework dependencies; the **application layer** orchestrates them through batch and streaming use cases via a dependency-injection container; the **infrastructure layer** implements all concrete I/O. The rule is enforced structurally rather than by convention: the domain package admits no framework names at runtime, a directly verifiable invariant. Ports are structural-typing protocols rather than abstract base classes, which decouples the port definition from the adapters entirely and makes back-end interchangeability an architectural property rather than a configuration convenience.

Live audio is denoised, then fanned out to the SER and transcription pipelines, which consume the same stream at different granularities: SER advances by 0.5 s hops over 3 s windows, transcription by 10 ms frames. Aggregation merges the two by *timestamp* rather than by arrival, since a first-in-first-out merge would emit an utterance before the windows overlapping it had been seen. A guard reinforces this, releasing a weighted frame centred at c_k only once the latest SER event ends at or after $c_k + w_s$, since emitting earlier would average over an incomplete set of windows and manufacture spurious transitions. It costs a minimum latency of one window duration and is bypassed at end of stream.

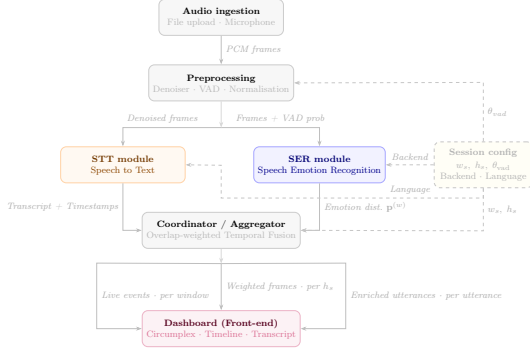


Figure 1: End-to-end data flow. Denoised audio feeds the transcription and SER branches in parallel; the aggregator fuses both into three output granularities.

4. Original Contributions

4.1. Probabilistic Emotion Representation

Every stage carries the full posterior $\mathbf{p}$ over the emotion set $\mathcal{E}$ rather than a reduced label. The Shannon entropy $H(\mathbf{p}) = -\sum_e p_e \log_2 p_e$ characterises how concentrated each prediction is: zero for a point mass, $\log_2 |\mathcal{E}|$ bits for a uniform distribution. Near the maximum no emotion dominates, and there an argmax reduction is not merely lossy but actively misleading, reporting a confident label where the model expressed none, which is why $\mathbf{p}$ is propagated through every stage instead of being collapsed at the first opportunity.

4.2. Weighted Aggregation Algorithm

The inference loop emits one event per hop h_s , each covering a window of duration w_s , so consecutive windows overlap by $w_s - h_s$. Aggregation turns this dense stream into two representations, both required to preserve the full distribution, to handle variable overlap, and to be produced online without access to future events. *Weighted frames* form a regular grid at resolution h_s : for a frame centred at c_k , the candidate set W_k collects every event whose window intersects $[c_k - h_s/2, c_k + h_s/2]$, capped at the four closest to c_k , and the frame distribution is their uniform average, uniform weighting being appropriate precisely because the construction of W_k already guarantees comparable overlap. *Enriched utterances* align with transcription

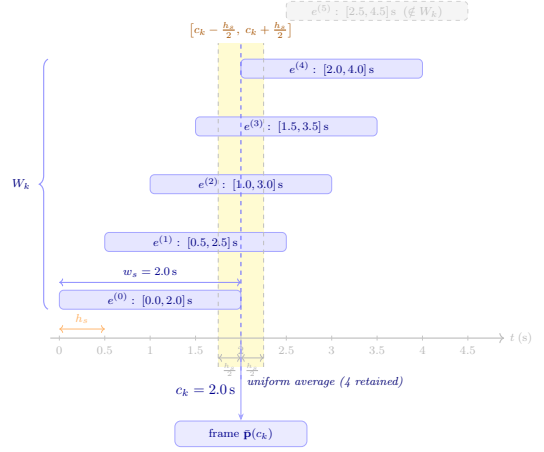


Figure 2: Frame aggregation at $c_k=2.0$ s ($w_s=2.0$ s illustrative, deployed default 3.0 s; $h_s=0.5$ s). Five windows intersect the frame interval (shaded); since $|W_k| > 4$, the four closest to c_k are uniformly averaged.

output, where overlap is genuinely variable: a window covering 0.2 s of a 3 s utterance should not count as much as one covering 1.8 s. The utterance distribution is therefore the overlap-weighted average

$$\bar{\mathbf{p}}_{\text{utt}} = \frac{\sum_e \text{ov}(e) \mathbf{p}_e}{\sum_e \text{ov}(e)}, \quad (1)$$

with $\text{ov}(e)$ the temporal overlap between event e and the utterance. Each frame and utterance also carries the count of contributing events, read downstream as a reliability signal.

