HammerTeam at SpeechMATICS 2026: Exploring Early and Late Fusion Strategies for Multimodal Irony Detection in Spanish Spoken Dialogue

Darián Santiago Llanes-Guilarte[1], Vitali Herrera-Semenets[2], Lázaro Bustio-Martínez[3] and Gabriel Hernández-Sierra[1]

[1]*Advanced Technologies Application Center (CENATAV), Havana, Cuba*
[2]*On The Dime S.r.l., Ancona, Italy*
[3]*Universidad Iberoamericana, Mexico City, Mexico*

Abstract

This report describes the participation of HammerTeam in the SpeechMATICS challenge at IberLEF 2026, which focuses on identifying irony in colloquial Spanish spoken dialogues. To address this task, two distinct systems integrating pre-trained transformer embeddings were developed. In the first configuration, mean-pooled textual representations from XLM-RoBERTa-base and temporal statistical features from Wav2Vec2-Large-XLSR-53 were extracted and combined through early feature-level concatenation followed by a Multi-Layer Perceptron (MLP) classifier. In the second configuration, a late fusion pipeline using L2-normalized features and a weighted probability-based ensemble of calibrated Logistic Regression models was assessed. The evaluation shows that while both approaches capture relevant ironic cues, the early fusion MLP framework achieved a higher macro F_1-score of 0.66827, securing 3rd place on the official challenge leaderboard.

Keywords

irony detection, multimodal classification, speech processing, XLM-RoBERTa, Wav2Vec2, early fusion, late fusion, Spanish spoken dialogue

1. Introduction

Irony detection is a highly challenging problem in natural language processing and speech signal processing, as the true communicative intent of an utterance frequently contradicts its literal meaning. While lexical markers and textual context provide initial indicators of irony, speakers often rely heavily on prosodic variations, intonation, and acoustic nuances to signal ironic intent. This makes spoken language a particularly complex domain for automated detection, requiring systems that can effectively process both lexical and acoustic modalities.

The SpeechMATICS[1] shared task at IberLEF 2026 [2] addresses this challenge using the MESironic dataset [3], the first multimodal corpus specifically dedicated to irony detection in Peninsular Spanish. Drawn from podcasts, livestreams, and stand-up comedy, the dataset represents natural, non-formal spoken Spanish. The challenge is formulated as a binary classification task (ironic vs. non-ironic) split into two complementary tracks:

- **Subtask 1: Text-only Irony Classification**, which prohibits acoustic information and serves to analyze the boundaries of text-based irony detection.
- **Subtask 2: Multimodal Irony Classification**, which permits the combined usage of audio and text, encouraging systems that leverage prosodic features to improve detection performance.

This work explores two straightforward fusion methodologies to combine textual and acoustic modalities. Rather than fine-tuning massive networks end-to-end, a resource-efficient paradigm based on extracting rich, frozen representations from pre-trained foundation models was adopted. For text, a

IberLEF 2026, September 2026, León, Spain
darian.llanes@cenatav.co.cu (D. S. Llanes-Guilarte); vitali.herrera@onthedime.it (V. Herrera-Semenets); lazaro.bustio@ibero.mx (L. Bustio-Martínez); gsierra@cenatav.co.cu (G. Hernández-Sierra)

multilingual masked language model was used, and for audio, a robust self-supervised speech processor was leveraged. An early fusion approach optimized via cross-validated grid search over a Multi-Layer Perceptron was then compared against a late fusion framework using weighted posterior probabilities of linear classifiers. The primary early fusion submission yielded a macro F_1-score of 0.66827 on the official test set, placing HammerTeam in 3rd place.

The remainder of this report is organized as follows: Section 2 describes the dataset and the feature extraction pipeline; Section 3 outlines the architecture of the early and late fusion systems; Section 4 analyzes the experimental results; and Section 5 concludes with brief remarks on future directions.

2. Methodology and Feature Extraction

To build a computationally efficient yet robust classification pipeline, end-to-end fine-tuning of large pre-trained models was avoided. Instead, deep, high-dimensional representations were extracted from frozen state-of-the-art transformer checkpoints, and lightweight downstream classifiers were trained. This section describes the feature extraction strategies for both the textual and acoustic modalities.

