Intelligent Systems with Applications 29 (2026) 200614

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications

# Enhancing token boundary detection in disfluent speech

Manu Srivastava [a,c,b,*], Marcello Ferro [a,1], Vito Pirrelli [a,1], Gianpaolo Coro [c,1]

[a] Istituto di Linguistica Computazionale "Antonio Zampolli" – CNR, Pisa, Italy
[b] Università di Pisa, Pisa, Italy
[c] Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo" – CNR, Pisa, Italy

ARTICLE INFO

*Keywords:*
Automatic Speech Recognition
Statistical analysis
Disfluencies
Voice Activity Detection

ABSTRACT

This paper presents an open-source Automatic Speech Recognition (ASR) pipeline optimised for disfluent Italian read speech, designed to enhance both transcription accuracy and token boundary precision in low-resource settings. The study aims to address the difficulty that conventional ASR systems face in capturing the temporal irregularities of disfluent reading, which are crucial for psycholinguistic and clinical analyses of fluency. Building upon the WhisperX framework, the proposed system replaces the neural Voice Activity Detection module with an energy-based segmentation algorithm designed to preserve prosodic cues such as pauses and hesitations. A dual-alignment strategy integrates two complementary phoneme-level ASR models to correct onset–offset asymmetries, while a bias-compensation post-processing step mitigates systematic timing errors. Evaluation on the READLET (child read speech) and CLIPS (adult read speech) corpora shows consistent improvements over baseline systems, confirming enhanced robustness in boundary detection and transcription under disfluent conditions. The results demonstrate that the proposed architecture provides a general, language-independent framework for accurate alignment and disfluency-aware ASR. The approach can support downstream analyses of reading fluency and speech planning, contributing to both computational linguistics and clinical speech research.

## 1. Introduction

Speech disfluencies, such as filled pauses and repetitions, represent disruptions in the natural flow of speech and are common across speakers and contexts. Their occurrence is shaped by speaker-specific and environmental factors, and their detection, classification, and localisation are relevant to both clinical and non-clinical domains. Prior research has shown that disfluencies provide valuable insights into language planning and cognitive load (Corley & Stewart, 2008; Lindström et al., 2008; Segbroeck et al., 2014). In this context, temporal annotation is often required; however, this work remains largely manual and time-intensive, and is typically performed with tools such as PRAAT (Gkoumas et al., 2024). This difficulty arises because accurate alignment between disfluent speech and its transcription remains imperfect. Although disfluencies are generally more frequent in spontaneous speech, in clinical research on disfluent speech, read speech is often preferred over spontaneous speech to ensure a quantifiable standard (Ash et al., 2023; Fiorin et al., 2015; Pakhomov et al., 2013; Sheikh & Kodrasi, 2024; Shriberg, 2001). In fact, while spontaneous speech generally contains more disfluencies, it also introduces uncontrolled linguistic and contextual variability, making it difficult to isolate the specific timing phenomena associated with reading difficulties. In contrast, read speech provides a stable textual reference and enables the measurement of disfluent events — such as hesitations, omissions, or repetitions — relative to a known linguistic target. This characteristic is essential in clinical and psycholinguistic investigations of dyslexia and other reading disorders, where the goal is to assess the reader's fluency and timing accuracy against the intended text (Dawson et al., 2019; Gala & Ziegler, 2016; Kearns & Whaley, 2019; Pedersen & Larsen, 2010; Ramlan et al., 2023). Accordingly, this study uses read speech to analyse disfluent speech, because read speech better reflects the cognitive and motor processes underlying reading performance rather than spontaneous language production.

To investigate the alignment between disfluent speech and its transcription, this study uses Automatic Speech Recognition (ASR) to generate precise word-level transcriptions and temporal markers for read speech.

* Correspondence to: via Moruzzi 1, 56124, Pisa, Italy.
*E-mail addresses:* manu.srivastava@cnr.it (M. Srivastava), marcello.ferro@ilc.cnr.it (M. Ferro), vito.pirrelli@ilc.cnr.it (V. Pirrelli), gianpaolo.coro@isti.cnr.it (G. Coro).
[1] via Moruzzi 1, 56124, Pisa, Italy.

https://doi.org/10.1016/j.iswa.2025.200614
Received 23 October 2025; Received in revised form 21 November 2025; Accepted 4 December 2025
Available online 6 December 2025

Recent developments in ASR, including Whisper (Radford et al., 2023) and its extensions — WhisperX, FasterWhisper (Systran, 2025), and Whisper-timestamped (Louradour, 2023) — have substantially advanced speech-to-text capabilities by adopting large end-to-end deep neural network architectures. This innovation is consistent with that observed across diverse scientific domains, including Medical Science, in which deep learning models have demonstrated effectiveness in modelling complex, nonlinear, and dynamic systems (Akkilic et al., 2024; Sabir, Abdelkawy, et al., 2025; Sabir et al., 2022; Sabir, Assaad, et al., 2025; Sabir et al., 2023, 2021; Sanchez et al., 2018).

Nonetheless, the performance of end-to-end ASRs remains limited for low-resource languages such as Italian, particularly when handling speech with natural or pathological disfluencies. Key challenges include: (i) isolating target speech in multi-speaker environments, (ii) mitigating hallucinated insertions (words not present in the audio), and (iii) preventing the omission of disfluent elements during transcription.

The Whisper models provide temporal information for detected words, a feature essential for analysing disfluent speech. The task of *token* boundary detection (identifying the smallest meaningful units within speech) has a long research history (Liu, Shriberg, et al., 2006; Liu et al., 2023; Stolcke et al., 1998; Yildirim & Narayanan, 2009; Zhou et al., 2024). Three main Whisper-based approaches to this task have been proposed (Yamasaki et al., 2023).

1. The application of the Wav2Vec2 ASR to generate character probabilities for each speech frame, followed by an alignment process based on Dynamic Time Warping (DTW) of these probabilities to the characters predicted by Whisper (Giorgino, 2009). This methodology is adopted in WhisperX (Bain et al., 2023).
2. The execution of DTW on cross-attention weights for each predicted token. This approach is employed in Whisper-Timestamped (Louradour, 2023), Faster Whisper (Klein, 2023), and OpenAI-Whisper (Radford et al., 2023).
3. The analysis of the probability distribution over the timestamps recorded at the end of each token, used in Stable Whisper (Jian, 2023) and Whisper.CPP (Gerganov, 2023).

