IJCoL
Italian Journal of Computational Linguistics

12-1 | 2026
Italian Journal of Computational Linguistics vol. 12, n. 1 June 2026

On Voice Activity Detection for Italian Spoken Language

Shibingfeng Zhang, Gloria Gagliardi and Fabio Tamburini

Electronic version
URL: https://journals.openedition.org/ijcol/2261
ISSN: 2499-4553

Publisher
Accademia University Press

Electronic reference
Shibingfeng Zhang, Gloria Gagliardi and Fabio Tamburini, "On Voice Activity Detection for Italian Spoken Language", *IJCoL* [Online], 12-1 | 2026, Online since 01 June 2026, connection on 06 August 2026. URL: http://journals.openedition.org/ijcol/2261

On Voice Activity Detection for Italian Spoken Language

Shibingfeng Zhang[‡]
Università di Bologna

Gloria Gagliardi[§]
Università di Bologna

Fabio Tamburini[¶]
Università di Bologna

Voice Activity Detection (VAD) refers to the task of identifying human speech in noisy settings, playing a crucial role in fields like speech recognition and audio surveillance. However, most VAD research has predominantly focused on English, leaving other languages — such as Italian —underexplored. This study aims to evaluate and improve VAD systems for Italian speech, with the ultimate goal of enhancing the speech segmentation component of the Digital Linguistic Biomarkers (DLBs) extraction pipeline for early mental disorder screening. We experimented with multiple VAD systems and proposed a novel ensemble approach that demonstrates improved speech event detection performance. This advancement provides a robust foundation for more accurate early detection of mental health conditions using DLBs in the Italian language.

1. Introduction

Voice Activity Detection (VAD) refers to the task of identifying the presence of human speech within noisy audio signals by classifying utterance segments as either "speech" or "non-speech". Typically, it involves making binary decisions for each frame of a noisy signal (Graf et al. 2015).

VAD has a broad range of applications, serving as a critical component in various areas such as telecommunications, speech recognition systems, and audio surveillance (Mehrish et al. 2023). Nevertheless, the vast majority of current works focus on applying VAD to English despite the fact that several factors can affect its cross-linguistic transfer, potentially leading to suboptimal results. For instance, Voice Onset Time can vary significantly across languages, affecting the system's ability to detect speech activity accurately (Cho, Whalen, and Docherty 2019). Additionally, differences in phonetic structures further complicate system effectiveness across languages. Given these factors, investigating and evaluating various VAD systems on Italian speech holds significant value for the academic community.

Dementia is a syndrome characterized by the impairment of multiple higher cognitive functions, resulting in a loss of functional independence. It poses a significant public

‡ Dept. of Classic Philology and Italian Studies - Via Zamboni 32, 40126 Bologna, Italy.
E-mail: shibingfeng.zhang@unibo.it
§ Dept. of Classic Philology and Italian Studies - Via Zamboni 32, 40126 Bologna, Italy.
E-mail: gloria.gagliardi@unibo.it
¶ Dept. of Classic Philology and Italian Studies - Via Zamboni 32, 40126 Bologna, Italy.
E-mail: fabio.tamburini@unibo.it

health challenge due to its widespread prevalence worldwide. Furthermore, projections indicate that the number of cases could rise to 139 million by 2050.

Extensive research has shown that language is one of the cognitive domains impacted by dementia (Boschi et al. 2017; Gagliardi 2024). Notably, linguistic changes often appear earlier than other clinical symptoms (Eyigoz et al. 2020), prompting a growing body of studies to investigate the potential of linguistic analysis as a screening tool (König et al. 2015; Gagliardi and Tamburini 2021, 2022; Themistocleous, Eckerström, and Kokkinakis 2018, 2020).

Digital Linguistic Biomarkers (DLBs) refer to linguistic features automatically extracted directly from patients' verbal productions that offer insights into their medical state (Gagliardi, Kokkinakis, and Duñabeitia 2021). In their study, Gagliardi and Tamburini (2022) proposed the first DLBs extraction pipeline for the early identification of mental disorders in Italian. The extraction of acoustic and rhythmic features in this tool relies heavily on a preprocessing step consisting of speech segmentation via VAD. The VAD system they adopted was a statistical VAD model called "SSVAD v1.0" (Mak and Yu 2014), which will be presented and compared with other VAD systems in Section 2. In this study, we focus on VAD for the Italian language — an area that remains largely unexplored — with the aim of identifying a system that performs more reliably than the one used in the original pipeline. The outcomes of this research will serve as a fundamental component in the DLB extraction pipeline, replacing the current VAD system. Moreover, this work provides a robust foundation for future projects in this field, enabling more accurate and earlier detection of mental health issues using linguistic biomarkers.

Our main contributions are as follows:

- Testing and evaluation of multiple VAD systems on Italian speech.
- Proposal of an ensemble VAD system that achieves superior performance.

The structure of this paper is as follows. Section 2 provides background on related work and describes VAD systems leveraged in this work. Section 3 details the data resources and performance evaluation metrics. Section 4 presents the experimental setup, including tests on individual VAD systems and ensemble methods applied, and discusses the results. Finally, Section 5 draws the conclusions.

