DOI: 10.61782/fa.2025.0850

FORUM ACUSTICUM
EURONOISE 2025

TEMPORAL SEGMENTATION STRATEGY TO IMPROVE SPEECH-BASED ANGER DETECTION USING YAMNET

Fangfang Zhu-Zhou* **Roberto Gil-Pita** **Diana Tejera-Berengué**
Gonzalo Corral-García **Manuel Zurera-Rosa** **Manuel Utrilla-Manso**
Signal Theory and Communications Department, University of Alcalá,
Pza. San Diego s/n, Alcalá de Henares, 28805, Madrid, Spain.

ABSTRACT

Anger detection in speech is a critical task in affective computing, especially in real-world environments where acoustic conditions are often suboptimal. This paper presents a robust anger detection system based on transfer learning using YAMNet, a convolutional neural network pre-trained on large audio datasets. Two key parameters are analyzed: the number of layers retained from the pre-trained model and the number of temporal segments per utterance introduced to capture temporally distributed emotional cues.

The system is evaluated on the emoDB dataset under a leave-one-speaker-out cross-validation scheme. Audio samples are augmented with noise and reverberation to simulate realistic acoustic conditions. The results show that increasing the number of temporal segments per utterance leads to significant performance improvements, with configurations using four or more segments achieving up to 60% relative reduction in the probability of false positives and significantly higher F1 scores. Moderate network depths provide a good trade-off between complexity and accuracy.

These results highlight the effectiveness of segment-wise processing and architectural optimization in improving speech-based emotion recognition. The proposed system demonstrates high robustness and generalizability, making it suitable for use in practical human-machine interaction scenarios.

Keywords: anger detection, speech emotion recognition, transfer learning, temporal segmentation

**Corresponding author*: *fangfang.zhu@uah.es.*

1. INTRODUCTION

Automatic emotion recognition from speech has gained increasing interest due to its applications in human-computer interaction, mental health monitoring, and behavioral analysis [1, 2]. Among emotional states, anger is particularly relevant given its association with conflict escalation and high-risk decision-making. Reliable anger detection can enhance voice-based systems in contexts such as call centers, driver assistance, and therapy [3].

However, detecting anger remains challenging, especially in real-world conditions with background noise, reverberation, and speaker variability [4]. Traditional methods relying on handcrafted acoustic features often lack robustness [5]. Deep learning approaches have improved performance by learning discriminative representations directly from audio [6, 7], but they typically require large annotated datasets, which are scarce in emotion recognition.

Transfer learning has emerged as a solution to this limitation. In particular, YAMNet [8], a CNN pre-trained on the large-scale AudioSet dataset [9], has shown excellent results in general audio classification. Yet, its potential for fine-grained emotion recognition, such as anger detection, remains underexplored.

Another key challenge is the temporal distribution of emotional cues within an utterance. Fixed-length sliding windows, as used by default in YAMNet, may dilute or

miss emotionally salient regions. Segment-wise strategies have shown promise in capturing localized emotional information and improving robustness [10].

In this work, we present a robust anger detection system based on transfer learning with YAMNet. We analyze the impact of two design factors: the number of retained layers from YAMNet and the number of fixed-length temporal segments per utterance. A customized segmentation strategy is applied to handle variable-length inputs, and data augmentation with noise and reverberation from the MUSAN corpus [11] is used to enhance generalization.

The remainder of this paper is structured as follows: Section 2 details the dataset and transfer learning setup; Section 3 presents the system architecture and preprocessing strategy; Section 4 reports experimental results; and Section 5 concludes the study.

2. MATERIALS AND METHODS

This section provides a detailed description of the dataset and methodological framework used for the development and evaluation of the proposed anger detection system. First, we present the Berlin Emotional Speech Database (emoDB), which serves as the basis for training and validation. We then describe the data augmentation strategy used to simulate real-world acoustic conditions and improve model robustness. Finally, we outline the neural architecture employed, which builds on YAMNet through a transfer learning approach adapted for binary classification.