All existing Whisper-based approaches show limitations when transcribing disfluent speech, largely due to the model's pre-training objectives, which prioritise transcription in a default *stylistically intended* manner (Romana et al., 2024). As described by Lea et al. (2023), ASR transcriptions can be categorised as either *intended* — where speech errors, false starts, or changes in thought are omitted — or *verbatim*, which faithfully represents all uttered content, including unintended elements. Whisper generally produces *intended* transcriptions, aiming to capture the speaker's communicative intent while suppressing disfluencies as unintentional deviations. Moreover, the model has been shown to exhibit hallucinations when processing aphasic speech (Koenecke et al., 2024), leading to misalignments between the acoustic signal and the token sequence and thereby reducing word-boundary accuracy. For analyses of disfluent speech, adopting a *verbatim* transcription style is therefore preferable. These challenges become more pronounced when the model processes longer audio sequences.

Several Whisper-based extensions, including WhisperX and Whisper -Timestamped, integrate a Voice Activity Detector (VAD) — such as Pyannote (Bredin & Laurent, 2021; Bredin et al., 2020) or Silero VAD (Silero Team, 2024) — to segment speech before alignment. While this method offers robustness against hallucinations, it often increases token omissions (Yamasaki et al., 2023) due to inherent VAD limitations, such as front/rear-end clipping and sensitivity to background noise (Louradour, 2023). Furthermore, VAD performance varies by application context (Patil & Patil, 2024). Deep neural network-based VADs, like those used in WhisperX, are particularly dependent on data availability and tend to underperform in low-resource languages (Coro et al., 2021). Their accuracy also degrades with atypical speech patterns (Jiang et al., 2024). In fact, their reliance on statistical patterns learned from large-scale, fluent datasets can result in the misinterpretation of hesitation-related pauses or non-sonorant phonemes as background noise, thereby truncating or merging speech segments at inappropriate points. Such premature or inaccurate segmentation can suppress or distort the acoustic evidence of disfluencies, which is the phenomenon of interest in this study. For these reasons, we explore alternative segmentation methods that allow more explicit control over silence thresholds and low-energy regions associated with hesitation and repair, thereby better aligning with the prosodic reality of disfluent speech. In this context, token boundary detection is particularly relevant because hesitations, repetitions, and prolongations disrupt the temporal regularity of the signal (Betz, 2020; Collard, 2009; Romana, 2024). Disfluencies alter the expected timing between lexical items and often introduce acoustic irregularities — such as elongated pauses or truncated word endings — that challenge conventional ASR models trained on fluent speech. When token boundaries are inaccurately estimated, these deviations are either smoothed out or misattributed, preventing a temporal reconstruction of the speaker's output. Therefore, improving token boundary precision is not only a technical enhancement but a prerequisite for the reliable detection and analysis of disfluent phenomena, which depend on millisecond-level timing information to reveal the structure and dynamics of speech planning.

Approaches to token boundary detection so far have included signal-processing systems based on prosodic features, which used pitch, energy, and duration cues to enhance word and syllable boundary detection, especially in noisy scenarios (Ananthakrishnan & Narayanan, 2008; Biron et al., 2021; Hasija et al., 2022; Kingsbury et al., 1998; Kocharov et al., 2019; Wu et al., 1998). Other approaches have used deep-learning models based on convolutional and recurrent artificial neural networks to improve acoustic modelling and token boundary detection, by learning hierarchical representations of prosodic and spectral features (Hau, 2014; Sabu et al., 2021; Stehwien et al., 2020; Vitale et al., 2024). Other studies have integrated syntactic, semantic, and prosodic features to improve sentence boundaries and transcription interpretation (Kocharov et al., 2019; Rehbein et al., 2020; Treviso et al., 2017; Widiaputri et al., 2023). However, these methods are generally not language-agnostic, poorly suited for low-resource languages, and do not ensure accurate detection of token boundaries in disfluent read speech.

In this work, we present an open-source Italian-language ASR pipeline designed to improve both transcription quality and token-boundary detection in disfluent read speech. Built upon the WhisperX framework (Bain et al., 2023), the proposed system replaces the neural VAD module with a syllabic-scale, energy-based segmentation algorithm that better captures prosodic cues such as hesitations, filled pauses, and self-corrections. It also introduces a dual-alignment strategy that integrates two complementary phoneme-based ASR models to refine onset–offset precision, together with a token-marker post-processing step that corrects systematic temporal biases and ensures local boundary consistency.

The pipeline is evaluated against WhisperX (version 3-large-weights) on two corpora of read Italian speech — READLET (child disfluent reading) and CLIPS (adult read speech) — using token transcription accuracy and boundary detection metrics. We demonstrate that the energy-based segmentation effectively substitutes the default neural VAD, improving temporal alignment in disfluent regions, while the dual-alignment strategy further enhances boundary precision. The pipeline yields a language-agnostic, modular ASR architecture that offers a more robust and interpretable solution for analysing disfluent read speech, particularly in low-resource language contexts where prosodic cues and temporal accuracy are critical for both linguistic and clinical research.

This paper is organised as follows: Section 2 details the assumptions, materials, and models used in our system. Section 3 presents performance comparisons across all proposed configurations, and Section 4 summarises and discusses the findings.

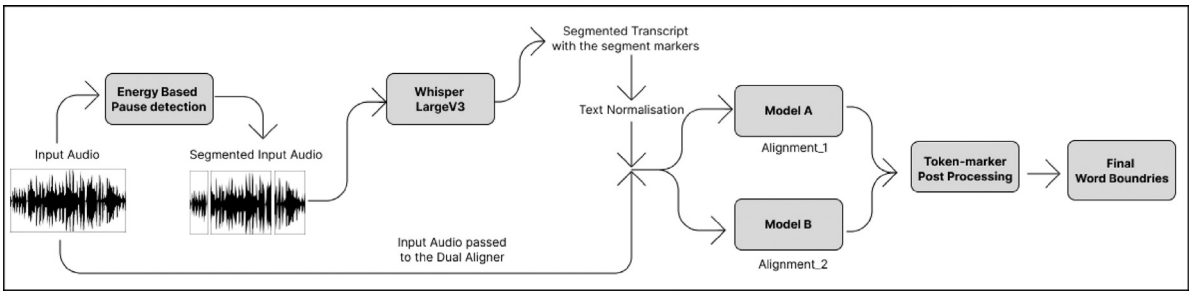


**Fig. 1.** Overview of the proposed ASR pipeline for disfluent read speech. The process begins with energy-based pause detection, which segments the input waveform into prosodically coherent units based on energy thresholds, replacing the default neural VAD of WhisperX. Then, each audio segment is transcribed through the WhisperX encoder–decoder ASR (Whisper Large V3), producing preliminary text with segment markers. The transcriptions undergo text normalisation before being passed to the dual aligner, which produces two alignments. The alignments are then merged using a token-marker post-processing module that corrects systematic onset–offset deviations and enforces temporal consistency between adjacent tokens. The resulting token boundaries provide temporally precise, prosodically aware alignments that better capture disfluent speech.