2. Background

This section provides an overview of the background, state-of-the-art developments, and architectures of VAD systems as presented in current literature

Most Voice Activity Detection (VAD) systems treat the task as a binary classification problem applied to each frame of a noisy audio signal, with or without overlapping frames. Depending on their architecture, these systems are typically categorized into two main types: statistical VAD systems and deep neural network (DNN)-based VAD systems.

Statistical VAD systems rely on probabilistic models and statistical signal processing techniques to distinguish between speech and non-speech segments. Common statistical approaches include Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs), and Bayesian frameworks. These methods typically utilize features such as spectral energy, zero-crossing rate, or Mel-Frequency Cepstral Coefficients to model speech and noise characteristics.

For example, in Sohn, Kim, and Sung (1999) the authors proposed a robust statistical VAD system that models the signal using a first-order two-state HMM. In this system, the VAD score of each frame is calculated based on the likelihood ratio between the probability density functions conditioned on two hypotheses: speech absent and speech present. Additionally, the state-transition probability is determined using the likelihood ratio from the previous frame, which helps in maintaining temporal coherence and improving the accuracy of the voice activity detection process.

On the other hand, DNN-based VAD systems leverage the capabilities of deep learning to model complex patterns in audio signals. These systems use neural network architectures, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), or more advanced structures with attention mechanism (Sehgal and Kehtarnavaz 2018). In many studies, CNN and RNN layers are combined to exploit their complementary strengths. Typically, the input audio is first segmented into short-duration frames. Convolutional Neural Networks (CNNs) are then used to extract local features from individual frames, while Recurrent Neural Networks (RNNs) capture the temporal dynamics across successive frames, enabling the modeling of long-range dependencies within the audio signal. For example, Wilkinson and Niesler (2021) employs a stack of convolutional layers to extract features from spectral representations of audio segments. These features are then fed into a bidirectional LSTM layer to produce a binary speech/non-speech classification.

Below, we present a list of the VAD systems evaluated in this project, along with a brief description of each one.

- ***SSVAD v1.0 (Baseline)*** (Mak and Yu 2014) is a statistical VAD system designed to handle low signal-to-noise-ratio (SNR), impulsive noise, and cross talks in interview-style speech files. This system achieved a significant reduction in VAD error rates on the NIST 2010 SRE interview speech dataset (Martin and Greenberg 2010), outperforming several conventional VADs of that time. The system enhances speech segments as a pre-processing step to improve SNR, thereby facilitating subsequent speech/non-speech decisions. SSVAD v1.0 was previously integrated into the older version of the DLBs extraction pipeline (Calzà et al. 2021; Gagliardi and Tamburini 2022) for speech segmentation and serves as the baseline for comparison with other systems in this study.
- ***rVAD*** (Tan, kr. Sarkar, and Dehak 2020) is an unsupervised model comprising two denoising steps followed by a final VAD stage. In the first denoising step, high-energy noise segments are identified and nullified. The second step uses a speech enhancement method to further denoise the signal. Experimental results on the Aurora-2 (Hirsch and Pearce 2000) and RATS (Walker and Strassel 2012) databases show that rVAD achieves very low frame error rate and significantly outperforms multiple supervised and unsupervised VAD methods across various noise types and SNR levels.
- ***WebRTC VAD*** is a system developed by Google for the WebRTC project[1]. Detailed information about this VAD system is limited as it is closed source and undocumented.

1 https://webrtc.org/

- ***Silero*** (Silero Team 2021) is a pre-trained CNN systems with a encoder-decoder architecture. Similar to the WebRTC VAD system, it is closed source and detailed information about its architecture are not publicly available. Evaluation results on benchmarks such as AliMeeting (Yu et al. 2022) show that both WebRTC and Silero are robust systems with strong performance, with Silero tending to achieve superior performance over WebRTC.
- ***GPVAD*** (Dinkel et al. 2021) is a 5-layer DNN composed of CNN and RNN layers. The proposed model employs a data-driven teacher-student learning paradigm for VAD, where a teacher model is initially trained on a source dataset with weak labels to handle vast and noisy audio data. The trained teacher model then provides frame-level guidance to a student model trained on various unlabeled target datasets. Experimental results demonstrate that GPVAD outperforms many other systems, including rVAD, in real-world noisy conditions. It achieves substantially lower frame error rates and higher event-based F1 scores on the DCASE18 benchmark (Dekkers et al. 2018) comparing to rVAD.
- ***Context-aware VAD*** (Jo et al. 2021) is a self-attentive VAD system based on the Transformer architecture (Vaswani et al. 2017). The proposed self-attentive VAD model processes acoustic features extracted from audio input, enhancing it with contextual information from surrounding frames. This is also the only system in this study known to have incorporated an attention mechanism. It was trained and tested on TIMIT (Garofolo et al. 1993) data augmented with diverse noise sources.
- ***Pyannote*** (Bredin et al. 2020) is a pre-trained open-source toolkit for audio processing that involves a VAD model. Similar to GPVAD and Silero, it is a DNN-based model with CNN and RNN modules. It outperforms Silero on various VAD benchmarks such as AMI (Carletta 2007).

