Paper accepted for presentation at CONVERSATIONS 2025 – the International Symposium on Chatbots and Human-Centred AI, hosted at TU Lübeck, Germany, November 12-13, 2025.

From Voice to Feeling: Lessons Learned from a Hybrid Deep-Learning Model for Spanish Emotion Recognition in Virtual Assistants

Rafael del-Hoyo-Alonso[1][0000−0003−2755−5500], Gorka Labata-Lezaum[1], Óscar Rubio-Martín[2], Sergio Pastor[1], Beatríz Franco[1], Javier Marro[2], Sergio Herrero[2], Óscar Peralta[2], Elisa Amorós[2], and Héctor Paz[2]

[1] Aragon Institute of Technology, 50008 Zaragoza, Spain
{rdelhoyo,glabata,spastor,bfranco}@ita.es
[2] IMASCONO, Zaragoza, Spain
{oscar,javier.marro,sherrero,oscarp,elisa,hp}@imascono.com

Abstract. We present a pragmatic, real-time Spanish emotion-recognition module for virtual assistants that fuses speech- and text-based predictions under a precision-first objective. Our pipeline combines a curated Spanish speech dataset (1,564 utterances; imbalanced) with 88k chat turns and integrates into the V·E·G·A avatar platform via REST. To prevent spurious affect displays, the assistant defaults to *Neutral* unless calibrated confidence thresholds are met, safeguarding user experience under a < 2 s constraint. On our splits, a Conv1D⇒BiLSTM audio model outperformed an AST variant within acceptable latency. Although the text branch underperforms in isolation, it adds value as a veto/fallback and in noisy ASR settings. We discuss limitations (dataset size and acted speech) and propose a calibrated late-fusion alternative plus an A/B evaluation blueprint. Future work prioritises ethically sourced in-the-wild Spanish data, multilingual training, and tri-modal fusion (audio–text–face).

Keywords: speech emotion recognition · text emotion recognition · calibrated fusion · virtual assistants · Spanish · precision-first · real-time

1 Introduction

Although chatbots have a long history, current usefulness surged with Transformer-based models [11]. Public interest spiked in 2023, and in a preregistered two-player Turing-test setting GPT-4 was judged human in 54% of dialogues [7]. Face-to-face interaction conveys much emotion through non-verbal cues (intonation, facial gestures), and overlooking these signals can reduce perceived empathy [9]. Given the scarcity of Spanish resources and possible cultural differences in affect expression, reliable emotion recognition is essential for warmer human–machine interaction. Passing a text-based Turing test does not guarantee a satisfying experience. Empathy—recognising and responding to a user's emotional state—becomes critical when interactions go beyond information retrieval.

In customer service it helps handle frustration; in healthcare and elder care it can prevent harmful mismatches; in education it supports adaptive pacing; and even in retail or smart-home scenarios it fosters trust. Emotionally oblivious systems (e.g., cheerful replies to stressed users) alienate people and erode adoption.

This position paper presents a precision-first Spanish emotion pipeline with neutral-by-default safeguards and calibrated thresholds under $< 2\,\mathrm{s}$ latency; its deployment in the V·E·G·A avatar platform where emotion signals adapt wording, prosody, and non-verbal behaviour; a transparent account of data constraints with per-class metrics and audio/text/fusion ablations; and a lessons-learned synthesis covering thresholding, data collection under constraints, analytics, and an A/B evaluation blueprint.

2 Our Approach

We built a two-branch pipeline that combines speech- and text-based emotion recognition. Each branch infers a label and confidence independently; the assistant adapts only when both exceed calibrated thresholds and agree, otherwise defaulting to *Neutral* to protect user experience under near-real-time constraints ($<$ 2 s). Branches were designed and validated separately and, at deployment, pass predicted labels and probabilities as context to the dialogue manager.

