# Ferrara at SatiSPeech-IberLEF 2025: Leveraging BETO and HuBERT for Multimodal Speech-Text Satire Recognition

Marco Bortolotti[1]

[1]University of Ferrara, Italy

**Abstract**

Satire is a complex and subtle form of communication that blends humor, irony, and criticism to address social, political, or cultural issues. Its interpretation often depends on nuanced linguistic and prosodic cues, making satire particularly difficult to detect—especially in multimodal settings that involve both textual and acoustic signals. This paper presents a system developed for the SatiSPeech@IberLEF 2025 shared task, which focuses on the binary classification of Spanish content as satirical or non-satirical using multimodal data. The proposed approach explores the interplay of linguistic patterns, vocal intonation, and rhythm to identify features most indicative of satire. Key challenges include the scarcity of rich multimodal satire datasets and the complexity of designing robust fusion strategies for heterogeneous modalities. By leveraging recent advances in deep learning and multimodal integration, the work aims to contribute to the development of more accurate and culturally aware satire detection systems.

**Keywords**

Satire Detection, Multimodal Classification, NLP, Speech Processing, BETO, HuBERT, Multi-Head Attention, Text-Audio Fusion, Spanish



## 1. Introduction

Satire is a sophisticated and multifaceted form of expression that employs irony, sarcasm, and exaggeration to critique social, political, or cultural phenomena. Unlike straightforward humor, satire often relies on implicit cues and contextual knowledge, making it inherently difficult to detect and interpret, even for human readers. In computational contexts, this complexity increases further, particularly when dealing with content that spans multiple modalities.

Recent advances in multimodal learning [1] [2] have opened new avenues for satire detection by enabling the fusion of textual and acoustic information. Prosodic features such as intonation, rhythm, and speech rate can provide essential signals for detecting the tone and intent behind a message. However, capturing the interplay between linguistic structure and vocal delivery remains a challenging task, especially in languages like Spanish, where cultural and regional nuances play a critical role in humorous expression.

The SatiSPeech@IberLEF 2025 shared task [3], held at IberLEF 2025 [4], addresses this research gap by promoting the development of systems capable of identifying satire in Spanish through a multimodal approach. The task consists of a binary classification challenge in which participants are required to determine whether a given audio-text pair is satirical or not. This setting enables the exploration of novel fusion techniques and the evaluation of various deep learning architectures suited for handling both sequential and acoustic inputs.

To address this challenge, the proposed system leverages two state-of-the-art pretrained models: BETO, a BERT-based language model trained on Spanish texts, for the textual modality, and HuBERT, a self-supervised speech representation model, for the audio modality. These representations are subsequently integrated through a late fusion architecture designed to combine complementary cues from both modalities.

*IberLEF 2025, September 2025, Zaragoza, Spain*
marco03.bortolotti@edu.unife.it (M. Bortolotti)
https://www.unife.it (M. Bortolotti)
0009-0002-8188-919X (M. Bortolotti)

## 2. Dataset

The dataset for this task was compiled to address the challenge of detecting satire in a multimodal context, combining text and audio. The data were obtained from a wide range of YouTube channels, including satirical programs such as El Intermedio, Zapeando, Homo-Zapping, and El Mundo Today, as well as non-satirical news programs such as Antena 3 Noticias, El Mundo, and BBC News. This ensures a broad representation of the varieties of Spanish spoken in different regions.

The compilation process involved extracting videos from these channels and segmenting them into manageable audio units using a diarization tool [5]. Segments longer than 25 seconds were discarded to maintain a consistent length. The audio segments were transcribed using Whisper [6] to ensure high-quality textual representations.

A semi-supervised approach was used to annotate the segments as either satirical or non-satirical, combining manual annotation by three experts and automatic classification techniques to increase efficiency and reliability. Next, a manual annotation process was conducted by the organizers to ensure high-quality labels for the dataset. It is worth noting that the dataset includes content from diverse Spanish-speaking regions, ensuring linguistic and cultural diversity while minimizing regional bias.

The final dataset consists of approximately 25 hours of annotated audio segments and their corresponding transcriptions. For the purposes of this competition, around 5,000–6,000 audio-text pairs will be selected and divided into training and test sets with an 80%-20% split.