3. Experiments

In this section, we outline the experiments carried out, describe the evaluation metrics used, and detail the resources adopted throughout the experimental process.

3.1 Evaluation Dataset

In this work, the CLIPS dataset—*Corpora e Lessici dell'Italiano Parlato e Scritto* ("Corpora and Lexicons of Spoken and Written Italian")[2] (Leoni et al. 2007)—is used to evaluate the VAD systems presented earlier.

The CLIPS dataset comprises approximately 100 hours of speech data, evenly balanced between male and female speakers. It includes a diverse range of regional and situational speech samples, providing a comprehensive representation of the Italian language across various contexts. The CLIPS dataset is divided into five subsets, among which the 'DIALOGICO' and 'LETTO' subsets provide complete temporal alignment between audio and textual transcriptions, totaling approximately 7.5 hours of test data

2 http://www.clips.unina.it/it/

comprised of 120 audio files, with a mean duration of 225.7 seconds. The 'DIALOGICO' group contains 30 files and has a mean duration of 658.57 seconds. The 'LETTO' group contains 90 files and has a much shorter mean duration of 80.60 seconds. The 'DIALOGICO' subset contains dialogues between two speakers, while the 'LETTO' subset consists of recordings of words read aloud from lists. These two subsets were used to evaluate the performance of the selected VAD systems on Italian speech.

3.2 Experiment Settings & Evaluation

To thoroughly evaluate the performance of the various VAD systems, we employed two sets of metrics: frame-level and event-level.

Frame-level metrics assess each 10 ms audio segment independently and decide if the prediction of each segment is true positive, true negative, false positive, or false negative by comparing it with the ground truth, from which we then compute measures such as F1 score, precision, and recall. Furthermore, in order to provide a more comprehensive and task-oriented evaluation that accounts for both missed detections and false alarms in a balanced manner, we adopted a metric named Detection Cost Function (DCF). DCF is a standard evaluation measure used in VAD and speaker diarization tasks, designed to combine the probabilities of misses and false alarms into a single cost value weighted by the prior probability of the target class. In the case of VAD evaluation, the target class is speech class. The DCF adopted in this study is defined as follows:

$$\text{DCF} = P_{\text{target}} \cdot P_{\text{miss}} + (1 - P_{\text{target}}) \cdot P_{\text{fa}}$$

where:

- P_{miss} is the probability of a miss, i.e., the proportion of speech segments incorrectly labeled as non-speech.
- P_{fa} is the probability of a false alarm, i.e., the proportion of non-speech segments incorrectly labeled as speech.
- P_{target} is the prior probability of the speech class, representing how frequently speech occurs in the dataset.

Event-level metrics treat each speech event as a distinct unit to be detected rather than evaluating frame by frame. A detection is considered a true positive if a predicted speech event temporally overlaps with a reference event of the same label by more than 50%. A false positive occurs when the system predicts a speech event for which no corresponding reference event exists, while a false negative refers to a reference speech event that the system fails to detect. From the counts of TP, FP, and FN, standard performance measures such as precision, recall, and F1 score are computed at the event level.

The tools used for evaluation are Evaluation toolbox for Sound Event Detection (Mesaros, Heittola, and Virtanen 2016) and NIST SAD scoring tool [3]. Experiments were conducted on CLIPS dataset using the VAD systems outlined in Section 2. To ensure optimal performance, systems make prediction using its default frame size. Additionally, we applied various ensemble methods to combine the systems' predictions

3 https://www.nist.gov/itl/iad/mig/tools

in order to further improve overall performance. Further details on these ensemble strategies are provided in Section 4.2.

4. Results

This section presents and analyzes the experimental results of the various VAD systems. Section 4.1 reports the results obtained from the individual systems described in Section 2, while Section 4.2 presents the outcomes of combining their predictions through ensemble methods. These ensembling approaches are applied with the goal of enhancing overall performance.

4.1 Single Systems Evaluation

The experimental results of the systems described in Section 2 are summarized in Table 1. These results were obtained using the evaluation methods detailed in Section 3.2.

Table 1
Results of VAD experiment on the systems presented in Section 2. For frame-level results, each 10ms is considered one frame.