Data. Spanish speech resources meeting our timeline and licensing needs were limited. EmoSPeech [10] was unavailable at project start and restricts commercial use; ELRA's Emotional Speech Synthesis Database [6] was costly and speaker-limited. We therefore recorded 40 chatbot-style sentences read by 12 speakers across six Ekman emotions plus *Neutral* [4], yielding 840 clips (2–4 s), and added 722 Creative Commons–licensed Spanish excerpts (2–5 s) that were isolated and manually labelled. After QC and de-duplication, the labelled YouTube subset was highly imbalanced: Neutral n=487 (68%), Angry 102 (14%), Sad 37 (5%), Happy 33 (5%), Surprise 31 (4%), Fear 18 (3%), Disgust 7 (1%). For text, lacking Spanish chatbot emotion corpora, we semi-automatically annotated proprietary IMASCONO logs using a multilingual BERT classifier [3] and collapsed 1–5 sentiment to a 1–3–5 scheme via sequence trends.

Acoustic modelling. We profiled features on public English corpora (IEMOCAP, RAVDESS) and our data [2, 8], selecting Mel spectrograms and MFCCs as core inputs after statistical screening [1]. An Audio Spectrogram Transformer baseline [12] underperformed our Conv1D⇒BiLSTM on our splits. The AST consumed Mel spectrograms; the Conv1D⇒BiLSTM captured local spectral patterns and long-range temporal dependencies, using Mel/MFCC features extracted with *librosa*, mono 16 kHz audio, and fixed-length stacking. Although roughly 10× slower than AST, it met the latency budget and With 2,836 samples, the model retrains in 108 seconds.

Evaluation, calibration, and fusion. We used a 70/15/15 speaker-level split to avoid leakage and optimised for precision with F_β at β=0.2. Thresholds were selected on validation PR curves under a cost sensitive to false positive emotions, and post-hoc temperature scaling improved probability calibration. The production policy requires agreement and confident predictions; a reliability-weighted late fusion (audio weight α, text $1-\alpha$) is also considered to increase responsiveness while preserving a precision-first objective.

3 Implementation in V·E·G·A

V·E·G·A is a platform for virtual avatars integrating speech-to-text, large language models, text-to-speech, and emotion recognition [5]. Avatars support more natural interactions, with real-time subtitles and conversation history across more than ten languages.

Avatars may be hyper-realistic or stylised (Figure 1), with dynamic gestures and lip synchronisation based on visemes. A web-based customisation environment (V·E·G·A Trainer) enables configuration of appearance, personality, knowledge, and content. Current implementations include roles such as tourism guide, children's storyteller, real-estate agent, secretary, and sales agent.

Fig. 1. Example V·E·G·A avatars with different styles

Sentiment/emotion analysis enables the avatar to adjust four aspects of behaviour: verbal expression (word choice), vocal tone, facial expressions, and body animations.

Emotion-driven adaptations in V·E·G·A. When the fused classifier meets the calibrated threshold, V·E·G·A applies a 4D adaptation: (i) **verbal** (lexical choice, directness, escalation policy), (ii) **vocal** (pitch range, speech rate, intensity), (iii) **facial** (FACS-inspired viseme-to-expression blends), and (iv) **body** (gesture amplitude and tempo). For example, on *Angry*, the system reduces speech rate, acknowledges frustration, shortens instructions, and increases hand openness; on *Sad*, it softens intensity, increases pausing, and uses supportive phrasing;

on *Happy*, it mirrors enthusiasm with higher pitch variability and positive reinforcement. A rule-based safety layer prevents incongruent combinations (e.g., cheerful voice with apologetic wording).

4 Results and Evaluation Plan

We restricted the audio branch to four emotions—*sad*, *happy*, *angry*, and *neutral*—as *fear*, *surprise*, and *disgust* were rare and increased confusion without practical benefit. To protect task performance, we adopted a precision-first objective using F_β with $\beta = 0.2$ and a neutral-by-default policy. With a 70/15/15 speaker-level split and confidence gating, the best Conv1D⇒BiLSTM achieved $F_{\beta=0.2} = 0.891$ and accuracy $= 0.773$ at a 0.8 threshold, whereas the best AST-style Transformer reached $F_{\beta=0.2} = 0.733$ and accuracy $= 0.674$ at 0.5. For text, we collapsed labels to *positive*, *neutral*, and *negative* and trained on 88,244 samples; overall precision was 0.50 and accuracy 0.56, with neutrality hardest to resolve but few opposite-pole errors. Audio-only delivered strong precision on *angry* and *happy* while struggling on *sad*; text-only was weaker but useful as a veto or fallback, particularly with unreliable ASR. Conservative agreement-based fusion preserved precision and modestly improved coverage, and a reliability-weighted late fusion increased responsiveness without compromising the precision-first target.