2.1. Textual Feature Extraction (XLM-RoBERTa)

For the textual transcriptions, the pre-trained `xlm-roberta-base` model [4] was leveraged, as it provides strong multilingual representations particularly suited for Spanish colloquial vocabulary.

Rather than relying on the default class token (`[CLS]`), a masked mean-pooling strategy was implemented over the sequence length to capture the semantic information distributed across the entire sentence. Given the output hidden states $H_T \in \mathbb{R}^{B \times L \times D_T}$ (where B is the batch size, L is the padded sequence length restricted to a maximum of 128 tokens, and $D_T = 768$ is the hidden dimension) and the binary attention mask $M \in \{0, 1\}^{B \times L}$, the pooled text representation $h_T \in \mathbb{R}^{768}$ is computed as:

$$h_T = \frac{\sum_{i=1}^{L} H_{T,i} \cdot M_i}{\max\left(\sum_{i=1}^{L} M_i,\, \epsilon\right)} \tag{1}$$

where $\epsilon = 10^{-9}$ is a small constant to prevent division by zero. This pooling mechanism ensures that padding tokens do not skew the semantic representation of shorter transcriptions.

2.2. Acoustic Feature Extraction (Wav2Vec2-XLSR)

To extract prosodic and acoustic cues from the raw spoken audio, the multilingual `wav2vec2-large-xlsr-53` model [5] was utilized. All audio signals are standardized to a sampling rate of 16,000 Hz. Waveform loading, resampling, and standardization are handled using the Librosa [6] and TorchAudio [7] processing toolkits. The model processes the audio waveforms to yield frame-level hidden representations $H_A \in \mathbb{R}^{B \times T \times D_A}$, where T is the temporal sequence length and $D_A = 1024$.

Because irony is highly dependent on transient speech dynamics such as intonation shifts, elongation of vowels, or changes in speaking rate, a simple temporal mean is often insufficient. To capture both the average acoustic state and its temporal variability, two pooling statistics are computed across the time dimension: the temporal mean (μ_A) and the temporal standard deviation (σ_A):

$$\mu_A = \frac{1}{T} \sum_{t=1}^{T} H_{A,t} \tag{2}$$

$$\sigma_A = \sqrt{\frac{1}{T} \sum_{t=1}^{T} (H_{A,t} - \mu_A)^2} \tag{3}$$

These vectors are subsequently concatenated to form a composite 2048-dimensional acoustic representation $h_A \in \mathbb{R}^{2048}$:

$$h_A = \left[\mu_A \parallel \sigma_A\right] \tag{4}$$

2.2.1. Temporal Truncation Strategy (System 2)

In spoken dialogue, podcasts, and stand-up comedy, the markers of irony or the "punchline" frequently cluster towards the end of an utterance. To investigate whether focusing on the concluding portion of the audio improves classification performance, an end-truncation strategy was implemented for the second system configuration.

For any audio file exceeding a maximum duration of $t_{\text{max}} = 15$ seconds, only the final $t_{\text{max}} \times 16{,}000 = 240{,}000$ samples are retained. Utterances shorter than this limit are kept in their entirety. Features are then extracted from this truncated waveform using the same temporal pooling configuration described above.

2.3. Embedding Preprocessing and Scaling

To prepare the extracted embeddings for downstream linear and non-linear classifiers, two distinct normalization methodologies are compared:

1. **Standardization (System 1):** Standard normal scaling (zero mean, unit variance) is applied independently to the text embeddings and the acoustic embeddings using parameters fitted solely on the training split:

$$x_{\text{scaled}} = \frac{x - \mu_{\text{train}}}{\sigma_{\text{train}}} \tag{5}$$

2. **L2 Normalization (System 2):** To preserve directional relationships and mitigate scale discrepancies between different features, L2 normalization is applied to project the embeddings onto a unit hypersphere:

$$x_{\text{norm}} = \frac{x}{\|x\|_2} \tag{6}$$

   This normalization is often highly compatible with cosine-similarity based relationships inherent in pre-trained transformer embeddings.