System	Frame-level				Event-level		
	F1	Precision	Recall	DCF	F1	Precision	Recall
Context-aware VAD	61.7	96.8	45.3	45.43	12.1	6.5	100
SSVAD (Baseline)	66.1	97.0	50.1	39.86	23.1	13.1	100
WebRTC	62.7	87.5	48.8	24.28	27.0	83.5	16.1
rVAD	73.4	95.1	59.7	34.94	72.2	56.5	99.7
GPVAD	89.5	96.8	83.3	16.81	72.3	58.4	94.9
Pyannote	92.3	94.4	90.3	15.27	**80.3**	84.6	76.5
Silero	**92.5**	94.6	90.6	**15.02**	80.1	75.2	85.6

As can be seen, the majority of the tested systems outperformed the baseline system SSVAD used in the current DLB pipeline at the frame level. A notable pattern from the experiment results is that DNN-based systems, such as Silero, GPVAD, and Pyannote, tend to achieve better results compared to traditional statistical systems like rVAD and SSVAD. However, context-aware VAD is an exception, with an segment-level F1 score of 61.7, which is lower than the baseline SSVAD score of 66.1. Another pattern that can be observed is that systems with relatively lower F1 score do not have lower precision, but rather low recall, meaning that they are more conservative in detecting speech segments and they tend to miss actual speech activity rather than falsely detecting non-speech as speech. As for event-level results, similar to the frame-level results, almost all systems outperformed the baseline. DNN-based systems tend to perform better, with Context-aware VAD being again an exception, as its F1 score is the lowest among all systems.

The poor performance of Context-aware VAD could be attributed to the fact that, unlike GPVAD and Pyannote, it is trained only on the TIMIT dataset (Garofolo et al. 1993) with additional background noise. The TIMIT dataset is a relatively small English speech dataset, containing only 5 hours of audio. This limited training data is likely insufficient for training an attention-based model with a large number of parameters, increasing the risk of overfitting. Another possible reason for this relatively poor performance could be that, while Pyannote and GPVAD are trained on multilingual datasets like DIHARD III (Ryant et al. 2020) and Audioset (Gemmeke et al. 2017), Context-aware VAD is trained solely on English speech. When tested on Italian speech, the system

could suffer a domain shift, resulting in diminished performance. This result further underscores the necessity and importance of conducting experiments to evaluate the performance of VAD systems on the Italian language.

To gain a better understanding of the differences in system performance, a Friedman test was conducted on the F1 scores. The results indicate that both the differences between frame-level results and event-level results are significant. A Nemenyi test with Bonferroni correction was then performed for post-hoc comparisons. The statistical analysis demonstrates that systems GPVAD, rVAD, Silero, and Pyannote exhibit similar performance at both the frame and event levels, while SSVAD, WebRTC, and Context-aware VAD show significantly lower performance at both levels. The former group consistently achieves mean ranks between 2.0 and 3.0, whereas the latter group has mean ranks between 5.7 and 6.2.

After considering the performance at different levels, we tested all combination of three systems to form an ensemble prediction system to generate more accurate VAD results. The architectures of these ensemble systems and the corresponding experimental results are discussed in the following section.

4.2 Ensemble Systems for VAD

This section describes the ensemble methods that combine predictions of systems tested in Section 4.1. It subsequently presents the experimental evaluation results and analysis.

Of the systems presented in Section 2, Silero, Pyannote, GPVAD, and Context-aware VAD assign a score to each frame with a threshold used for making predictions. The other systems do not generate such scores, either due to differences in their architecture or because they are closed-source. This score can be interpreted as the probability of the frame being speech or not. We attempted to ensemble system's predictions using both the probability scores and the final predictions of VAD systems. The major challenge faced by these ensemble methods is that each system uses a different frame size, which complicates achieving alignment for the ensemble system.

We proposed and tested several ensemble strategies:

- **Probability Voting (PV)**: This method involves summing and averaging the probability scores from different predictions. This is the most straightforward ensemble method to address the frame size differences of VAD systems by bypassing alignment requirements entirely. Since different models use different frame sizes, their predictions are not temporally aligned. Rather than attempting to explicitly align frames, PV directly aggregates the probability scores by identifying overlapping prediction windows and averaging the corresponding probabilities. This approach enables straightforward ensemble without requiring alignment of frames across models.
- **Probability Voting with Frame (PV_f)**: In this approach, each audio is first segmented into frames. For each frame, we identify all overlapping frames from all predictions, average their probability scores, and use this average as the probability score for the frame. The frame size of PV_f is 200 ms. The probability score for a frame at time t is computed as:

$$\mathrm{F}_{\text{score}}(t) = \frac{1}{N_t} \sum_{i=1}^{N_t} p_i(t) \tag{1}$$

where $\mathrm{F}_{\text{score}}(t)$ represents the probability score for the frame at time t, N_t represents the number of prediction that overlap with the frame at time t, and $p_i(t)$ represents probability score assigned to the frame at time t by the i^{th} prediction. Note that N_t does not necessarily equal the number of models involved in the ensemble method, but it could be larger.