To examine scalability, we trained the same Conv1D⇒BiLSTM on RAVDESS (English), observing clearer confusion-matrix diagonals and improved per-class scores relative to our Spanish set, which supports the view that performance is chiefly data-limited rather than architecture-limited. For deployment validation, we instrument session-level emotion distributions, ASR confidence, latency, and duration, and we plan controlled A/B tests that randomly enable the empathy module and compare session duration and task completion with robust confidence intervals to verify that perceived empathy increases engagement without harming task success.

5 Lessons Learned, Limitations, and Outlook

Defaulting to *Neutral* and triggering emotions only when calibrated confidence thresholds are met prevents mis-empathic responses, where even modest false positives harm user experience. With limited data, the audio branch drives quality, while text adds value as a veto or fallback when speech recognition is unreliable. Probability calibration—temperature scaling with class-specific thresholds—reduced over-triggering without latency costs. Architectural simplicity proved advantageous at our scale: the Conv1D⇒BiLSTM model was more robust than the AST variant, and results on RAVDESS suggest performance is constrained chiefly by data, not model class.

These lessons sit alongside clear constraints: a small, imbalanced, partly acted speech corpus limits generalisability. For deployment, we will use explicit opt-in consent within V·E·G·A, apply anonymisation or pseudonymisation with purpose limitation, and retrain in short cycles as diverse material accrues. Looking ahead, we will expand ethically sourced Spanish data, maintain precision-first, neutral-by-default policies, and progress towards calibrated tri-modal fusion, validating impact with analytics and controlled A/B tests. We plan to fuse emotional information from images, voice and text into a single, calibrated tri-modal model. This approach will improve real-time tracking of user affect and enable more timely, empathic adaptations throughout the interaction.

Acknowledgments. This research was funded by the Department of Big Data and Cognitive Systems at the Aragon Institute of Technology, under Retech Tourism-Spain Living Lab Agreement and by the Government of Aragon.

References

1. Aizawa, K., Nakamura, Y., Satoh, S.: Advances in Multimedia Information Processing – PCM 2004. Lecture Notes in Computer Science, Springer (2004), used as a standard reference for MFCC feature extraction overview
2. Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J., Lee, S., Narayanan, S.: Iemocap: Interactive emotional dyadic motion capture database. Language Resources and Evaluation **42**, 335–359 (2008)
3. Devlin, J., Chang, M., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proc. NAACL-HLT. pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423
4. Ekman, P.: Facial expressions of emotion: New findings, new questions. Psychological Science **3**, 34–38 (1992)
5. Imascono Art S.L.: Virtual Enterprise Generative Avatars (V·E·G·A) (2024), https://imascono.com/en/vega-product/
6. Interface EU: Emotional speech synthesis database. ELRA Catalogue (2014), https://www.islrn.org/resources/477-238-467-792-9/
7. Jones, C.R., Bergen, B.K.: People cannot distinguish gpt-4 from a human in a turing test (2024), https://arxiv.org/abs/2406.XXXXX, arXiv preprint
8. Livingstone, S.R., Russo, F.A.: 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**, 1–35 (2018)
9. Mehrabian, A., Wiener, M.: Decoding of inconsistent communications. Journal of Personality and Social Psychology **6**, 109–114 (1967)
10. Pan, R., García-Díaz, J.A., Rodríguez-García, M.Á., García-Sánchez, F., Valencia-García, R.: Overview of emospeech at iberlef 2024: Multimodal speech-text emotion recognition in spanish. Procesamiento del Lenguaje Natural **73**(0), 359–368 (2024), http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6623
11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 30, pp. 5998–6008. Curran Associates, Inc., Long Beach, CA, USA (2017), https://papers.nips.cc/paper/7181-attention-is-all-you-need
12. Yuan Gong, Y.A.C., Glass, J.: Ast: Audio spectrogram transformer. INTERSPEECH (Sep 2021)