2.1 Database Overview

The Berlin Emotional Speech Database (emoDB) is a well-established and widely used resource in the field of speech emotion recognition [12]. It contains recordings of ten German native speakers (five male and five female) who acted out seven different emotions –anger, boredom, disgust, fear, happiness, sadness, and neutral– while uttering a predefined set of sentences. This resulted in a total of 528 utterances, with 126 corresponding to anger and the remaining 402 distributed among the other six emotional categories. In this study, we reframe the task as a binary classification problem, where anger is treated as the target class and all other emotions are grouped as non-anger. While this introduces a slight class imbalance, the dataset remains sufficiently balanced to support robust and reliable evaluation.

In addition to its emotional diversity, emoDB exhibits considerable variability in audio sample duration, which is an essential feature to consider in audio-based detection tasks. Figure 1 shows the distribution of audio durations in the dataset. The average duration of the recordings is 2.79 seconds, with a median of 2.62 seconds, a minimum of 1.23 seconds, and a maximum of 8.98 seconds. Most of the utterances are concentrated between 2 and 4 seconds, although a few extend to nearly 9 seconds.

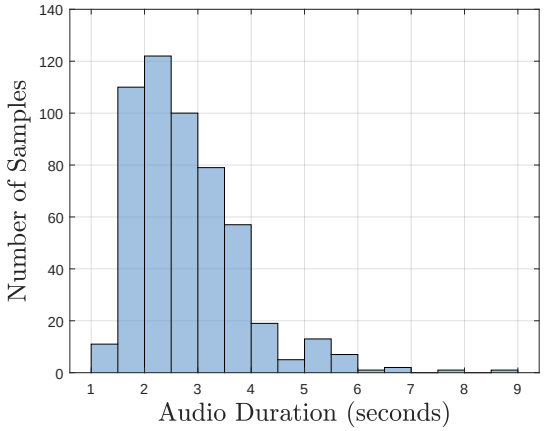

Figure 1. Histogram of audio durations in emoDB, showing the distribution across utterances.

This variability highlights the importance of considering utterance length when developing speech-based emotion recognition systems, as differences in duration can affect the performance and stability of feature extraction and classification processes.

To address the limited size of the emoDB dataset and to increase the robustness of the proposed system under real-world conditions, we applied a targeted data augmentation strategy inspired by the approach introduced in [10]. This process involved artificially altering the acoustic environment of the original recordings by adding both reverberation and background noise.

Reverberation was simulated by convolving each audio signal with a room impulse response (RIR) configured to produce a reverberation time of RT60 = 0.7 seconds, emulating realistic indoor spaces. Background noise was introduced at multiple signal-to-noise ratio (SNR) levels ranging from 3 dB to 39 dB in 3 dB increments, effectively generating multiple augmented versions of each utterance with varying degrees of acoustic degradation.

To incorporate realistic noise sources, we used the

MUSAN dataset [11], which contains a diverse collection of real-world audio, such as ambient sounds and music. From this dataset, non-speech audio samples were randomly selected without repetition and superimposed on the original recordings to simulate challenging and varied acoustic scenarios.

This augmentation approach significantly increases the variability and realism of the training data, promoting the development of a model capable of generalizing to a wide range of unpredictable acoustic environments.

2.2 Transfer Learning using YAMNet pre-trained CNN network

To develop an effective anger detection system in speech, we adopt a transfer learning strategy based on YAMNet [8], a deep convolutional neural network (CNN) pre-trained on AudioSet, a large dataset containing over two million human-labeled audio clips categorized into more than 500 classes [9]. YAMNet is built on the MobileNetV1 architecture, which is optimized for efficient classification of audio events, and its pre-training on a wide variety of sound classes makes it a powerful starting point for emotion recognition tasks.

In this work, we adopt YAMNet as a pre-trained backbone network that is truncated at different depths (N_{layers}) to analyze the impact of model complexity on the anger detection task. Instead of using the full model, only the first N_{layers} are retained, with four configurations evaluated in this study: 40, 52, 64, and 76 layers, to study how much of the pre-trained network is needed to capture relevant emotional patterns in speech.