- **Simple Voting with Frame(SV_f)**: Similar to PV_f, this method segments audio into frames. However, instead of averaging probability scores, it performs simple majority voting based on the predictions of overlapping frames. The frame size of SV_f is 200 ms.
- **Probability Voting with Weight (PV_w)**: This method is akin to PV_f but with a twist: probability scores of overlapping frames from the system predictions are weighted according to their overlap percentage. For each frame, contributions from overlapping predictions are scaled as:

$$w_i(t) = \frac{o_i(t)}{\sum_{j=1}^{N_t} o_j(t)} \tag{2}$$

$$\mathrm{F}_{\text{score}}(t) = \sum_{i=1}^{N_t} w_i(t) \cdot p_i(t) \tag{3}$$

where N_t is the number of prediction that overlap with the frame at time t, $o_i(t)$ is the temporal overlap between the i^{th} prediction and the frame at time t, $w_i(t)$ is the weight of the i^{th} prediction, and $p_i(t)$ is the probability score from the i^{th} prediction. $\mathrm{F}_{\text{score}}(t)$ is the final weighted probability score for the frame at time t.
These weighted scores are then summed to determine the probability score for each frame. Comparing to PV_f, this method accounts for partial overlaps more rigorously, assigning higher weight to predictions with greater temporal overlap.

- **Probability Voting with Sampling (PV_s)**: For a given audio, this method samples timestamps. For each timestamp, it calculates the mean of the probability scores from the three systems, using this mean as the probability score for the timestamp. The sampling rate of PV_s is approximately 33.33 Hz, meaning that one point is sampled every 0.03 seconds.
- **Probability Voting with Bézier curve modelling (PV_b)**: For each prediction from each system, a Bézier curve is generated using control points sampled from the prediction. This approach aims to use a smooth

curve to model the prediction and address the alignment issues caused by different frame sizes of the systems.
Similar to PV_f, each audio segment is divided into frames, and the probability score for each frame is the average of the scores estimated by the Bézier curves. The sampling rate of control points that are used to generate Bézier curve in PV_b is 5 Hz (0.2 seconds).

Table 2
Results of VAD experiment using the SV_f method. For frame-level results, each 10ms is considered one frame.

System	Frame-level				Event-level		
	F1	**Precision**	**Recall**	**DCF**	**F1**	**Precision**	**Recall**
Pyannote	92.3	94.4	90.3	15.27	80.3	84.6	76.5
Silero	**92.5**	94.6	90.6	15.02	80.1	75.2	85.6
GPVAD, Silero, Pyannote	91.7	94.1	89.4	18.2	**84.7**	78.2	92.4
GPVAD, C-a, WebRTC	59.1	92.3	43.5	24.6	62.1	46.7	92.6
GPVAD, SSVAD, C-a	67.4	91.4	53.4	41.9	17.6	9.6	100
GPVAD, SSVAD, WebRTC	59.5	92.4	43.8	24.3	76.6	71.7	82.5
Pyannote, C-a, WebRTC	61.3	91.7	46.0	24.4	69.6	61.3	80.7
Pyannote, GPVAD, C-a	84.2	93.1	76.8	24.5	42.9	26.6	99.0
Pyannote, GPVAD, SSVAD	86.2	92.8	80.5	23.7	57.8	40.8	99.1
Pyannote, GPVAD, WebRTC	62.0	92.8	46.5	20.0	55.3	98.5	38.4
Pyannote, SSVAD, C-a	70.1	91.4	56.9	39.3	17.6	9.7	100
Pyannote, SSVAD, WebRTC	61.5	91.6	46.3	21.6	72.0	85.3	62.3
SSVAD, C-a, WebRTC	47.2	89.9	32.0	43.2	29.6	17.7	99.2
Silero, C-a, WebRTC	61.6	92.4	46.3	23.6	69.8	60.5	82.4
Silero, GPVAD, C-a	84.6	93.5	77.3	23.6	42.9	27.3	99.5
Silero, GPVAD, SSVAD	86.6	93.1	80.9	23.0	57.4	40.3	99.7
Silero, GPVAD, WebRTC	62.3	93.7	46.7	19.0	59.8	93.4	43.9
Silero, Pyannote, C-a	87.3	93.2	82.1	20.4	52.2	35.5	98.7
Silero, Pyannote, SSVAD	89.0	92.9	85.4	20.2	68.4	52.6	97.7
Silero, Pyannote, WebRTC	62.9	92.8	47.6	21.9	47.7	97.9	31.5
Silero, SSVAD, C-a	70.6	91.8	57.3	38.5	17.3	9.5	100
Silero, SSVAD, WebRTC	61.8	92.3	46.4	20.6	72.8	82.4	65.2
rVAD, C-a, WebRTC	52.9	91.2	37.3	34.8	41.1	26.4	93.0
rVAD, GPVAD, C-a	73.1	92.3	60.5	36.7	28.9	16.9	99.8
rVAD, GPVAD, SSVAD	76.5	91.9	65.5	34.6	42.3	26.8	99.6
rVAD, GPVAD, WebRTC	59.1	93.4	43.2	25.6	79.4	91.6	70.1
rVAD, Pyannote, C-a	75.6	91.9	64.2	34.4	27.3	15.8	99.6
rVAD, Pyannote, GPVAD	86.7	92.5	81.6	23.9	74.8	60.1	98.9
rVAD, Pyannote, SSVAD	78.9	91.7	69.2	32.4	43.0	27.4	99.4
rVAD, Pyannote, WebRTC	61.7	92.4	46.2	22.8	58.5	98.1	41.7
rVAD, SSVAD, C-a	57.7	89.7	42.6	50.2	18.0	9.9	100
rVAD, SSVAD, WebRTC	54.5	91.5	38.8	34.9	62.7	48.3	89.2
rVAD, Silero, C-a	75.9	92.4	64.4	33.6	27.0	15.6	100
rVAD, Silero, GPVAD	86.9	92.7	81.8	23.3	73.2	57.9	99.7
rVAD, Silero, Pyannote	89.7	92.8	86.8	20.1	82.1	72.6	94.4
rVAD, Silero, SSVAD	79.3	92.0	69.6	31.7	41.9	26.5	100
rVAD, Silero, WebRTC	62.0	93.2	46.4	21.8	63.3	92.7	48.0