3. Classification and Fusion Architectures

Two distinct architectural configurations were designed to address the challenges of Subtasks 1 and 2. The first system utilizes feature-level concatenation (Early Fusion) paired with a non-linear neural network, while the second system relies on decision-level probability blending (Late Fusion) of linear models. All linear models, scaling pipelines, and neural structures evaluated in this study were implemented using the Scikit-learn framework [8].

3.1. Subtask 1: Text-Only Classifiers

For the text-only classification subtask, linear models designed to handle high-dimensional embedding spaces robustly and efficiently were employed.

1. **Linear Support Vector Classifier (System 1):** In the first configuration, a Linear Support Vector Classifier (LinearSVC) is trained on the standardized text features. To address potential class imbalances, balanced class weighting is applied, which penalizes mistakes on the minority class proportionally to its under-representation. The objective is optimized using a grid search over

the regularization parameter $C \in \{0.001, 0.01, 0.1, 1, 10\}$. Through 5-fold cross-validation, the optimal parameter was determined to be $C = 0.001$.

2. **Logistic Regression (System 2):** In the second configuration, a Logistic Regression classifier trained on L2-normalized text features is utilized. This model offers the advantage of producing calibrated class posterior probabilities, which are essential for the late fusion pipeline. The inverse regularization strength $C \in \{0.1, 1, 10, 100\}$ is optimized using 5-fold cross-validation, selecting $C = 100$ as the best-performing parameter.

3.2. Subtask 2: Multimodal Fusion Schemes

To integrate the textual and acoustic modalities, two contrasting fusion paradigms are evaluated.

3.2.1. Early Fusion via Multi-Layer Perceptron (System 1)

The primary multimodal system uses an early fusion strategy, wherein normalized unimodal features are concatenated prior to classification. This approach allows the classifier to learn complex cross-modal interactions at the representation level.

Let $x_T \in \mathbb{R}^{768}$ and $x_A \in \mathbb{R}^{2048}$ denote the standardized text and audio embeddings, respectively. The joint multimodal representation $x_{\text{joint}} \in \mathbb{R}^{2816}$ is constructed via concatenation:

$$x_{\text{joint}} = \left[x_T \parallel x_A\right] \tag{7}$$

To capture non-linear relationships between prosodic variations and literal transcriptions, x_{joint} is passed into a Multi-Layer Perceptron (MLP). The network is trained with early stopping based on validation loss to prevent overfitting. A grid search is performed over key hyperparameters:

- Hidden layer architectures: $\{(256,), (512, 128)\}$
- L2 regularization penalty (alpha): $\{0.0001, 0.001\}$
- Initial learning rate: $\{0.001, 0.0001\}$

Using 3-fold cross-validation, the best architecture was found to be a single hidden layer of 256 units, an alpha of 0.001, and an initial learning rate of 0.0001.

3.2.2. Late Fusion via Calibrated Probability Blending (System 2)

As an alternative to early concatenation, a late fusion paradigm is explored in the second system. This strategy assumes that the textual and acoustic modalities can be modeled independently, reducing the risk of overfitting to high-dimensional concatenated spaces.

The late fusion workflow is structured as follows:

1. An independent Logistic Regression classifier is trained on the L2-normalized audio features x_A, searching over $C \in \{0.1, 1, 10, 100\}$. The optimal parameter found was $C = 100$.
2. The calibrated posterior probability vectors are extracted from both the text-only model ($P_T \in \mathbb{R}^2$) and the audio-only model ($P_A \in \mathbb{R}^2$).
3. Decision-level fusion is performed by computing a weighted linear combination of these probabilities:

$$P_{\text{final}} = w_T \cdot P_T + w_A \cdot P_A \tag{8}$$

   where w_T and w_A represent the modal weights, subject to $w_T + w_A = 1$. The fusion weights were selected empirically from internal validation experiments, with a higher weight assigned to the textual classifier because it provided a more stable unimodal baseline. The final configuration used $w_T = 0.65$ and $w_A = 0.35$.
4. The final predicted class $\hat{y}$ is determined via argmax:

$$\hat{y} = \arg\max_{c \in \{0,1\}} P_{\text{final},\, c} \tag{9}$$

4. Experimental Evaluation and Discussion

This section presents the evaluation framework, reports the official results on the IberLEF 2026 SpeechMATICS test set, and analyzes the performance of the proposed configurations.