## 2. Material and methods

This section outlines the preliminary observations that informed the design of our system (Section 2.1), followed by the methodology for extracting token boundary markers and optimising speech transcription (Sections 2.2–2.5). It then presents the benchmark data and evaluation metrics used for performance comparison (Section 2.6.2). The overall pipeline is illustrated in Fig. 1.

### 2.1. Observations

The development of our pipeline was guided by the following key observations:

**Segmentation**: End-to-end ASR systems such as WhisperX and Speechbrain Ravanelli et al. (2021) typically segment the input speech signal into fixed-size windows (e.g., 30 s; Radford et al. (2023)) due to architectural context-length constraints. This segmentation can disrupt contextual dependencies between speech units, particularly when boundaries occur mid-unit, thereby reducing transcription accuracy across segments with abrupt transitions. To address this, we adopted a heuristic segmentation strategy based on Inter-Pausal Units (Prakash & Murthy, 2024) and Tone Unit Segmentation (Coro et al., 2022; D'Anna & Petrillo, 2003), providing a more contextually coherent and stable alternative.

**Use of fine-tuned models for alignment**: As emphasised by the authors of WhisperX (Bain et al., 2023), language-specific fine-tuned models significantly improve token alignment. Accordingly, we systematically evaluated multiple fine-tuned models in a plug-and-play configuration to identify performance trends and limitations within our framework.

**Leveraging aligner tendencies**: Through experimentation with various fine-tuned aligners, we observed consistent behavioural tendencies that could be exploited to improve accuracy. Based on these insights, we designed a *dual aligner* framework that combines complementary aligners to counterbalance their respective biases and enhance overall token-boundary precision.

### 2.2. Audio segmentation

The WhisperX speech recognition and token boundary extraction pipeline relies on a Voice Activity Detection (VAD) module for input audio segmentation (Bain et al., 2023). This VAD is implemented as an artificial neural network that models the function $\Omega : A \rightarrow y_t$, where $A$ denotes the input audio waveform and $y_t$ the probability of speech presence at time $t$. The probabilistic outputs are subsequently binarised through a *smoothing* and *decision* phase. Additionally, the VAD constrains segment duration to a maximum of 30 s; if exceeded, truncation occurs at the $y_t$ point with the lowest speech probability. Segments identified as *voiced* are then passed to the transcription modules.

Although generally effective (Bain et al., 2023), this approach shares a common limitation with other VAD-based methods: it frequently misclassifies non-sonorant phonemes at word onsets or offsets as noise. This occurs because such phonemes (e.g., */s/*) lack fundamental frequency and exhibit broad spectral distributions similar to noise within the typical 10–20 ms phonetic-length analysis window. Consequently, the VAD tends to contract token boundaries toward sonorant regions, a bias that worsens under noisy conditions. This limitation is particularly critical in the presence of disfluencies, where pauses, false starts, and self-repairs frequently produce low-energy regions that do not correspond to actual silence (Evangelopoulos & Maragos, 2006; Hamzah & Jamil, 2019; Marzinzik & Kollmeier, 2002). VADs trained to optimise for speech presence rather than linguistic continuity tend to excise these regions, thereby fragmenting the disfluent event and reducing transcription fidelity (Hughes & Mierle, 2013; Zhang & Wang, 2015). In contrast, analyses based on energy profiles, conducted at the syllabic scale, have shown improved robustness to noise and reverberation, suggesting their suitability for more accurate boundary detection (Greenberg & Kingsbury, 1997; Kingsbury et al., 1998; Vitale et al., 2024; Wu et al., 1998). Several studies have demonstrated that the syllabic-scale energy profile effectively smooths short-term fluctuations and emphasises sustained energy islands that correspond to tone-unit and inter-pausal structure, making pause markers easier to detect under moderate non-stationary noise (Coro et al., 2022; Cutugno et al., 2002; Ludusan et al., 2011; Wu et al., 1998). This profile also preserves prosodic and hesitation cues that frame-level neural VADs may suppress when trained on predominantly fluent or broadcast corpora (Coro et al., 2021; Vitale et al., 2024). This approach is better suited to ensure that segmentation is guided by acoustic continuity rather than by categorical speech/non-speech classification. It can highlight energy transitions that mark disfluencies, such as prepauses, filled pauses, and hesitative elongations, ultimately improving both word boundary and disfluent speech detection.

A concept used in energy and prosodic analyses is *tone unit* (TU), an acoustic correlate of a dialogue unit defined as a segment of speech produced with a coherent intonation contour (D'Anna & Cutugno, 2003). Tone units often represent complete and meaningful utterances, making them valuable for analysing verbal communication quality (Coro et al., 2022) and improving ASR performance on salient speech segments (Coro et al., 2021; Ludusan et al., 2014). A TU typically ends following a period of high-energy speech. Previous TU detection studies have demonstrated that syllabic-scale analysis windows (100–200 ms) yield robust segmentation under diverse noise conditions (Coro et al., 2022; Cutugno et al., 2005, 2002). Given the limitations of the WhisperX-embedded VAD explained in Section 1, we replaced it with a syllabic-scale energy-based segmentation algorithm. This algorithm

computes the RMS energy within 100-ms frames to identify pause markers that delineate dialogue unit sequences of predefined length. The underlying assumption is that such sequences preserve sufficient contextual information to enhance ASR accuracy.

Following this rationale, we implemented an energy-based segmentation method that identifies low-energy regions succeeding continuous high-energy segments. The low-energy threshold is treated as a tunable hyperparameter. This process effectively detects syllabic-scale energy drops corresponding to pause markers, in line with established TU and Inter-Pausal Unit (IPU) detection approaches (Kane et al., 2014; Nahar & Kai, 2020; Yang et al., 2022). Unlike conventional VAD systems, our method targets the detection of pauses rather than the onsets or codas of speech units.