The truncated YAMNet model is extended with a global average pooling layer and a fully connected layer with a single output neuron. This architecture produces a continuous output score between 0 and 1, indicating the probability that the input audio segment expresses anger. A regression layer is used during training to provide soft probabilistic outputs that are more informative for post-processing and threshold-based decision making.

Importantly, the entire model—including the retained part of YAMNet—is fine-tuned during training. This full-network optimization allows the system to adapt pre-learned audio features to the specific task of anger detection, leveraging both the general knowledge captured from large audio datasets and the specificity required for emotion recognition in speech.

This approach provides an efficient way to adapt a general-purpose audio network to a specific emotion recognition task, while facilitating a systematic analysis of how network depth influences performance in anger detection.

3. PROPOSED APPROACH

This section presents the overall framework designed for speech-based anger detection. The system is structured as a modular pipeline that transforms raw audio into a probabilistic anger score. We first describe the architecture and components of the model, including input preprocessing, segmentation, feature extraction, and training. We then provide a detailed explanation of the segmentation strategy, which plays a central role in adapting the system to utterances of different durations and enhancing its ability to capture localized emotional cues.

3.1 System Architecture

The proposed anger detection system is designed as a modular pipeline that processes raw audio recordings and outputs a continuous score representing the probability of anger. It consists of the following main components:

1. Audio Input and Normalization: Each input consists of a raw audio signal sampled at 16 kHz. The signals are normalized to zero mean and unit variance to reduce loudness variability and ensure consistency across utterances.

2. Temporal Segmentation: To account for the variability in audio duration found in the emoDB dataset, each utterance is divided into a fixed number of temporal segments (denoted as N_{segs}). These segments, approximately 1 second in duration, are extracted from specific positions along the utterance (e.g., beginning, middle, end), as described in detail in Section 3.2. This strategy ensures uniform input size and allows the model to focus on local emotional cues.

3. Spectrogram Generation and YAMNet Feature Extraction: Each segment is first converted into a log-mel spectrogram using YAMNet's built-in preprocessing pipeline, and then passed through a truncated version of the network that retains only the first N_{layers} of its architecture. Four configurations were evaluated in this study: 40, 52, 64, and 76 layers. This truncation allows analysis of the effect of model depth on emotion recognition performance.

4. Output Head and Score Estimation: The extracted features are passed through a global average pooling layer, followed by a fully connected layer with a single output neuron. A regression layer is used to produce a soft output score between 0 and 1 representing the likelihood of anger in the segment.

5. Score Aggregation: Since each utterance is divided into multiple segments, the final anger score is obtained by averaging the scores across all segments, providing a robust summary that accounts for temporal variation in emotional expression.

6. Training Strategy: The entire model, including the truncated YAMNet layers, is fine-tuned during training. This enables the system to adapt the pre-trained representations to the specific task of anger detection. A leave-one-speaker-out cross-validation scheme is used to assess generalization to unseen speakers. Each experiment is repeated ten times to ensure robust performance estimation.

This architecture provides both flexibility and interpretability, allowing systematic evaluation of the effects of network depth and segmentation granularity on anger detection performance.

3.2 Preprocessing and Segmentation Strategy

Handling variable-length audio is a key challenge in speech-based emotion recognition. The emoDB dataset contains utterances ranging from 1.23 to 8.98 seconds, which complicates processing with fixed-size neural networks such as YAMNet. To ensure uniform and robust feature extraction, we adopt a segmentation strategy that divides each utterance into a fixed number of 1-second segments, denoted by N_{segs}, which is treated as a tunable hyperparameter.

The baseline case (N_{segs} = 1) corresponds to YAMNet's default segmentation, which uses overlapping 0.975-second windows with a hop size of 0.48 seconds. In contrast, our custom strategy selects non-overlapping (or partially overlapping) 1-second segments from predefined positions, with automatic overlapping applied for shorter utterances to maintain coverage.

This approach standardizes input length and increases the system's robustness to temporal variability by allowing the model to focus on localized emotional cues. Segment-level scores are averaged to obtain the final prediction, integrating temporal diversity while maintaining consistent network input.