## 3. System Description

The system developed for the SatiSPeech@IberLEF 2025 shared task addresses two distinct subtasks: the first (Task 1) focuses on satire detection using only textual modality, while the second (Task 2) employs a multimodal approach that combines both textual and acoustic information. Each task is tackled using a specific preprocessing pipeline and tailored classification models.

For the textual modality, the BETO model, a Spanish pre-trained BERT variant, is used to extract deep contextual embeddings. For the audio modality, HuBERT is employed to capture prosodic and phonetic representations. In the multimodal task, these two modalities are fused to enhance classification performance by integrating information from both text and audio.

## 4. Text Modality (Task 1)

In this task, the aim is to classify satirical content based solely on textual transcriptions. The system relies on the BETO model, a Spanish BERT-based model, which is fine-tuned using the SentenceTransformers framework.

### 4.1. BETO: Spanish BERT-based Model

BETO [1] is a transformer-based language model pre-trained on large Spanish corpora. It follows the same architecture as BERT-Base [7] (12 layers, 768 hidden units, 12 attention heads, and 110 million parameters), but it is trained exclusively on Spanish data, such as Wikipedia and news articles.

By focusing on the Spanish language, BETO captures language-specific nuances, making it more suitable than multilingual models for tasks such as satire detection, where understanding idiomatic expressions, tone, and cultural references is crucial.

**Tokenization Process.** The BETO model uses the WordPiece tokenization algorithm, which splits the input text into subword units, handling out-of-vocabulary words and morphological variations effectively. During preprocessing, the input text is tokenized, with special tokens like [CLS] and [SEP] added at the beginning and end of the sequence. The tokenized input is then transformed into token IDs, attention masks, and segment embeddings, which are used in the downstream fine-tuning process.

### 4.2. Model Architecture

The core of the system is based on the pre-trained BETO model (dccuchile/bert-base-spanish-wwm-uncased). For this task, the BETO model was directly utilized for classification purposes.

The model was integrated into the SentenceTransformers framework, which efficiently converts text into sentence embeddings. The architecture consists of three components: the transformer encoder (BETO), a pooling layer that computes sentence embeddings, and a dense layer with a Tanh activation function to project the embeddings into a 768-dimensional space.

Tokenization and preprocessing were performed using the tokenizer associated with the BETO model. The dataset was tokenized and prepared without manual text preprocessing, as BETO's tokenizer handles typical cleaning operations such as lowercasing, punctuation removal, and tokenization.

The model was trained for three epochs on the training dataset. This training process enabled the model to adjust its weights to improve its ability to classify satirical content in Spanish. Despite the simplicity of the configuration, the pre-trained BETO model yielded strong performance, highlighting the effectiveness of transfer learning.

This architecture leverages the advantages of a pre-trained language model, providing a robust and data-efficient solution for the classification task.

## 5. Multimodal Task (Task 2)

The goal of the second task is to identify satire by combining both textual and vocal features in a multimodal approach, leveraging the strengths of both data types. First, the audio features are extracted and used to train an independent audio classifier. Then, a fusion strategy is applied to integrate predictions from both modalities.

### 5.1. Audio Modality

The objective of the audio subtask is to perform satire classification using only vocal characteristics, assuming that acoustic patterns such as tone, rhythm, or prosody may signal satirical intent.

### 5.2. Feature Extraction with HuBERT

To extract high-level speech features, we employed the pre-trained HuBERT model [2] HuBERT (Hidden-Unit BERT) is a self-supervised model that learns acoustic representations by predicting masked audio segments based on unsupervised cluster assignments.

Raw audio files were resampled to 16 kHz using `librosa`. We used the `Wav2Vec2Processor` and the pre-trained HuBERT model from Hugging Face to compute hidden representations. The final feature vector for each audio file was obtained by averaging the last hidden states across the time dimension, resulting in a fixed-length 1024-dimensional embedding that captures prosodic, rhythmic, and timbral features.

### 5.3. Audio Classification Architecture

The extracted embeddings were subsequently processed by a feed-forward neural network designed for classification. The architecture begins with an input layer that receives the 1024-dimensional embeddings generated by the HuBERT model.

This is followed by two hidden layers. The first hidden layer consists of a fully connected layer with 1024 units, followed by Batch Normalization, a ReLU activation function, and a Dropout layer with a dropout rate of 0.4. The second hidden layer applies a similar structure, with a fully connected layer reduced to 256 units, again followed by Batch Normalization, ReLU activation, and Dropout with the same rate.