**Algorithm 1** Energy-Based Audio Segmentation - Phase 1

**Require:** Audio file sampled at 16 kHz, maximum chunk duration

**procedure** COMPUTE RMS
  Load the audio file as an array.
  Set frame duration to $F = 100$ms and window-shift to $H = 50$ ms.
  Compute RMS values.
**end procedure**

**procedure** IDENTIFY SILENT FRAMES
  Set relative silence threshold to $\theta = 0.1\%$.
  Create a boolean array (`silentFrames`) where:

$$\texttt{silentFrames}[i] = \begin{cases} \text{True} & \text{if relative RMS}[i] < \theta \\ \text{False} & \text{otherwise} \end{cases}$$

  Store the `timeMarkers` vector containing the frame start times.
**end procedure**

**procedure** DETECT SILENT SEGMENTS
  Set `startTime` to 0.
  Initialise an empty object list: `energyIslands`
  **for** each frame $i$ **do**
    Accumulate `silentFrames`[$i$] if it is equal to TRUE
    **if** Accumulation > 200ms **then**
      Set `endTime` to `timeMarkers`[$i$]
      Append the `[startTime;endTime]` energy island to `energyIslands`
      Set `startTime` = `endTime`
    **end if**
  **end for**
**end procedure**

Our process consists of two main phases. The first (Algorithm 1) operates a frame-by-frame calculation of Root Mean Square (RMS), the continuous power of an audio signal frame $k$:

$$\text{RMS}_k = \sqrt{\frac{1}{F}\sum_{n=0}^{F-1} x_k[n]^2} \tag{1}$$

with $\text{RMS}_k$ being the RMS value of the $k$th frame, $F$ the frame length in samples, and $x_k[n]$ the $n$th signal sample in the $k$th frame. $\text{RMS}_k$ measures the average loudness of the sound over time. In our analysis, we used a syllabic-scale frame length of 100 ms with 50% overlap between consecutive frames to produce a smoother RMS curve. This choice is grounded in psycho-acoustic analysis: A 100 ms window provides a temporal resolution that aligns with linguistically meaningful syllables and is close to the average syllable duration in Italian (Coro et al., 2023; Cutugno et al., 2002). The 50% overlap reduces frame-edge effects and produces a smoother RMS envelope, thereby improving the detection of gradual energy transitions, such as hesitations and filled pauses, without excessively increasing computational cost (Wu et al., 1998).

As a further step, the process creates a Boolean representation of the frames, where zeros indicate low-energy or silent frames and ones indicate energetic frames (containing either speech or noise). A fixed threshold on $\text{RMS}_k$ distinguishes between these two cases. This threshold was calculated based on a statistical analysis over a development corpus (Section 2.6.1) and was set, as default, to 0.1% of the maximum $\text{RMS}_k$ value. A sensitivity analysis on this parameter is reported in Section 3.2.

A scan of this Boolean vector can reveal continuous regions of energetic signal with a 50 ms temporal resolution (given the 50% overlap between the frames). Our process searches for signal energy "islands" ending with low-energy frames having at least the length of a long-syllable (200 ms). This length is short enough to be findable within a fast-speech dialogue (which might not contain word-length pauses), and long enough to guarantee that it is unlikely a pause inside a word (Ludusan et al., 2011; Wu et al., 1998). The algorithm's output is a report of the start and end markers for the signal energy islands.

**Algorithm 2** Energy-Based Audio Segmentation - Phase 2

**procedure** SEGMENT AUDIO
  Initialize empty object list `validSegmentMarkers`
  **for** each `island` in `energyIslands` **do**
    Compute `island` duration ($d_i$)
    Check if $d_i \leq n$, with $n$ =`maximum chunk duration`
  **end for**
  **if** At at least one island did not satisfy the condition **then**
    Discard all `energyIslands`
    Divide the entire signal into segments of $n$ duration
    Save the `[startTime;endTime]` markers of each segment to `validSegmentMarkers`
  **else**
    Create an empty `audio sub-segment` ($a$)
    **for** each `island` in `energyIslands` **do**
      Compute $d_i$
      **if** $(d_a + d_i) \leq n$ **then**
        Attach the `island` to the `audio sub-segment`
      **else**
        Save the current `[startTime;endTime]` markers of $a$ to `validSegmentMarkers`
        Start a new $a$ containing `island`
      **end if**
    **end for**
  **end if**
  Return `validSegmentMarkers`
**end procedure**

The second phase (Algorithm 2) applies a greedy algorithm to iteratively process the identified islands and generate audio segments of up to $n$ seconds (*maximum segment duration*). The parameter $n$ has a notable impact on WhisperX transcription performance, as demonstrated in the results section. The algorithm systematically evaluates each island to verify that its duration remains within the $n$-second limit. When this condition is met, the islands are merged sequentially into segments that do not exceed $n$ seconds. This procedure eliminates the reliance on the probabilistic truncation mechanism employed by the ANN-based VAD described earlier.

If this condition is not satisfied, the algorithm divides the entire audio signal into uniform intervals of $n$ seconds, disregarding previously detected islands. Such cases are rare, especially for $n$ values near 30 s, and typically occur when background noise exhibits energy levels comparable to speech at the syllabic scale. Under these conditions, energy-based features alone cannot reliably distinguish between speech and silence. Alternative approaches, such as adaptive energy thresholding or pitch-curve analysis (Kolář & Liu, 2010; Kumari et al., 2022; Liu,

Chawla, et al., 2006; Liu, Shriberg, et al., 2006; Shezi & Reddy, 2020), could be used but would risk introducing additional biases, including internal word or syllable segmentation errors. Therefore, uniform signal segmentation serves as a non-informative prior assumption regarding token boundaries, deferring the differentiation between consecutive tokens to the ASR stage.

### 2.3. Audio transcription

As a subsequent step, our methodology applies the WhisperX encoder–decoder ASR module (Whisper Large V3) to the segments extracted in the preceding phase. It iteratively applies the WhisperX ASR to the individual segments to obtain segment-focused transcriptions with associated word markers. This operation enhances transcription by utilising the contextual acoustic information derived from the TUs. The underlying hypothesis is that providing the ASR system with islands of speech that feature no abrupt ending cuts, encompass finite and meaningful dialogue interactions, and have a duration optimised for WhisperX performance would enhance transcription accuracy for each island. Consequently, the concatenation of the individual island transcriptions is anticipated to yield an overall improvement in performance.

Our audio transcription approach comprises three primary operations: (i) audio cutting and preparation based on the energy-island markers, (ii) transcription of individual sub-audio segments, and (iii) merging of sub-audio transcriptions to produce a final audio transcription with associated word-boundary markers in the WhisperX format. Algorithm 3 illustrates these steps in pseudocode.

**Algorithm 3** Audio Transcription

**Require:** Audio file, `validSegmentMarkers` from the previous step.

**procedure** PREPARE AUDIO SEGMENTS FOR TRANSCRIPTION
  Prepare an empty object list: `audioSegments`
  **for** each `segment` in `validSegmentMarkers` **do**
    Cut the audio from `startTime` to `endTime` in the original audio file
    Store the audio to a file
    Update `audioSegments` with the audio path
  **end for**
**end procedure**
**procedure** TRANSCRIBE AUDIO
  Prepare an empty object list: `audioTranscriptions`
  **for** each `audioFile` in `audioSegments` **do**
    Execute WhisperX on the `audioFile` to generate a transcription
    Store the audio transcription to `audioTranscriptions`
  **end for**
**end procedure**
**procedure** MERGE TRANSCRIPTIONS
  Prepare an empty transcription file in the WhisperX output format: `finalTranscription`
  **for** each `audioTranscription` in `audioTranscriptions` and `segment` in `validSegmentMarkers` **do**
    Create a sub-transcription with work markers in the WhisperX format
    Add the sub-transcription to `finalTranscription`
  **end for**
**end procedure**
Save `finalTranscription` to a file (`finalTranscriptionFile`)
Return `finalTranscriptionFile`.