Figures 2 and 3 show the distribution of analysis windows for different values of N_{segs} using short and long utterances, respectively. The comparison highlights the contrast between YAMNet's dense, overlapping windows and our fixed-position, 1-second segments. While short utterances introduce overlap to ensure coverage, longer utterances distribute segments to capture temporal diversity. This design reduces redundancy and improves the model's ability to detect localized emotional cues across utterance lengths.

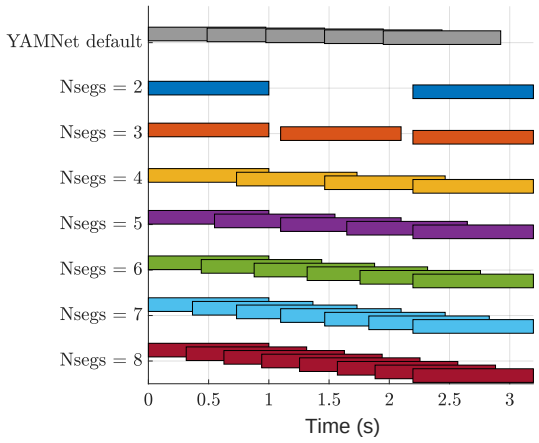

Figure 2. Illustration of segmentation positions for different values of N_{segs} in a short utterance.

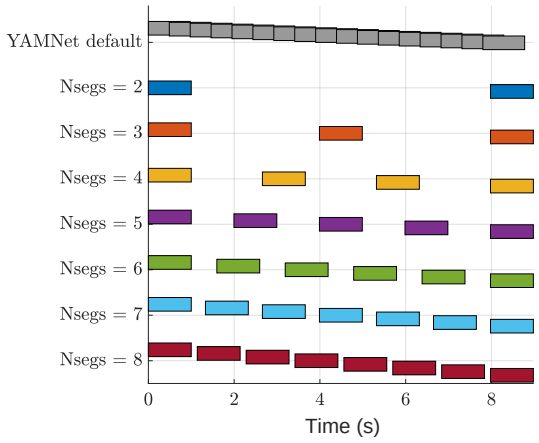

Figure 3. Illustration of segmentation positions for different values of N_{segs} in a long utterance.

4. EXPERIMENTAL RESULTS

This section presents the experimental validation of the proposed speech-based anger detection system. We describe the training protocol, evaluation setup, and methodology used to assess the impact of architectural depth and temporal segmentation. We then report a detailed analysis of the system's performance in different configurations, under both clean and noisy conditions.

The results focus on two main aspects: (1) the influence of the number of retained YAMNet layers (N_{layers}) on classification performance, and (2) the effectiveness of the proposed temporal segmentation strategy controlled by the hyperparameter N_{segs}. In addition, we provide insight into the robustness of the system to acoustic degradation by analyzing its behavior at different signal-to-noise ratios (SNRs), as well as its sensitivity specificity trade-offs using ROC curves.

Overall, the goal of this evaluation is to demonstrate the importance of temporal segmentation and controlled model complexity in improving detection accuracy, robustness, and generalization to unseen speakers.

4.1 Training and Evaluation Setup

The proposed anger detection system was trained and evaluated using the Berlin Emotional Speech Database (emoDB). To assess generalization to unseen speakers, we employed a leave-one-speaker-out cross-validation strategy, in which each speaker was iteratively excluded from the training set and used exclusively for testing. This protocol ensures that the system's performance is robust to inter-speaker variability, a key requirement for real-world emotion recognition applications.

For each fold, the training set consists of all utterances from the remaining nine speakers, while the test set contains all utterances from the excluded speaker. Each experiment was repeated 10 times to account for randomness due to initialization and optimization dynamics, and performance metrics were averaged over these repetitions.

All experiments were conducted using the augmented dataset described in Section 2.1. This data augmentation strategy introduces reverberation and real-world noise to simulate diverse acoustic environments, thereby improving the robustness of the model under challenging conditions.