Finally, the network concludes with an output layer consisting of a fully connected layer, making it suitable for binary classification.

**Training Details.** The model was trained using `CrossEntropyLoss` and optimized with the `Adam` optimizer (learning rate = 0.001). A `StepLR` scheduler decreased the learning rate by a factor of 0.1 every 10 epochs. Training was performed for 40 epochs.

# 6. Multimodal Fusion Strategy

To exploit both textual and acoustic information, we adopted a late fusion strategy, which combines the class probability distributions generated independently by the text and audio classifiers.

## 6.1. Fusion Mechanism

Each classifier outputs a probability vector over the target classes. These vectors are combined via element-wise average (arithmetic mean), giving equal weight to each modality. The final prediction corresponds to the class with the highest combined score.

## 6.2. System Overview

- **Text Classifier**: A transformer-based model (BETO) generates semantic representations from text transcriptions and outputs class probabilities.
- **Audio Classifier**: A HuBERT-based pipeline generates vocal embeddings, which are classified using a feed-forward network.
- **Fusion Step**: Class probabilities from both classifiers are averaged to produce the final decision.

# 7. Final Results for Competition

The system was initially evaluated during the SatiSPeech@IberLEF 2025 competition. The model achieved notable performance on the test set, with the following results:

- F1 Score (Textual Task): 0.832
- F1 Score (Multimodal Task): 0.837

## 7.1. Results Overview

In this subsection, we present the performance rankings for each subtask of the competition. The results are evaluated using the MACRO F1-Score, which reflects the average performance across all classes.

### 7.1.1. Task 1: Textual Task Ranking Results

Below is the ranking of teams for the Textual Task, based on the MACRO F1-Score:

| Ranking | User Team | Task 2: Multimodal F1-Score |
|---|---|---|
| 1 | mcastro UPV-ELiRF | 85.63760 |
| 2 | ITST ITST | 84.54580 |
| 3 | edu_valero UMU-Ev | 84.45500 |
| 4 | nguyenminhbao5032 | 83.27390 |
| 5 | MarcoBortolotti Ferrara | 83.21010 |
| 6 | AnGladun UKR | 83.20180 |
| 7 | ngocan0987 | 82.04130 |
| 8 | klagos1875 UAE | 81.63100 |
| 9 | angelalm LACELL | 81.46950 |
| 10 | cespinr EcuPLN | 79.47870 |
| 11 | deniscedeno UTP | 77.93640 |

#### 7.1.2. Task 2: Multimodal Task Ranking Results

Below is the ranking of teams for the Multimodal Task, based on the MACRO F1-Score:

| Ranking | User Team | Task 1: Textual F1-Score |
|---|---|---|
| 1 | edu_valero UMU-Ev | 88.34030 |
| 2 | mcastro UPV-ELiRF | 86.44420 |
| 3 | MarcoBortolotti Ferrara | 83.70410 |
| 4 | nguyenminhbao5032 | 83.27390 |
| 5 | ITST ITST | 83.27140 |
| 6 | ngocan0987 | 82.77640 |
| 7 | klagos1875 UAE | 81.49610 |
| 8 | angelalm LACELL | 81.46950 |
| 9 | AnGladun UKR | 80.13050 |
| 10 | cespinr EcuPLN | 79.47870 |
| 11 | deniscedeno UTP | 76.47580 |

## 8. Results and Discussion

The results obtained in the SatiSPeech@IberLEF 2025 competition demonstrate the effectiveness of the developed model. The model achieved an F1-Score of 0.832 for the *Multimodal Task* and 0.837 for the *Textual Task*, highlighting its capability to effectively handle both text recognition and multimodal integration of text and audio. The model ranked third in the *Multimodal Task* and fifth in the *Textual Task*.

The competitive performance in the rankings, with a score of 83.7 for the *Multimodal Task* and 83.2 for the *Textual Task*, indicates a good balance between model complexity and accuracy, showing strong generalization capabilities on the challenging task of satire detection.