We experimented with all possible system combinations using the SV_f ensemble method, as well as all possible combinations of Silero, Pyannote, GPVAD, and Context-aware VAD using other probability-based ensemble methods, as these are the only systems that generate probability scores. For all probability-based methods, the "speech/non-speech" prediction for each frame is determined by applying a threshold of 0.5 to the probability score.

Table 2 presents the results of all possible combinations to compose the ensemble system using SV_f method. Table 3 shows the results of all possible combinations to compose the ensemble systems using probability score related methods. The evaluation results are derived using the methods presented in Section 3.2.

As shown in Table 2, the ensemble created using the SV_f method did not yield better results than the individual systems at the frame level. The highest frame-level score of 91.7 was achieved by the combination of GPVAD, Silero, and Pyannote, which is still lower than the best performance of the Silero system alone. The DCF scores achieved by ensemble systems are also consistently higher than those of single systems. However, at the event level, the same combination achieved the highest score among all ensemble systems, with an F1 score of 84.7, which is higher than the best score achieved by a single system. Meanwhile, all other combinations yielded scores lower than the best performance of the individual systems.

Table 3
Results of VAD experiment on probability-based ensemble systems. For frame-level results, each 10ms is considered one frame.

System	Method	Frame-level				Event-level		
		F1	Precision	Recall	DCF	F1	Precision	Recall
Pyannote	-	92.3	94.4	90.3	15.27	80.3	84.6	76.5
Silero	-	92.5	94.6	90.6	15.02	80.1	75.2	85.6
Pyannote, GPVAD, Silero	PV	91.5	96.8	86.7	14.5	67.9	51.6	99.4
Pyannote, GPVAD, Silero	PV_f	91.9	95.6	88.6	14.4	**85.9**	84.7	87.2
Pyannote, GPVAD, Silero	PV_s	**93.1**	94.6	91.7	14.4	81.8	81.5	82.2
Pyannote, GPVAD, Silero	PV_w	91.8	96.0	88.0	14.3	85.6	80.0	92.0
Pyannote, GPVAD,Silero	PV_b	93.0	90.5	95.7	18.5	9.5	100	5.0
Pyannote, GPVAD, C-a	PV	90.6	96.6	85.3	15.6	78.9	66.6	96.8
Pyannote, GPVAD, C-a	PV_s	92.6	94.0	91.3	15.4	78.9	90.4	70.0
Pyannote, GPVAD, C-a	PV_w	90.6	96.6	85.3	15.6	78.9	66.6	96.8
Pyannote, GPVAD, C-a	PV_b	92.8	90.7	95.0	18.6	10.6	100.0	5.6
Silero, GPVAD, C-a	PV	88.5	97.5	81.0	17.6	50.4	33.7	100.0
Silero, GPVAD, C-a	PV_s	87.2	96.3	79.6	20.0	67.0	50.4	99.8
Silero, GPVAD, C-a	PV_w	88.7	97.2	81.5	17.4	71.4	55.8	99.1
Silero, GPVAD, C-a	PV_b	92.5	91.1	93.8	18.5	11.1	100.0	5.9
Silero, Pyannote, C-a	PV	92.8	94.7	91.0	14.6	70.1	57.6	89.6
Silero, Pyannote, C-a	PV_s	92.9	93.7	92.2	15.4	77.2	93.6	65.7
Silero, Pyannote, C-a	PV_w	93.0	94.5	91.6	14.4	81.3	88.5	75.1
Silero, Pyannote, C-a	PV_b	93.2	90.0	96.7	18.9	9.3	100.0	4.9

As shown in Table 3, the ensemble systems related to probability score did not achieve results that are prominently better than single systems at the frame level either; PV_s and PV_b methods scores when applied to the combination Pyannote, GPVAD, Silero are only slightly higher by a small margin of 0.6 compared to Silero. Unlike the ensemble systems of the SV_f method, some probability-based ensembles achieve lower DCF scores. This suggests that probability-based fusion helps balancing false alarms and missed detections, even if the overall F1 improvement remains marginal. However,

at the event level, several evident improvements can be observed in the performance of the ensemble systems. Probability-based ensemble systems combining Pyannote, GPVAD, Silero, except for PV_b and PV, outperformed the simple systems at event level, with PV_f achieving an F1 score of 85.9, which is 5.6 points higher than that of Pyannote. This result demonstrates that the ensemble approach can lead to substantial performance gains in detecting the temporal interval in which speech takes place. It is worth noticing that the ensemble method PV_b consistently shows great disparity between its performance at frame level and event level across all combinations. Despite its good performance on frame level, PV_b achieves rather low F1 score on event level, far lower than all other methods, with very high precision but very low recall, meaning that the system is overly conservative in making predictions. The disparity of performance at different levels is likely to be caused by the insufficient number of control points adopted for generating the Bézier curve. However, increasing the number of control points is infeasible due to the high computational complexity of the curve calculation, that is $O(n^2)$ with n being the number of control points.