4.3. Dimensional SER via Circumplex Inversion

Integrating a dimensional model requires reconciling its continuous output with the categorical representation shared by the rest of the pipeline. The *distances-to-probabilities* algorithm does this in two steps over the anchor set on the circumplex, from which *Other* is excluded as having no geometric interpretation. First, the Neutral mass is set from the radius $r = \sqrt{v^2 + a^2}$ of the model output, as $p_{\text{Neutral}} = \max(0, 1 - rs)$. A radius-based treatment is a necessity, not a convenience: the origin is equidistant from every peripheral anchor, so a distance-based formula would assign uniformly high probability to all peripheral emotions exactly where the output should concentrate on Neutral. Second, the remaining mass is distributed by a temperature-scaled softmax

over negative distances,

$$p_e = (1 - p_{\text{Neutral}}) \frac{\exp(-d_e/\tau)}{\sum_{e'} \exp(-d_{e'}/\tau)}, \quad (2)$$

with d_e the Euclidean distance from (v, a) to anchor e .

The temperature is bracketed from both sides: too low and the peripheral mass collapses onto the nearest anchor, degenerating into Voronoi assignment; too high and even confident predictions come out near-uniform. The resulting semantic property is that a point midway between two adjacent anchors yields a *bimodal* distribution over those two emotions rather than peaking on whatever lies at the geometric centroid between them, correctly representing uncertainty between neighbouring affective states.

4.4. Interpretability-First Visualisation

The operator-facing layer keeps distributional information visible at every granularity. The **Emotion Circumplex** plots the probability-weighted centroid of all active anchors rather than the argmax emotion, so a distribution split between *Anger* and *Sadness* places the crosshair in the upper-left quadrant instead of snapping to either. The **interpolated-colour timeline** maps each frame’s posterior onto one perceptual channel by averaging the canonical colours of the emotions present: concentrated frames appear as saturated hues while frames whose mass is split across opposing hues desaturate toward grey, giving an immediate reading of spread with no numeric readout. The **transcript panel** attaches to each utterance its overlap-weighted distribution from equation (1). All three 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

5.1. Experimental Setup

Two complementary kinds of material drive the evaluation. Real speech exercises the full pipeline: four Italian sessions from two speakers reading a call-centre-style text under Happiness, Neutral, and Stress conditions, each 27.3–37.8 s long, the conversational-turn regime a monitoring deployment actually observes; short acted

clips of two to three seconds fall below the analysis window and would not drive those stages at all. Synthetic input is used where a contribution is defined by deterministic geometry: the conversion is evaluated directly on valence–arousal coordinates, so every expected property follows analytically from its construction, and the gating experiment embeds a session between two stretches of digital silence so that each inference event is attributable to a voiced or unvoiced region.

Distributions are compared by Jensen–Shannon divergence (JSD) in bits and uncertainty reported as Shannon entropy. Stream stability is measured by *inter-frame jitter*, the mean JSD between consecutive distributions, and by the *dominant-label flip rate* per minute, both quantifying the visual restlessness an operator would experience; feasibility by the *real-time factor* (RTF), which must stay below one. All experiments run the production code path on an Apple M1 Pro laptop, as notebooks regenerating every figure and table in one execution, with $w_s = 3.0$ s, $h_s = 0.5$ s, $\theta_{vad} = 0.5$, $s = 2.0$, and $\tau = 0.45$ unless a sweep varies them.

5.2. Geometric Validation of the Conversion Algorithm

Evaluated over the full circumplex plane, the conversion satisfies each property its construction requires. At all seven peripheral anchors the corresponding emotion is the unique mode, carrying on average 4.1 times the uniform peripheral baseline, and the residual mass is semantically placed rather than scattered: the anchor and its angularly adjacent neighbour together hold 78 % of the distribution. Neutral mass follows the analytic radial law to within numerical precision and is direction-invariant. At all seven midpoints between adjacent anchors the distribution is bimodal over exactly the two flanking emotions (balance 1.000), confirming that genuine ambiguity is represented as such. Across the plane the mapping is smooth, neighbouring grid points differing by a mean JSD of 1.4×10^{-4} bits, and the operating temperature lies inside an admissible region of $[0.18, 0.55]$, so the adopted value is interior rather than marginal.

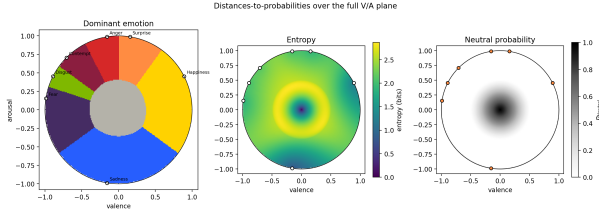


Figure 3: Conversion output over the valence-arousal disc: dominant emotion, entropy, and Neutral probability, with the seven anchors overlaid.

Backend	Cls.	Latency	RTF
EmoDB BiLSTM	7	78 ± 20 ms	0.43
Whisper categ.	9	427 ± 29 ms	0.75
Whisper dimens.	-	427 ± 20 ms	0.74

Table 1: Per-window inference latency (mean $\pm$ standard deviation over fifteen repetitions) and end-to-end real-time factor for the three backends on the same session.