### 8.1. Discussion of the Results

The results obtained show that the model has been successfully applied to the task of satire detection in both textual and multimodal formats. The performance is competitive, ranking in the top positions, reflecting a well-optimized and balanced model. These results suggest that the approach used is capable of addressing complex problems involving both natural language processing and audio analysis, achieving a high level of generalization. It can be observed that the inclusion of the audio modality has had a positive impact on performance, as evidenced by a significant improvement in the ranking of a participant who moved from a macro F1 score of 84.4 in the textual task to 88.3 in the multimodal task. This indicates that audio plays a beneficial role in enhancing the model's ability to detect satirical content. Nevertheless, there is still room for improvement in both tasks, particularly in the audio component, where further refinements to the proposed models in this paper could lead to better classification accuracy.

## 9. Improvements After the Competition

Following the conclusion of the competition, several enhancements were introduced to improve overall model performance by refining both the textual and audio modalities. Modifications to the text pipeline included training multiple BETO models with varied splits and epochs, while the audio component was enhanced through different fusion strategies and attention-based mechanisms.

### 9.1. Improvements After the Competition - Task 1

After the competition, several strategies were applied to enhance the model's performance in the Textual Task. These improvements focused on both refining the model architecture and enhancing the

training dataset. The main approaches explored were the use of multiple BETO models with varying training-validation splits and data augmentation techniques to enrich the textual data.

The first approach involved training multiple models on different subsets of the data, each with varying numbers of epochs. This method aimed to increase the model's generalization capability by exposing it to different data splits and training settings. Additionally, the incorporation of soft labeling and max labeling strategies was evaluated to determine which method would yield better performance.

The second improvement focused on augmenting the textual dataset to increase its size and variability. By employing various techniques such as synonym replacement, back-translation, and random deletion/insertion, the training data was made more diverse, which was expected to help the model generalize better to unseen data.

### 9.1.1. Multiple BETO Models with Different Training-Validation Splits

Several BETO models were trained with different training-validation splits. For each configuration, the number of epochs was varied, and the F1 scores were calculated for each model. Below are the results for the models trained with 3, 4, 5, and 6 epochs.

| Number of Models | Epochs | F1 Score |
|---|---|---|
| 5 Models | 6 | 0.842 |
| 4 Models | 5 | 0.853 |
| 3 Models | 4 | 0.843 |
| 2 Models | 6 | 0.847 |
| 1 Model | 3 | 0.832 |
| | 4 | 0.839 |

**Table 1**
F1 Scores for Different Configurations of BETO Models and Epochs

From the table above, it can be observed that the highest F1 score was achieved using 4 models trained for 5 epochs, outperforming configurations with 3 and 6 epochs. Additionally, using multiple models led to an improvement in performance compared to using a single model, with the F1 score increasing as more models were included. Between soft labeling and max labeling, the soft labeling approach was found to slightly outperform the max labeling technique, which is why it was chosen for the experiments.

### 9.1.2. Data Augmentation for Textual Data

To further improve performance, data augmentation techniques were explored to increase the diversity of the training data. The previously identified best-performing configuration (four models trained for five epochs) served as the baseline for these experiments. The following augmentation techniques were applied:

- **Synonym Replacement**: Words were randomly replaced with their synonyms to increase vocabulary diversity.
- **Back-Translation**: Sentences were translated into another language (e.g., English) and then translated back to Spanish, creating paraphrased versions of the original text.
- **Random Deletion/Insertion**: Words were randomly removed or added to sentences to create variations in the textual structure and lexical content.

Despite the application of these augmentation techniques, a slight decrease in performance was observed on the official test set when compared to the original configuration. As a result, the data augmentation approach was not included in the final submission. While the performance differences were not large, the original training setup without augmentation was found to be more reliable for generalization, avoiding the risk of overfitting introduced by artificial noise in the data.

## 9.2. Improvements After the Competition for Task 2

After the competition, further improvements were made to the multimodal classification model by exploring two main strategies for combining the predictions from the audio and text classifiers.

The first approach involved a weighted averaging strategy, where various combinations of weights were applied to the predictions of the two modalities. The objective was to identify the most effective weight distribution capable of enhancing classification performance.

The second approach consisted in the development of a neural network based on a multimodal attention mechanism. In this model, the class probabilities generated independently by the audio and text classifiers are used as input and fused through an attention-based mechanism to obtain the final prediction. The use of attention allows the network to selectively focus on the most informative aspects of each modality, thereby improving the accuracy of the resulting label prediction.