The final transcription is a text file containing sentence and word transcriptions. It follows the same format used by the standard WhisperX pipeline, which includes detailed information for each segment: the transcription text, start time, and end time. The transcription file is passed to the next step along with the original audio file. The two files are the basis for the subsequent alignment process.

### 2.4. Transcription post-processing

As an additional step, our methodology produces a revised transcription that more closely conforms to human transcription standards, thereby facilitating downstream word alignment. To this end, a post-processing procedure is applied to re-evaluate and refine the initial transcription, improving both performance assessment and automatic alignment. The resulting output adheres to the orthographic conventions adopted in major Italian corpora, as described in Section 2.6.

The post-processing procedure begins by converting the transcription to lowercase and removing punctuation, enabling subsequent alignment to focus exclusively on individual word matches. In the second step, numerical symbols are converted into their verbal equivalents, reflecting standard Italian annotation practices. The third step addresses truncation and elision phenomena characteristic of Italian, following conventions commonly adopted by human annotators. For instance, phrases such as *l'ingresso* (*the entrance*) and *d'immaginazione* (*of imagination*) are often rendered by WhisperX as single tokens, though they comprise two distinct lexical units. In corpus-based Italian transcription, these would typically be annotated as *l ingresso* and *d immaginazione* to explicitly mark word boundaries and facilitate fine-grained disfluency analysis (Dovetto et al., 2022; Leoni et al., 2007). Our post-processing replicates this rationale by appropriately separating such terms, thereby emphasising individual word boundaries and any potential pauses between them during subsequent alignment. Finally, pause annotations are removed to ensure that alignment focuses solely on lexical items rather than treating pauses as additional tokens.

After completion, the refined transcription, formatted with segment-level markers consistent with the WhisperX standard, is passed to the next stage for word alignment.

### 2.5. Transcription alignment and token-marker post-processing

Token boundary detection is a crucial process that involves the accurate identification of the start and end times of each token transcribed from an audio signal. Traditionally, this task has been achieved through *Forced Alignment*, which optimally aligns the token sequence in a transcription with the audio at the word or phoneme level. This is usually accomplished using a Hidden Markov Model (HMM) in conjunction with the Viterbi Algorithm (Brugnara et al., 1993; McAuliffe et al., 2017). More recent methodologies have used deep neural networks and Connectionist Temporal Classification (CTC) for this task (Fan et al., 2023; Kürzinger et al., 2020).

Among the various token-alignment strategies that have been integrated with Whisper to date (Bain et al., 2023; Jian, 2023; Louradour, 2023), two methods have exhibited superior performance (Yamasaki et al., 2023): (i) processing the ASR transcription through a phoneme-based ASR, such as Wav2Vec2 (Baevski et al., 2020), and (ii) employing Dynamic Time Warping (DTW) between the Whisper cross-attention scores and the speech transcription (Louradour, 2023).

In our methodology, we have chosen to use a phoneme-based ASR, as it is particularly suitable for analysing fine-grained speech structures and can effectively account for both coarticulation and contextual effects (Siohan, 2017; Weise et al., 2025), thereby resulting in reliable synchronisation between the text and the audio signal. We conducted tests on various phoneme-based ASRs, including those specifically fine-tuned for Italian (Conneau et al., 2020; DBDMG, 2022; Grosman, 2022a, 2022b; Prakash & Murthy, 2024), to identify a suitable replacement for the default aligner employed by WhisperX for Italian.

WhisperX currently embeds a phoneme-based Wav2Vec2 ASR (Baevski et al., 2020) pre-trained on the VoxPopuli corpus (Wang et al., 2021). However, based on evaluations conducted on our development

data (Section 2.6), it exhibits inferior performance compared to alternative ASRs in accurately detecting the onset and offset boundaries of disfluent speech. Generally, all models evaluated (including the default WhisperX aligner) demonstrated good performance in detecting either token onsets or offsets, but not both concurrently. No single ASR model consistently outperformed the others in detecting both token onsets and offsets.

Consequently, our pipeline addresses token onset and offset detection through two distinct models. To achieve this, we independently selected the optimal token-onset aligner and the optimal token-offset aligner from the ASRs tested. Our alignment process operates on the transcription generated in the preceding step of the pipeline (Section 2.4). Thus, it involves passing the transcription as input to two models, yielding two sets of token onset and offset markers. Finally, a token-level merging and consistency-check procedure is employed to integrate the two alignments, yielding a unified set of token boundary annotations. We will hereafter refer to this methodology as the *dual aligner*, as it capitalises on the unique, complementary characteristics of two distinct aligners. This dual-alignment strategy is particularly advantageous for disfluent speech, where timing irregularities can cause a single aligner to drift systematically toward early or delayed token boundaries (Diwakar & Karjigi, 2020; Knowles et al., 2015). Disfluent events produce variable acoustic patterns that challenge the temporal assumptions of most alignment models. By pairing two distinct phoneme-based ASRs with complementary strengths (one more accurate for token onsets, the other for offsets), our method compensates for these opposing biases. The merged alignment ensures that both the beginnings and endings of tokens remain temporally consistent, even in the presence of hesitation or self-correction.

The optimal aligner for token onsets on our development data was the "fine-tuned XLS-R 1B model for speech recognition in Italian" by Grosman (2022a). This model is trained on various Italian corpora, including Common Voice 8.0 (Ardila et al., 2020), Multilingual TEDx (Salesky et al., 2021), Multilingual LibriSpeech (Pratap et al., 2020), and Voxpopuli (Wang et al., 2021). The base ASR, prior to tuning, is the XLS-R, a large-scale multilingual ASR model developed by Facebook, comprising up to 2 billion parameters pre-trained on 436,000 h of audio data. It employs an internal Transformer architecture incorporating CTC and Attention mechanisms. Hereafter, we will refer to this model as *MODEL A*.

Notably, we identified a systematic bias in onset detection within this model, necessitating a 60 ms backward shift of all annotations to optimise the alignments. This correction was derived empirically from a distributional analysis of onset deltas computed on a subset of the READLET corpus. The resulting histogram exhibited a clear unimodal peak centred at 60 ms (indicating systematic delay of the automatic boundaries) with an approximately Gaussian spread of 20 ms. Applying a uniform backward shift of 60 ms effectively recentered this distribution around zero, eliminating the systematic offset while leaving the residual variance unchanged. This shift was demonstrated to optimise performance on the evaluation corpora as well (Section 3.3). Furthermore, it is worth noting that this shift is a configurable parameter in our pipeline, which can be adjusted if differing settings are required for other corpora.