5.3. Platform Demonstration: Interchangeable Backends

One session was processed by three backends through the identical code path, the backend selected by a single configuration field. The pipeline emitted exactly the same event structure in all three runs, 42 live events, 48 weighted frames on the same 0.5s grid, and two enriched utterances, because temporal behaviour is owned by the pipeline stages rather than by the model. What the swap determines is the label space, the inference cost, and the character of the output distribution. That last point is the substantive one: 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 genuine difference in classifier character that survives intact precisely because no stage collapses the distribution.

Table 1 reports the cost. Each window inference must complete within one hop of 0.5s and the pipeline must consume audio at least as fast as it arrives; both budgets are met, and even the 1.5-billion-parameter models sustain live operation on a consumer laptop at roughly 1.3 times real time. This measures the SER pipeline alone, since transcription runs on a separate thread, so

the margin is headroom that co-resident workload must fit within, and as a single-run measurement it establishes feasibility on the target hardware rather than a hardware-independent bound.

5.4. Validation of Probabilistic Temporal Aggregation

Aggregation was assessed on the four sessions against two alternatives: the raw inference stream it replaces, and the majority-vote reduction it refuses. Against the raw stream it stabilises the output, inter-frame jitter falling by 70 % (minimum 64 %) and the dominant-label flip rate dropping from 10 to 7 per minute, a directly perceptible reduction in the restlessness of the operator’s view. Against argmax it retains what the reduction would destroy: utterance entropy averages 1.08 bits and secondary-emotion mass 0.33 where majority voting yields exactly zero for both, and the JSD between each aggregated distribution and its own one-hot reduction averages 0.20 bits, which quantifies the information a label-only pipeline discards at every utterance. The release guard behaves as designed, every released frame being final, whereas an eager policy would have revised already-displayed frames by a mean of 0.07 bits.

5.5. Fidelity of the Visualisation Layer

Because the circumplex view draws a closed-form function of the distribution, its fidelity can be characterised exactly, over every reachable model state and independently of any audio. Along every azimuth covered by the anchor set the rendered radius grows strictly monotonically with the model radius (per-azimuth Spearman $\rho = 1.00$), so order relations of emotional intensity are always preserved, while the absolute radius is compressed within 0.90 of the unit disc, not an artefact but the confidence coding of the view: a broad distribution pulls the centroid inward exactly when the model is uncertain. Rendered azimuth tracks model azimuth with a mean absolute deviation of 17.6° , bounded by half the local angular gap between anchors. Exercised on real material, the map localises the three sessions into distinct and correct regions: Happiness on the high-valence side (mean valence +0.38), Neutral near the origin, and Stress

in 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.

6. Discussion and Conclusions

6.1. Discussion

The connections between the results reinforce one another. The platform result gains its weight from the aggregation result: because temporal behaviour is owned by the pipeline rather than by the model, the stability and preservation guarantees were established once and then held, unchanged, for backends of radically different output character, exactly the property an open architecture must offer to admit future models without re-deriving its guarantees. The visualisation claim, in turn, rests on mathematics rather than preference: all three operator-facing components are closed-form functions of the preserved distributions, with the timeline colour a provably faithful encoding of secondary-emotion mass ($r \approx 0.8$). Taken together, the five contributions substantiate a single claim, that distributional, interpretability-first SER is not a laboratory aspiration but a working system on commodity hardware.

The strength of these results is that the guarantees an operational deployment most depends on are precisely the ones that generalise. The temporal, computational, and geometric properties are established by construction and independently of any speaker, language, or emotional content, so they carry over unchanged to new domains; the geometric conformance of the conversion, in particular, holds exactly across the whole circumplex plane rather than for a sampled subset. What the present evaluation deliberately scopes for in-domain confirmation is the one model-specific dimension, the specific emotional content reported on spontaneous traffic, which is exactly what the fine-tuning path of the next section is designed to sharpen. The evaluation therefore establishes the architectural and algorithmic foundations firmly, and hands a clear, well-defined agenda to the deployment that builds on them.

6.2. Conclusions and Future Work

This thesis has presented a modular real-time SER platform for call-centre deployment, bridging the gap between the research community’s best models and the operational contexts that motivate them, and the evidence above supports the claim: the five contributions taken together demonstrate that distributional, interpretability-first SER is deployable today on commodity hardware.

The platform is a foundation rather than an end point. The most immediate consolidation is backends fine-tuned on call-centre speech in the target language, together with the per-utterance calibration an annotated in-domain corpus would enable; beyond that, its ability to run entirely within a deployment’s own boundary positions it as a quantitative analytics instrument under privacy constraints. Both build on the principle organising this work: that emotion in speech is most faithfully represented not as a label but as a distribution evolving over time.

References

- [1] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.
- [2] Shaoling Jing, Xia Mao, and Lijiang Chen. Automatic speech discrete labels to dimensional emotional values conversion method. *IET Biometrics*, 8(2):168–176, 2019.
- [3] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.
- [4] James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.
- [5] Taryn Wong, Zeerak Talat, Hanan Aldarmaki, and Anjalie Field. Unrequited emotions: Investigating the gaps in motivation and practice in speech emotion recognition research, April 2026.