Given that the ensemble systems composed of GPVAD, Silero, and Pyannote consistently outperformed other combinations across all ensemble methods, a Friedman test, followed by Bonferroni corrected Nemenyi's post-hoc test, was conducted to assess the differences in F1 score performance between the ensemble methods and the individual systems of GPVAD, Silero, Pyannote. At the frame level, the Friedman test test indicates that the differences are not significant. However, at the event level, the results reveal that PV_b's performance is significantly lower compared to the other systems, with a mean rank of 8.77 among nine systems.

In summary, given the performance of the systems, we plan to adopt PV_f as the speech segmentation component of the DLBs extraction pipeline, leveraging the combined predictions of Pyannote, Silero, and GPVAD. While PV_f shows slightly lower frame-level performance compared to the top-performing individual system, it enhances the accuracy in identifying speech intervals. This trade-off is justified by the substantial improvement in the speech event detection performance.

5. Conclusions

In this study, we investigated and improved Voice Activity Detection (VAD) systems for the Italian language — an area that remains relatively underexplored in speech processing. We evaluated a range of systems and developed an ensemble approach to enhance detection accuracy. Our findings demonstrate that aggregating predictions from multiple models leads to improved identification of speech intervals.

This ensemble method will be integrated into the Digital Linguistic Biomarkers extraction pipeline where accurate speech segmentation is essential for the reliable identification of linguistic features used in diagnosing cognitive impairments. By increasing segmentation accuracy, the proposed approach offers a more dependable foundation for downstream clinical analyses.

Future research may focus on refining the ensemble strategy by incorporating additional linguistic features directly into the VAD systems and exploring their synergistic impact. Furthermore, extending the application of this approach to other languages and dialects could broaden its utility.

Acknowledgements

This study was funded by the European Union – NextGenerationEU programme through the Italian National Recovery and Resilience Plan – NRRP (Mission 4 – Education and research), as a part of the project ReMind: an ecological, costeffective AI platform for early detection of prodromal stages of cognitive impairment (PRIN 2022, 2022YKJ8FP – CUP J53D23008380006).

CrediT Author Statement

SZ: Investigation, Software, Formal analysis, Visualization, Writing - Original Draft. GG: Writing - Review & Editing, Project administration, Funding acquisition. FT: Conceptualization, Methodology, Supervision, Writing - Review & Editing.

References

Boschi, Veronica, Eleonora Catricalà, Monica Consonni, Cristiano Chesi, Andrea Moro, and Stefano F. Cappa. 2017. Connected Speech in Neurodegenerative Language Disorders: A Review. *Frontiers in Psychology*, 8:1–21.

Bredin, Hervé, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. Pyannote. audio: neural building blocks for speaker diarization. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020)*, pages 7124–7128, Virtual Barcelona, May 4-8. IEEE.

Calzà, Laura, Gloria Gagliardi, Rema Rossini Favretti, and Fabio Tamburini. 2021. Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia. *Computer Speech & Language*, 65:101113.

Carletta, Jean. 2007. Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. *Language Resources and Evaluation*, 41(2):181–190.

Cho, Taehong, Douglas H. Whalen, and Gerard Docherty. 2019. Voice onset time and beyond: Exploring laryngeal contrast in 19 languages. *Journal of Phonetics*, 72:52–65.

Dekkers, Gert, Lode Vuegen, Toon van Waterschoot, Bart Vanrumste, and Peter Karsmakers. 2018. Dcase 2018 challenge-task 5: Monitoring of domestic activities based on multi-channel acoustics. *arXiv preprint arXiv:1807.11246*.

Dinkel, Heinrich, Shuai Wang, Xuenan Xu, Mengyue Wu, and Kai Yu. 2021. Voice activity detection in the wild: A data-driven approach using teacher-student training. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1542–1555.

Eyigoz, Elif, Sachin Mathur, Mar Santamaria, Guillermo Cecchi, and Melissa Naylor. 2020. Linguistic markers predict onset of Alzheimer's disease. *EClinicalMedicine*, 28:100583.

Gagliardi, Gloria. 2024. Natural language processing techniques for studying language in pathological ageing: A scoping review. *International Journal of Language & Communication Disorders*, 59:110–122.

Gagliardi, Gloria, Dimitro Kokkinakis, and Jon Andoni Duñabeitia. 2021. Editorial: Digital linguistic biomarkers: Beyond paper and pencil tests. *Frontiers in Psychology*, 12:752238.