4.2 Numerical Results

We investigated the influence of two key parameters on the performance of the system: the number of YAMNet layers (N_{layers}) and the number of temporal segments per utterance (N_{segs}). Segmenting the audio into smaller chunks helps the model capture local emotional cues and improves robustness under varying utterance lengths and noise conditions.

In this context, the value $N_{segs} = 1$ refers to the default segmentation mechanism of YAMNet, which uses overlapping windows of 0.975 seconds with a hop size of 0.48 seconds. We call this configuration the baseline.

Figure 4 provides a 3D representation of the mean classification error across different configurations. As observed, increasing both N_{layers} and N_{segs} generally leads to a decrease in the mean error. However, the benefit of increasing N_{segs} begins to saturate beyond six segments, and similarly, the performance gain from adding more layers diminishes at higher depths. These trends suggest that both architectural complexity and temporal granularity play an important role in improving the accuracy of the system, but optimal values must be carefully chosen to avoid unnecessary computational overhead.

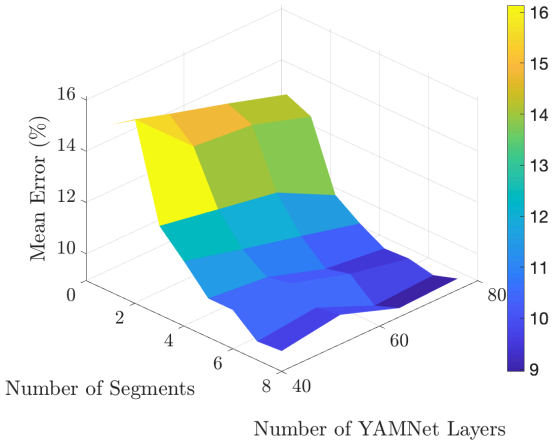

Figure 4. Mean classification error (%) as a function of the number of YAMNet layers N_{layers} and the number of audio segments N_{segs}. Note that the value $N_{segs} = 1$ corresponds to the baseline configuration, which uses YAMNet's default segmentation strategy based on overlapping windows of 0.975 seconds with a 0.48-second hop size.

To better visualize the impact of segmentation, Figure 5 presents the mean classification error for different values of N_{segs} across four representative configurations of network depth. This figure complements the previous 3D plot by providing a clearer view of selected architectures. The results indicate that deeper networks achieve better overall performance, especially when combined with multi-segment strategies. Among the configurations tested, the $N_{layers} = 64$ model offers an excellent compromise between performance and computational efficiency. It consistently outperforms shallower variants and shows near-optimal accuracy compared to more complex configurations.

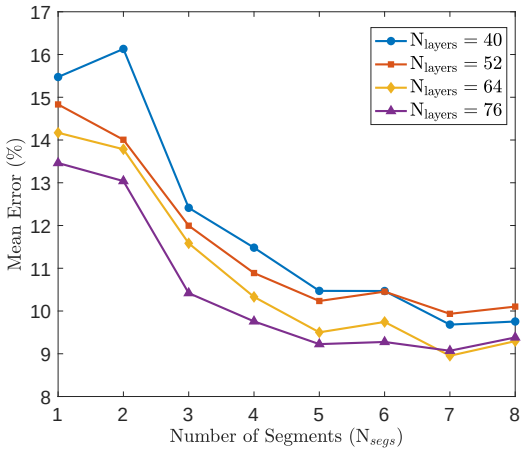

Figure 5. Mean classification error (%) versus number of segments N_{segs} for selected values of N_{layers}. The case $N_{segs} = 1$ corresponds to the baseline configuration, which employs YAMNet's default segmentation using overlapping windows of 0.975 seconds and a 0.48-second hop size.

Based on this analysis, we select the $N_{layers} = 64$ configuration as the reference architecture for the remaining experiments, as it provides a favorable trade-off between performance and complexity.

Figure 6 shows the evolution of the detection error at different SNR levels. As expected, lower SNRs lead to increased error rates, especially for the baseline configuration. In contrast, models using multiple segments show improved robustness across all noise conditions. This confirms the advantage of temporal segmentation, especially under severe acoustic degradation.