4.1. Dataset and Evaluation Setup

The systems were evaluated on the MESironic dataset, which is split into a training set of 5,100 samples and a test set of 903 samples. The dataset maintains a highly balanced label distribution (50.72% ironic and 49.28% non-ironic).

The official evaluation metric is the macro-averaged F_1-score (F_1-macro) for binary classification, which weights both classes equally. Additionally, Macro Precision and Macro Recall are reported to provide a comprehensive understanding of system behavior.

4.2. Experimental Results

The performance of our proposed systems is summarized in Table 1. System 1 represents our primary submission (ID 657286) using standardized features, LinearSVC for Subtask 1, and the Early Fusion MLP for Subtask 2. System 2 represents our alternative late-fusion submission (ID 657292) using L2-normalized features, Logistic Regression for Subtask 1, and weighted probability blending for Subtask 2. Note that the official challenge platform only reported the macro F_1-score for secondary submissions; consequently, the detailed Macro Precision and Macro Recall breakdowns for System 2 were not available.

Table 1
Official evaluation results on the SpeechMATICS test set. Bold values indicate our best performance for each subtask. Due to challenge platform constraints, Macro Precision and Recall metrics are omitted for the secondary submission (System 2).

Configuration	Subtask	Macro Precision	Macro Recall	Macro F_1-score
System 1	Subtask 1 (Text-only)	0.60984	0.60994	**0.60906**
(Early Fusion / MLP)	Subtask 2 (Multimodal)	0.66855	0.66855	**0.66827**
System 2	Subtask 1 (Text-only)	—	—	0.60155
(Late Fusion / LogReg)	Subtask 2 (Multimodal)	—	—	0.62237

On the official challenge leaderboard, System 1 achieved a final score of **0.66827**, securing **3rd place** overall.

4.3. Discussion

Several key insights can be drawn from the experimental results:

- **Impact of Multimodal Integration:** For both configurations, the integration of acoustic features significantly improved classification performance compared to the text-only baselines. For System 1, the addition of Wav2Vec2 representations yielded an absolute increase of approximately 5.92% in F_1-macro (rising from 0.60906 to 0.66827). This improvement confirms that lexical cues alone are insufficient to resolve ambiguous irony, and that prosodic features captured by Wav2Vec2's temporal statistics (μ_A and σ_A) provide critical contextual grounding.
- **Early vs. Late Fusion:** The feature-level concatenation paired with a non-linear MLP (System 1) substantially outperformed the decision-level probability blending (System 2), which scored 0.62237. A plausible explanation is that early fusion allows the hidden layers of the MLP to learn cross-modal correlations (e.g., associating specific sarcastic lexical terms with low-energy or monotonous acoustic frames). In contrast, the late fusion model treats the modalities as strictly independent, which restricts its ability to capture these joint interactions.

- **Influence of the Truncation Strategy:** System 2 incorporated a temporal truncation strategy designed to focus exclusively on the final 15 seconds of each audio clip, assuming that irony markers would cluster near the punchline. However, this strategy may have inadvertently discarded early-onset acoustic markers, such as a shift in the speaker's baseline pitch or mock-politeness cues established early in the dialogue.
- **Feature Normalization:** While L2 normalization is theoretically robust for cosine-similarity in embedding spaces, the combination of standardization (StandardScaler) and a standard neural network classifier in System 1 proved more effective at preserving the statistical variance needed to distinguish subtle stylistic variations in speech.