These enhancements were aimed at boosting overall model performance by fully exploiting the complementary nature of audio and text information.

## 9.3. Weighted Averaging Results

In this section, the results of the weighted averaging approach for combining the predictions from the audio and text classifiers are presented. The goal of this approach was to determine the optimal balance between the two modalities by experimenting with different weight distributions for the class probabilities generated by the models.

For the weighted averaging, a variety of combinations of weights for the text and audio classifiers were used, and the corresponding performance was evaluated using the F1 Score. Below are the results for several weight combinations:

| $w_{\text{text}}$ | $w_{\text{audio}}$ | F1 Score |
|---|---|---|
| 0.50 | 0.50 | 0.857 |
| 0.90 | 0.10 | 0.853 |
| 0.10 | 0.90 | 0.826 |
| 0.30 | 0.70 | 0.839 |
| 0.45 | 0.55 | 0.848 |
| 0.55 | 0.45 | 0.857 |
| 0.60 | 0.40 | 0.858 |
| 0.70 | 0.30 | 0.856 |
| 0.65 | 0.35 | 0.858 |

**Table 2**
F1 scores for various weighting combinations.

From these results, it is clear that a balanced combination of the text and audio modalities generally leads to better performance compared to using extreme values for either modality. The best performing configuration was when the text modality was weighted slightly more heavily ($w_{\text{text}} = 0.65$, $w_{\text{audio}} = 0.35$), achieving an F1 score of 0.858, outperforming the text-only baseline (0.853) and demonstrating that a weighted fusion strategy can enhance classification accuracy by effectively leveraging the complementary strengths of both modalities.

## 9.4. Multimodal Attention Network

To further improve the multimodal classification model, a **Multimodal Attention Neural Network** [8] was employed, which combines the probability distributions obtained from the audio and text models. The approach aims to better capture the relationships between the two modalities by leveraging attention mechanisms.

#### 9.4.1. Data Preparation

The dataset was split into **training** and **validation** sets with an 80-20% ratio. The audio model previously described was retrained exclusively on the training portion (4800 examples) to ensure that the data used for training the subsequent Multi-Head Attention fusion mechanism remained unseen during this phase. This prevented any label leakage and ensured a fair evaluation when learning optimal fusion weights.

### 9.5. Multimodal Attention Network Model

In order to improve the multimodal classification performance, a **Multimodal Attention Network** that combines the probability distributions output by the audio and text models has been adopted. The core idea behind this approach is to leverage the attention mechanism to learn the optimal fusion strategy between the two modalities, enhancing the model's ability to capture and integrate complementary information from both audio and text.

The architecture is structured around several interconnected components. First, the model takes as input the probability distributions generated independently by the audio and text classifiers. These distributions, which express the models' confidence in the classification task, constitute the foundation for the multimodal fusion.

Subsequently, the input features are projected into three separate spaces—Queries, Keys, and Values—through fully connected layers. These projections enable the model to build rich internal representations and effectively guide the attention mechanism in focusing on the most relevant aspects of each modality. The dimensionality of these projections is governed by the number of attention heads (4) and the hidden dimension size.

The core attention mechanism is based on the scaled dot-product attention, which computes similarity scores between Queries and Keys. These scores are normalized using a softmax function, producing attention weights that modulate the relative importance assigned to each modality. The resulting weighted sum of the Values captures the most salient information from both audio and text sources.

To preserve the original input features while enabling the learning of useful interactions, the attention output is combined with the input through a residual connection and passed through a linear projection layer. This ensures that the learned patterns do not override essential input characteristics.

The fusion of modalities is then performed through a learnable weighted combination of the audio and text outputs, governed by two parameters, $\alpha$ and $\beta$, which determine their respective contributions to the final decision. This dynamic fusion mechanism allows the model to adaptively weigh the importance of each modality based on the context.

Finally, Layer Normalization is applied to promote training stability, and dropout is used with a probability of 0.1 on the attention weights to mitigate overfitting. This combination ensures that the model maintains robustness and generalizes well across different inputs.

### 9.6. Multimodal Attention Network Results

The resulting F1 score was **0.859**, slightly higher than the best weighted average score of 0.858.