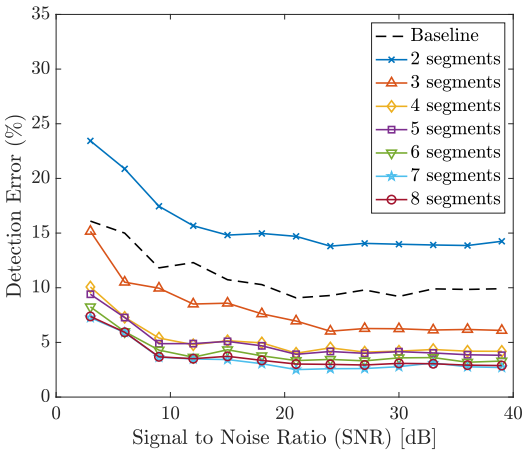

Figure 6. Detection error (%) versus Signal-to-Noise Ratio (SNR) for different segmentation configurations using $N_{layers} = 64$.

The best-performing configuration (N_{segs} = 8) maintains a low detection error even at 3 dB, demonstrating the effectiveness of the segmentation strategy in preserving discriminative information despite significant acoustic degradation. This confirms the benefit of segment-wise representation in handling real-world variability and further supports the findings discussed in Section 2 regarding the impact of our data augmentation methodology.

To gain a deeper understanding of the detection trade-offs at different decision thresholds, we analyze the Receiver Operating Characteristic (ROC) curves obtained for different levels of segmentation, with the model complexity fixed at $N_{layers} = 64$. Figure 7 shows the Probability of Misdetection (P_{MD}) versus the Probability of False Alarm (P_{FA}) for $N_{segs} \in 2, 3, \ldots, 8$, compared to the baseline configuration. As can be seen, most multi-segment models outperform the baseline over the entire range of false alarm rates. However, the N_{segs} = 2 configuration yields inferior results, highlighting that splitting into only two segments may be insufficient to capture the complexity of emotional dynamics. Given the duration distribution of emoDB utterances-mostly around 2-4 seconds-this configuration may fail to effectively isolate emotionally salient regions.

The results show a consistent improvement in recognition as the number of segments increases. In particular, segmentations with $N_{segs} \geq 4$ achieve significantly lower P_{MD} over the entire operating range, indicating

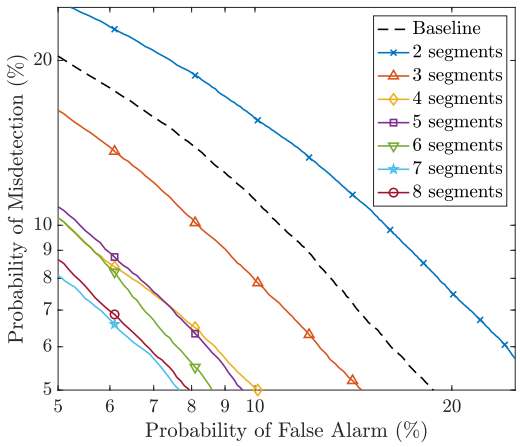

Figure 7. ROC curves comparing different segment configurations for $N_{layers} = 64$. The x-axis represents the Probability of False Alarm (P_{FA}) and the y-axis shows the Probability of Misdetection (P_{MD}), both expressed as percentages.

more reliable detection performance even when the decision threshold is adjusted to control false alarms.

Moreover, the performance gap between configurations narrows at low P_{FA} levels, but remains visible throughout, underscoring the robustness of the multi-segment approach, especially under strict false alarm constraints.

To quantify these observations, Table 1 presents the mean and standard deviation of the probability of false positives (P_{MD}) and the F1 score at a fixed false alarm rate of 5%, along with the relative improvement in P_{MD} over baseline and the statistical significance of the results. These metrics provide complementary insights: while the P_{MD} reflects the proportion of anger events that the system fails to detect, the F1-score captures the balance between precision and recall.