For token offsets, the optimal aligner was XLSR-53 by Conneau et al. (2020). This model has a Wav2Vec 2.0 architecture trained on 53 languages, derived from diverse corpora, including Babel (Gales et al., 2014), Multilingual LibriSpeech (Pratap et al., 2020), Common Voice (Ardila et al., 2020), VoxPopuli (Wang et al., 2021), and VoxLingua (Valk & Alumäe, 2021). The model acquires a cross-lingual speech representation in its encoder module and is subsequently fine-tuned on language-specific data to enhance transcription performance. In our case study, we employed the fine-tuned version for Italian. Similar to *MODEL A*, this ASR model is also based on an architecture combining Transformer modules with CTC and Attention mechanisms. Hereafter, we will refer to this model as *MODEL B*. Unlike *MODEL A*, this model did not require bias correction; however, our pipeline allows users to implement correction for *MODEL B* if deemed necessary.

We integrated the Italian fine-tuned *MODEL A* and *MODEL B* into our pipeline as two additional processes executed on the output of the preceding step to generate token onset and coda boundaries separately (Algorithm 4). A subsequent merging process generates a comprehensive transcription and token boundary sequence. Finally, a consistency check is conducted between the onsets and offsets of consecutive tokens, rectifying any inconsistencies such as a overlap of temporal boundary markers between subsequent tokens.

**Algorithm 4** Alignment with dual aligner

**Require:** Original audio file; the transcribed segments with token boundaries, in the WhisperX format, from the previous step (`transcript`)

**procedure** AUDIO ALIGNMENT WITH DUAL ALIGNMENT
  Initialize objects `alignedA` and `alignedB` to empty objects
  Pass `transcript` to `Model A` and store the results (transcription and markers) in `alignedA`
  Repeat with ModelB and store the results in `alignedB`
**end procedure**
**procedure** ONSET-BIAS CORRECTION
  **for** each `token` in `alignedA` **do**
    Subtract 60 ms from the `token` onset
    Update `alignedA`
  **end for**
**end procedure**
**procedure** OUTPUT MERGING
  Initialize object `alignedMerge` to an empty object list
  **for** each `token` in `transcript` **do**
    Take the `token` onset (`On`) from `alignedA`
    Take the `token` offset (`Off`) from `alignedB`
    Create the new token boundary `mergedBoundary=[On, Off]`
    Add `token` and `mergedBoundary` to `alignedMerge`
  **end for**
**end procedure**
**procedure** CORRECT OVERLAP BETWEEN SUBSEQUENT TOKENS
  **for** each `token` in `alignedMerge` **do**
    Get the index `i` of `token` in the token sequence
    **if** (`token[i]_onset` < `token[i-1]_offset`) **then**
      Calculate the difference $d$ between `token[i]_onset` and `token[i-1]_offset`
      Reduce `token[i-1]_offset` by $d/2$
      Increase `token[i]_onset` by $d/2$
      Update `alignedMerge`
    **end if**
  **end for**
  Save `alignedMerge` to `alignedMergeFile` file in the WhisperX format.
  Return `alignedMergeFile`.
**end procedure**

The output of this process is the final audio transcription of our pipeline, along with the token boundary report, in the WhisperX format. This output was the subject of our benchmarking.

### 2.6. Data and benchmarking

This section describes the data we used as development and test sets for our pipeline (Section 2.6.1). Then it explains the systems and metrics we used for performance benchmarking (Section 2.6.2).

#### 2.6.1. Data

We conducted a benchmarking study of our pipeline using two distinct Italian corpora: READLET (Ferro et al., 2024) and CLIPS (Leoni, 2014). These corpora differ significantly in size, scope, and audio

**Table 1**
Statistical attributes of the used corpora.

| Attribute | READLET | CLIPS |
|---|---|---|
| Total duration (min) | 12.52 | 56.69 |
| SNR (mean ± dtd Dev) | 15.72 ± 4.72 dB | 7.38 ± 36.20 dB |
| Number of tokens | 1506 | 10,613 |
| Average length (s) | 75.14 | 5.65 |
| Min. duration (s) | 54.76 | 1.73 |
| Main speaker type | Children (7–10 years) | Adults |

quality, as outlined in Table 1, and they facilitated a comprehensive evaluation of ASR performance across a diverse range of speech variability.

The READLET corpus was created by the Communication Physhiology lab of the Institute for Computational Linguistics of the National Research Council of Pisa (Italy: CNR-ILC), in collaboration with the Cognitive Neuroscience lab of the International School for Advanced Studies of Trieste (Italy: SISSA) as part of the ReadLet Italian Strategic project (2017W8HFRX). The corpus consists of audio and tactile recordings of silent and aloud reading conducted by children and adults. All subjects were instructed to read a text on a tablet, while concurrently pointing to the text with a finger. The tablet facilitated recording finger positions corresponding to the text, and synchronising them with speech recordings in oral reading sessions. The collection was mainly geared towards early identification of reading problems through a multimodal analysis integrating both movement and speech data (Marzi et al., 2023, 2022).

Given that our primary focus was on read speech and the disfluencies common to readers with dyslexia or similar impairments, we selected a relevant subset of READLET tailored to this context. Notably, 80% of our selection comprised recordings of oral reading by early graders, specifically children aged 7–10 years, who were reported to be difficult readers (occasionally with special education needs). Thus, these audio recordings predominantly featured young readers exhibiting frequent disfluencies.

Our READLET selection was manually annotated with temporal boundaries for all uttered words, using PRAAT (Boersma, 2011) as the annotation software to generate separate onset and offset annotation layers. The transcriptions adhered to a verbatim style, wherein disfluencies and mispronunciations were recorded as they were articulated, even if the word transcriptions did not conform to grammatical norms. This approach aligns with the transcription conventions typical of end-to-end ASRs (Coro et al., 2025). Filler sounds (e.g., *aa*, *um*, *ehm*) were not explicitly transcribed but were denoted with a $< filler >$ tag alongside their temporal boundaries. Similarly, pauses were annotated with a $< pause >$ tag and reported with their respective temporal boundaries. Given its foundational intent to study disfluent speech, we used the READLET selection as a development set to parameterise the heuristic sub-processes within our pipeline. For completeness, we also calculated ASR performance measurements on this corpus.