This result suggests the potential effectiveness of the attention mechanism in combining probabilities from both modalities, which could lead to improved overall classification performance. The model's ability to learn optimal fusion strategies appears promising and warrants further investigation.

### 9.7. Post-Competition Results

After the competition, the multimodal system was further refined and improved by testing different techniques to optimize its performance.

For Task 1, the main improvements came from:

- **Increased Training Epochs for the BETO models:** The number of training epochs for the BETO models was extended, allowing them to converge better and improve performance.

- **Training Multiple BETO Models:** Multiple BETO models were trained on different validation and training splits, which helped increase the robustness and generalization capability of the system.

For Task 2, several fusion strategies were explored to improve performance. The following techniques were implemented:

- **Weighted Averaging:** A weighted averaging strategy was experimented with, where the weights of the predictions from the audio and text classifiers were adjusted. This helped find the optimal balance between the modalities.
- **Multimodal Attention Network:** A neural network using a multimodal attention mechanism was developed. The model takes class probabilities from both the audio and text classifiers and combines them through attention-based fusion. This attention mechanism allows the model to focus on the most relevant parts of each modality, improving prediction accuracy.

The main performance improvements were achieved by training four BETO models for five epochs each for Task 1. For Task 2, the implementation of the Multi-Head Attention strategy showed promising results and contributed to a small but measurable performance increase.

**Post-competition Results:**

- **Task 1 (Binary Satire Detection - Text):** F1-score of **0.853**
- **Task 2 (Multimodal Satire Detection):** F1-score of **0.859**

## 10. Conclusions and Future Improvements

The results obtained clearly indicate that the textual information, captured through the BETO language model, plays a predominant role in the classification task, offering the highest individual performance among the modalities. However, the inclusion of acoustic features extracted using HuBERT contributed to measurable improvements, reinforcing the hypothesis that audio and text provide complementary information for the detection of satirical content.

Despite the competitive results achieved, several avenues for future work remain open to further enhance the system. A first line of improvement concerns the audio classifier itself, which in the current approach is based on a relatively simple architecture. Employing more sophisticated neural architectures or fine-tuning pre-trained speech models could potentially lead to better acoustic representations and thus improved classification accuracy.

A second promising direction involves refining the multimodal attention mechanism. In the present work, attention is applied to the class probability distributions generated by the unimodal classifiers. Future research could explore the design of a more expressive Multi-Head Attention module that operates directly on the intermediate features extracted from both modalities. This would enable the model to capture richer cross-modal interactions and possibly uncover deeper patterns relevant to satire detection.

Moreover, the current attention-based fusion approach was trained using a limited set of approximately 1200 multimodal examples. Increasing the amount of training data would likely enhance the stability and robustness of the fusion model, especially in complex or ambiguous cases. Additionally, systematically tuning the hyperparameters of the attention mechanism—such as the number of heads and the size of the hidden layers—may yield configurations that better support the integration of multimodal information and further improve the overall performance.

In summary, this study shows that combining textual and acoustic cues is a viable strategy for satire detection. Further improvements in fusion strategies, acoustic modeling, and training data augmentation are expected to yield even better results in future research.

## Declaration on Generative AI

During the preparation of this work, the author(s) used DeepL in order to Grammar and spelling check.

## References


[1] J. Cañete, G. Chaperon, R. Fuentes, J. Pérez, B. Poblete, BETO: Spanish BERT pretrained model, https://github.com/dccuchile/beto, 2020. Accessed June 2025.

[2] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, A. Mohamed, Hubert: Self-supervised speech representation learning by masked prediction of hidden units, IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021) 3451–3463.

[3] R. Pan, J. A. García-Díaz, T. Bernal-Beltrán, F. García-Sánchez, R. Valencia-García, Overview of SatiSPeech at IberLEF 2025: Multimodal Audio-Text Satire Classification in Spanish, Procesamiento del Lenguaje Natural 75 (2025).

[4] J. Á. González-Barba, L. Chiruzzo, S. M. Jiménez-Zafra, Overview of IberLEF 2025: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org, 2025.

[5] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, M.-P. Gill, Pyannote. audio: neural building blocks for speaker diarization, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 7124–7128.

[6] A. Radford, et al., Whisper: Openai's speech recognition system, https://github.com/openai/whisper, 2023. Accessed June 2025.

[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2019, pp. 4171–4186.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.