Gagliardi, Gloria and Fabio Tamburini. 2021. Linguistic biomarkers for the detection of mild cognitive impairment. *Lingue e linguaggio*, (1):3–31.

Gagliardi, Gloria and Fabio Tamburini. 2022. The automatic extraction of linguistic biomarkers as a viable solution for the early diagnosis of mental disorders. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5234–5242, Marseille, France, June. European Language Resources Association.

Garofolo, John S., Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, and Victor Zue. 1993. Timit acoustic phonetic continuous speech corpus. *Linguistic Data Consortium*.

Gemmeke, Jort F., Daniel P.W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *Proceedings of the 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780, New Orleans, LA, USA, March 5-9. IEEE.

Graf, Simon, Tobias Herbig, Markus Buck, and Gerhard Schmidt. 2015. Features for voice activity detection: a comparative analysis. *EURASIP Journal on Advances in Signal Processing*, 2015:1–15.

Hirsch, Hans-Günter and David Pearce. 2000. The aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions. In *ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenges for the Next Millennium"*, Paris, France, September 18-20.

Jo, Yong Rae, Young Ki Moon, Won Ik Cho, and Geun Sik Jo. 2021. Self-attentive vad: Context-aware detection of voice from noise. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021)*, pages 6808–6812, Toronto, Ontario, Canada, June. IEEE.

König, Alexandra, Aharon Satt, Alexander Sorin, Ron Hoory, Orith Toledo-Ronen, Alexandre Derreumaux, Valeria Manera, Frans Verhey, Pauline Aalten, Phillipe H. Robert, and Renaud David. 2015. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimers Dement (Amst)*, 29:112–124.

Leoni, Federico Albano, Francesco Cutugno, Renata Savy, Valentina Caniparoli, Leandro D'Anna, Ester Paone, Rosa Giordano, Olga Manfrellotti, Massimo Petrillo, and Aurelio De Rosa. 2007. Corpora e lessici dell'italiano parlato e scritto.

Mak, Man-Wai and Hon-Bill Yu. 2014. A study of voice activity detection techniques for nist speaker recognition evaluations. *Computer Speech & Language*, 28(1):295–313.

Martin, Alvin F. and Craig S. Greenberg. 2010. The nist 2010 speaker recognition evaluation. In *Interspeech*, volume 2010, page 2726.

Mehrish, Ambuj, Navonil Majumder, Rishabh Bharadwaj, Rada Mihalcea, and Soujanya Poria. 2023. A review of deep learning techniques for speech processing. *Information Fusion*, 99:101869.

Mesaros, Annamaria, Toni Heittola, and Tuomas Virtanen. 2016. Metrics for polyphonic sound event detection. *Applied Sciences*, 6(6):162.

Ryant, Neville, Prachi Singh, Venkat Krishnamohan, Rajat Varma, Kenneth Church, Christopher Cieri, Jun Du, Sriram Ganapathy, and Mark Liberman. 2020. The third dihard diarization challenge. *arXiv preprint arXiv:2012.01477*.

Sehgal, Abhishek and Nasser Kehtarnavaz. 2018. A convolutional neural network smartphone app for real-time voice activity detection. *IEEE access*, 6:9017–9026.

Silero Team. 2021. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. *GitHub repository*. `https://github.com/snakers4/silero-vad`.

Sohn, Jongseo, Nam Soo Kim, and Wonyong Sung. 1999. A statistical model-based voice activity detection. *IEEE signal processing letters*, 6(1):1–3.

Tan, Zheng-Hua, Achintya kr. Sarkar, and Najim Dehak. 2020. rvad: An unsupervised segment-based robust voice activity detection method. *Computer speech & language*, 59:1–21.

Themistocleous, Charalambos, Marie Eckerström, and Dimitrios Kokkinakis. 2018. Identification of Mild Cognitive Impairment From Speech in Swedish Using Deep Sequential Neural Networks. *Frontiers in Neurology*, 9:975.

Themistocleous, Charalambos, Marie Eckerström, and Dimitrios Kokkinakis. 2020. Voice quality and speech fluency distinguish individuals with Mild Cognitive Impairment from Healthy Controls. *PLoS ONE*, 15(7):e0236009.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Walker, Kevin and Stephanie M. Strassel. 2012. The rats radio traffic collection system. In *Odyssey 2012 The Speaker and Language Recognition Workshop*, pages 291–297, Singapore, June.

Wilkinson, Nicholas and Thomas Niesler. 2021. A hybrid cnn-bilstm voice activity detector. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021)*, pages 6803–6807, Toronto, Ontario, Canada, June. IEEE.

Yu, Fan, Shiliang Zhang, Pengcheng Guo, Yihui Fu, Zhihao Du, Siqi Zheng, Weilong Huang, Lei Xie, Zheng-Hua Tan, DeLiang Wang, et al. 2022. Summary on the ICASSP 2022 multi-channel multi-party meeting transcription grand challenge (M2MeT). In *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, Singapore, May. IEEE.