The Corpora and Lexicons of Spoken and Written Italian (CLIPS) is a data collection developed between 1999 and 2004, supported by the Italian Ministry of Education, Universities, and Research. Its primary aim was to facilitate advances in automatic speech processing and linguistic analysis of the Italian language (Università di Napoli Federico II, 2024). CLIPS encompasses audio recordings of dialogic speech (DIALOGICO) and read speech (LETTO) performed by native speakers from 15 different Italian cities. It captures a wide array of speech variability, encompassing regional, social, stylistic, and individual differences. However, the DIALOGICO segment of CLIPS comprises spontaneous speech, which could have introduced additional complexities to the ASR task and hindered the analysis of disfluent speech (Coro et al., 2007; Cutugno et al., 2005; Dutta et al., 2022; Southwell et al., 2022). Moreover, spontaneous speech is not the target modality for modern ASRs. Consequently, we focused our evaluation on the LETTO section of CLIPS, which contains instances of disfluent speech accounting for at least 1% of all uttered words. This section includes meticulous annotations of long and short pauses, breathing/exhaling/inhaling sounds, throat-clearing, background noise, and filler sounds. To ensure consistency, we standardised the annotations to match those in READLET. We employed this corpus, which was an order of magnitude larger than READLET, as the primary test set for performance benchmarking.

#### 2.6.2. Benchmarking

Our pipeline is a modification of the off-the-shelf WhisperX ASR pipeline. Therefore, we used WhisperX as a *baseline* for performance benchmarking. Our pipeline and WhisperX share the same encoding-decoding module (Whisper version 3-large-weights) but use different methods for audio segmentation (Section 2.2), transcription alignment, and token-marker post-processing (Section 2.5). Table 2 reports a summary of the principal differences between our pipeline (*Reference*) and WhisperX.

We organised our pipeline evaluation as an ablation study: we incrementally substituted modules in the baseline (from segmentation to token-marker post-processing) to measure performance changes in token transcription and boundary detection each time. Table 3 summarises the configurations we tested and the label assignment of each configuration.

For performance evaluation, we calculated the Word Error Rate (**WER**) as the percentage of tokens that were incorrectly transcribed compared to the manual annotation:

$$\text{WER} = \frac{S + D + I}{N}$$

where $S$, $D$, and $I$ stand for the number of token substitutions, deletions, and insertions, respectively; and $N$ is the total number of tokens in the manual transcriptions. We also calculated the rates of the individual error types, i.e., the Insertion Error Rate (**IER**) $(I/N)$, the Deletion Error Rate (**DER**) $(D/N)$, and the Substitution Error Rate (**SER**) $(S/N)$. While IER and DER quantified potential hallucinations and missed tokens, SER measured token mis-transcription due to phonetic similarity with other words or phoneme sequences, a common error in end-to-end ASRs that do not incorporate externally-provided language models (Coro et al., 2021).

As a further performance measure, we calculated the Character Error Rate (**CER**), which provides fine-grained insights into errors by accounting for substitutions, insertions, deletions, and correct matches at the character level.

For the evaluation of token boundary detection, we initially computed the deltas ($\delta(milliseconds) = |\texttt{ASR\_t} - \texttt{Annotated\_t}|$) of the boundaries of correctly transcribed tokens in relation to the manual annotations. Subsequently, we analysed the deltas for the onsets and offsets independently, by examining their distributions. Thirdly, we calculated the Intersection-over-Union (**IOU**) distribution for the token segments. Finally, as a comprehensive measure of quality, we assessed the percentage of tokens that were accurately transcribed and exhibited boundaries with less than a 50 ms difference (phonetic-size discrepancy) compared to the manual annotations. We refer to this measure as the Collared Matching Rate (**CLMR**). Formally, $CLMR = \frac{N_{\text{predicted}}}{N_{\text{annotated}}}$; with $N_{\text{annotated}}$ being the total number of manually *annotated* tokens, and $N_{\text{predicted}}$ being the number of *predicted* tokens with onset and offset times satisfying the condition $|\text{onset}_{predicted} - \text{onset}_{annotated}| < 50$ ms AND $|\text{offset}_{predicted} - \text{offset}_{annotated}| < 50$ ms.

## 3. Results

This section presents the results of the benchmarking processes described in Section 2.6.

**Table 2**
Summary of the principal differences between our pipeline (*Reference*) and the off-the-shelf WhisperX pipeline (*Baseline*).

| Model | Segmentation | Aligner | Token-marker post-processing |
|---|---|---|---|
| Baseline | ANN-based VAD | Phoneme-based Wav2Vec2 ASR | No |
| Reference | Energy-based | Dual Aligner | Yes |

**Table 3**
Configurations Used in the Ablation Study.

| Model | Segmentation | Aligner | Token-marker post-processing |
|---|---|---|---|
| $Baseline_1$ | ANN-based VAD | Phoneme-based Wav2Vec2 ASR | No |
| $Baseline_2$ | ANN-based VAD | Dual Aligner | No |
| $Baseline_3$ | ANN-based VAD | Dual Aligner | Yes |
| $Reference_1$ | Energy-based | Phoneme-based Wav2Vec2 ASR | No |
| $Reference_2$ | Energy-based | Dual Aligner | No |
| $Reference_3$ | Energy-based | Dual Aligner | Yes |

**Table 4**
Performance comparison between the baseline off-the-shelf WhisperX ASR ($Baseline_1$) and our pipeline using energy-based segmentation ($Reference_1$) on the READLET and CLIPS corpora.

| Model | READLET | | | | | CLIPS | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | CER (%) | WER (%) | IER (%) | DER (%) | SER (%) | CER (%) | WER (%) | IER (%) | DER (%) | SER (%) |
| $Baseline_1$ | **2.61** | **4.24** | **0.07** | 2.98 | **1.19** | 4.62 | 6.06 | **2.76** | 1.07 | **2.23** |
| $Reference_1$ | 3.14 | 4.61 | 0.69 | **2.66** | 1.26 | **4.49** | **6.05** | **2.76** | **0.88** | 2.41 |

### 3.1. Performance comparison on speech transcription

As a preliminary analysis, we assessed the differences in speech transcription performance between the off-the-shelf WhisperX ASR ($Baseline_1$) and our pipeline variant utilising energy-based segmentation as the sole alternative module ($Reference_1$). This comparison aimed to evaluate changes in transcription performance following the substitution of the initial signal segmentation module. Given that the token boundary detection module did not influence transcription results, we excluded the other benchmark pipelines from this analysis. The performance metrics were compared using the READLET and CLIPS corpora (Table 4). To avoid bias in the overall results caused by the substantial differences in corpus sizes, we reported the performance metrics for each corpus separately. This approach ensures a fair assessment of performance. Notably, the READLET corpus, being more aligned with our goal of analysing disfluent speech, balances quantity with quality and was therefore regarded with equal importance as the CLIPS corpus.