5. Conclusion and Future Work

This work presented the participation of HammerTeam in the SpeechMATICS challenge at IberLEF 2026. Two distinct approaches for classifying ironic spoken dialogue in colloquial Spanish were explored. The experiments demonstrate that integrating deep acoustic features extracted from Wav2Vec2-Large-XLSR-53 with textual embeddings from XLM-RoBERTa-base yields a substantial performance improvement over text-only baselines. This confirms the critical role of prosodic cues in disambiguating ironic statements. Furthermore, the results suggest that an early fusion scheme utilizing a non-linear Multi-Layer Perceptron (MLP) is more effective at capturing cross-modal interactions than a rigid, probability-based late fusion approach. The primary system achieved a macro F_1-score of 0.66827, placing 3rd on the official leaderboard.

While these results are encouraging, the current pipeline has several limitations to be addressed in future work:

- **Cross-Modal Attention:** Instead of simple feature-level concatenation, cross-attention mechanisms (such as multimodal transformers) should be investigated to dynamically align and weigh the textual and acoustic sequences.
- **End-to-End Fine-Tuning:** The current approach relied entirely on frozen pre-trained embeddings to remain computationally efficient. Fine-tuning the outer transformer layers of both XLM-RoBERTa and Wav2Vec2 on the MESironic dataset could help the models learn task-specific features.
- **Spanish-Specific Acoustic Models:** Exploring self-supervised models pre-trained primarily on Spanish speech corpora (such as WavLM variants or Whisper encoder representations) may provide more fine-grained acoustic and prosodic details tailored to the dialects present in the dataset.
- **Robust Temporal Pooling:** Instead of using global mean and standard deviation statistics across the entire audio file, recurrent layers (e.g., GRUs) or temporal convolutional networks should be explored to model the sequence of prosodic changes over time.

Acknowledgments

The authors would like to thank the organizers of the SpeechMATICS 2026 shared task and IberLEF for coordinating this competition and providing the MESironic corpus.

Declaration on Generative AI

During the preparation of this work, the authors used Google AI Studio (models `gemini-3.5-flash` and `gemini-3.1-pro`) in order to: improve English phrasing and stylistic clarity; assist in structuring the LaTeX document; and support software implementation and script debugging. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] V. Ahuir, A. Barceló-Milkova, A. Casamayor-Segarra, M. J. Castro-Bleda, L. F. Hurtado, Overview of SpeechMATICS at IberLEF 2026: Speech Multimodal Audio-Text Irony Classification in Spanish, Procesamiento del Lenguaje Natural 77 (2026). To appear.

[2] A. Bonet-Jover, J. Á. González-Barba, L. Chiruzzo, Overview of IberLEF 2026: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS.org, 2026, p. to appear.

[3] V. Ahuir, A. J. Barceló-Milkova, A. Casamayor-Segarra, V. González Verdegais, L.-F. Hurtado, C. Periñán-Pascual, M. J. Castro-Bleda, MESironic: A Multimodal Dataset and Benchmark for Irony Detection in Spontaneous Spoken Spanish, Procesamiento del Lenguaje Natural 77 (2026). To appear.

[4] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 8440–8451.

[5] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, M. Auli, Unsupervised cross-lingual representation learning for speech, in: Proceedings of Interspeech 2021, ISCA, 2021, pp. 2426–2430.

[6] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, O. Nieto, librosa: Audio and music signal analysis in python, in: Proceedings of the 14th Python in Science Conference, 2015, pp. 18–25. doi:10.25080/majora-7b98e3ed-003.

[7] Y.-Y. Yang, M. Hira, Z. Ni, A. Chourdia, A. Astafurov, C. Chen, C.-F. Yeh, C. Puhrsch, D. Pollack, D. Genzel, D. Greenberg, E. Z. Yang, J. Lian, J. Mahadeokar, J. Hwang, J. Chen, P. Goldsborough, P. Roy, S. Narenthiran, S. Watanabe, S. Chintala, V. Quenneville-Bélair, Y. Shi, Torchaudio: Building blocks for audio and speech processing, arXiv preprint arXiv:2110.15018 (2021).

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, Journal of Machine Learning Research 12 (2011) 2825–2830.