The baseline configuration yields a P_{MD} of 20.37% and an F1 score of 0.807, which is the reference for comparison. As the number of segments increases, there is a consistent improvement in both metrics. In particular, the P_{MD} is nearly reduced by half when increasing to 4 or more segments, reaching a minimum of 8.11% with $N_{segs} = 7$, which is a relative improvement of 60.19%. This configuration also achieves the highest F1-score (0.876), indicating a robust detection capability in both precision and recall dimensions. On the other hand, all improvements beyond baseline were statistically significant with p-values < 0.001, confirming the reliability of the observed differences.

These results confirm that segmentation not only improves average performance, but also yields statistically significant gains in robustness and reliability, validating the effectiveness of the proposed approach.

5. CONCLUSIONS

This paper presents a robust anger detection system for speech that is designed to operate reliably under variable and acoustically adverse conditions. The proposed approach uses transfer learning with YAMNet and systematically analyzes the impact of two key factors: the number of layers retained from the pre-trained model (N_{layers}) and the number of temporal segments extracted per utterance (N_{segs}).

Through extensive experimentation on the emoDB dataset, we have shown that increasing the temporal granularity by splitting utterances into multiple fixed-length segments leads to significant improvements in detection performance. In particular, configurations with $N_{segs} \geq 4$ achieved considerably lower false alarm rates and higher F1-scores, with statistically significant gains over the baseline. These results confirm the value of segment-wise representation in capturing localized emotional cues that may be diluted in global analyses.

The main scientific contribution of this work is to demonstrate that combining fine-grained temporal processing with controlled architectural depth yields significant gains in both accuracy and robustness for speech-based anger detection. The proposed system maintains strong performance even under noisy conditions, highlighting its potential for real-world deployment.

Future work will explore the integration of these strategies into real-time systems and their extension to broader emotion recognition tasks in multilingual and cross-domain contexts.

6. ACKNOWLEDGMENTS

This work was funded by the Spanish Ministry of Science and Innovation under Project PID2021-129043OB-I00 (funded by MCIN/AEI/10.13039/501100011033/FEDER, EU), and by the Community of Madrid and University of Alcalá under project EPU-INV/2020/003.

Table 1. Mean and standard deviation of the probability of misdetection (%) and the F1-score obtained for a probability of false alarm of 5%, for different values of N_{segs}, using a YAMNet network configuration with $N_{layers} = 64$.

N_{segs}	Prob. of misdetection mean (%) [SD (%)]	F1-score mean (%) [SD (%)]	Relative improvement (%)	p-value
Baseline	20.37 [1.23]	0.807 [0.008]	-	-
2	25.61 [1.74]	0.774 [0.011]	-25.71 %	< 0.001
3	16.21 [1.06]	0.831 [0.006]	20.42 %	< 0.001
4	10.30 [0.99]	0.865 [0.005]	49.44 %	< 0.001
5	10.81 [0.67]	0.862 [0.004]	46.93 %	< 0.001
6	10.32 [1.16]	0.864 [0.006]	49.33 %	< 0.001
7	8.11 [0.91]	0.876 [0.005]	60.19 %	< 0.001
8	8.66 [1.69]	0.873 [0.009]	57.50 %	< 0.001

7. REFERENCES

[1] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[2] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[3] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE transactions on multimedia*, vol. 20, no. 6, pp. 1576–1590, 2017.

[4] S. Latif, H. Cuayáhuitl, F. Pervez, F. Shamshad, H. S. Ali, and E. Cambria, "A survey on deep reinforcement learning for audio-based applications," *Artificial Intelligence Review*, vol. 56, no. 3, pp. 2193–2240, 2023.

[5] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462, 2010.

[6] 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 *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5200–5204, IEEE, 2016.

[7] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[8] G. Research, "Yamnet." `https://github.com/tensorflow/models/tree/master/research/audioset/yamnet`. Accessed: 2025-03-01.

[9] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780, IEEE, 2017.

[10] F. Zhu-Zhou, D. Tejera-Berengué, R. Gil-Pita, M. Utrilla-Manso, and M. Rosa-Zurera, "Computationally constrained audio-based violence detection through transfer learning and data augmentation techniques," *Applied Acoustics*, vol. 213, p. 109638, 2023.

[11] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[12] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Interspeech*, pp. 1517–1520, 2005.