On the READLET corpus, $Baseline_1$ exhibited a lower Word Error Rate (WER) compared to $Reference_1$ (4.24% vs 4.61%), suggesting superior transcription accuracy for $Baseline_1$. Conversely, $Reference_1$ demonstrated a slightly lower WER on the CLIPS corpus (6.05% vs. 6.06%), indicating that the two systems achieved comparable overall accuracy. The highlighted significant discrepancies, as well as the others reported herein, were checked for significance using a paired t-test ($p < 0.05$).

A notable advantage of the energy-based segmentation in $Reference_1$ was its ability to deliver tone-unit-like audio segments to the ASR model for recognition. A TU generally encompasses a complete and meaningful sentence, thereby enabling the ASR to rely on grammar (automatically abstracted during training) to enhance sentence transcription. This mechanism resulted in a reduction of the deletion rate across both corpora (2.98% vs 2.66% for READLET and 1.07% vs 0.88% for CLIPS). However, the use of grammatical structures led to an increase in word insertions, which were likely hallucinations, particularly evident in the READLET corpus for $Reference_1$ (0.07% vs 0.69% IER). This effect was balanced on the CLIPS corpus, where the IER remained equal at 2.76%.

Furthermore, grammar correction contributed to an increase in SER between $Baseline_1$ and $Reference_1$ (1.19% vs 1.26% on READLET and 2.23% vs 2.41% on CLIPS), particularly in cases of disfluent speech, as words were occasionally re-corrected according to expected pronunciation and grammatical rules. The increases in IER and SER for $Reference_1$ were also reflected in CER, with a more pronounced discrepancy observed in the READLET corpus (2.61% vs 3.14%) compared to CLIPS (6.06% vs 6.05%).

One notable example of the advantages and challenges introduced by energy-based segmentation is illustrated in Fig. 2. This figure emphasises the tendency of $Reference_1$ to generate a grammatically consistent sentence despite the presence of a false start. In this instance, $Reference_1$ endeavours to produce the intended word sequence (*"che passano"*), which consequently leads to an increase in substitution and character error rates. Conversely, $Baseline_1$ eliminates such inconsistencies, resulting in a coherent yet abbreviated sentence.

### 3.2. Performance comparison on token marker detection

For the second performance comparison, we assessed the token-boundary detection effectiveness of all benchmarked pipelines on the READLET (Table 5) and CLIPS corpora (Table 6). A discernible gradient of increasing performance improvement was observed for the baseline and reference pipelines as we replaced the off-the-shelf modules with our custom modules. This assertion was substantiated by analysing the distributions of the deltas across the corpora (Fig. 3), which showed a notable leftward shift, indicating a general reduction in token boundary detection errors upon integrating our modules. Furthermore, a decrease in standard deviation was evident, suggesting an overall enhancement in precision. This corroborated the significant impact of the dual aligner and the token-marker post-processing combination on token boundary detection.

The energy-based segmentation module also contributed further to handling disfluent speech in the READLET corpus. This enhancement enabled $Reference_3$ to achieve superior performance in boundary

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Manual Annotation** | ••• | da | li | guarda | le | macchine | che | pass | che | passano | veloci | ••• |
| **Baseline_1** | ••• | da | li | guarda | le | macchine | --- | --- | che | passano | veloci | ••• |
| **Reference_1** | ••• | da | li | guarda | le | macchine | che | passano | che | passano | veloci | ••• |

**Fig. 2.** Example of word and character deletions and insertions from the compared baseline and developed (reference) pipelines. Red lines indicate missing tokens in the ASR transcription. Green letters indicate character insertions by the ASR.

**Table 5**
Performance comparison on the READLET corpus between the benchmarked pipelines in the detection of token onsets, offsets, and boundary-transcription matching. The delta values presented here are absolute values $\delta = |\texttt{ASR\_t} - \texttt{Annotated\_t}|$, reported in milliseconds. IOU and CLMR are reported as percentages.

| Metric | $\text{Baseline}_1$ | $\text{Baseline}_2$ | $\text{Baseline}_3$ | $\text{Reference}_1$ | $\text{Reference}_2$ | $\text{Reference}_3$ |
|---|---|---|---|---|---|---|
| Onset delta *mean* | 89 | 77 | 42 | 94 | 72 | **38** |
| Onset delta *median* | 71 | 63 | **23** | 74 | 62 | **23** |
| Offset delta *mean* | 45 | 41 | 40 | 49 | **36** | **36** |
| Offset delta *median* | 24 | 20 | 20 | 24 | 20 | **19** |
| IOU *mean* | 63.51 | 66.46 | **80.77** | 62.83 | 66.82 | **80.86** |
| IOU *median* | 70.94 | 74.64 | **85.76** | 68.76 | 73.82 | **86.04** |
| CLMR *mean* | 24.19 | 30.71 | **70.46** | 23.25 | 31.12 | 70.20 |
| CLMR *median* | 19.78 | 27.38 | 68.35 | 19.63 | 26.45 | **69.74** |

**Table 6**
Performance comparison on the CLIPS corpus between the benchmarked pipelines in the detection of token onsets, offsets, and boundary-transcription matching. The delta values presented here are absolute values $\delta = |\texttt{ASR\_t} - \texttt{Annotated\_t}|$, reported in milliseconds. IOU and CLMR are reported as percentages.

| Metric | $\text{Baseline}_1$ | $\text{Baseline}_2$ | $\text{Baseline}_3$ | $\text{Reference}_1$ | $\text{Reference}_2$ | $\text{Reference}_3$ |
|---|---|---|---|---|---|---|
| Onset delta *mean* | 71 | 60 | **21** | 70 | 60 | **21** |
| Onset delta *median* | 66 | 57 | **16** | 66 | 56 | **16** |
| Offset delta *mean* | 27 | **20** | **20** | 25 | **20** | **20** |
| Offset delta *median* | 20 | **15** | **15** | 20 | **15** | **15** |
| IOU *mean* | 58.83 | 64.39 | 82.75 | 59.93 | 64.77 | **82.86** |
| IOU *median* | 65.07 | 69.96 | **86.98** | 66.27 | 70.38 | 86.78 |
| CLMR *mean* | 21.79 | 36.10 | 87.23 | 22.99 | 37.41 | **87.89** |
| CLMR *median* | 21.42 | 35.71 | 88.88 | 22.22 | 36.36 | **89.47** |

detection and token-segment similarity as measured by the IOU. Additionally, $\text{Reference}_3$ attained the highest performance in terms of combined correct transcription and boundary detection (measured by the CLMR) on the CLIPS corpus, while demonstrating performance comparable to $\text{Baseline}_3$ on the READLET corpus. Notably, the IOU and CLMR discrepancies were checked to be statistically significant ($p < 0.05$) through a Wilcoxon signed-rank test for both READLET and CLIPS.

We also conducted a sensitivity analysis for the RMS silence threshold of 0.1%, indicated as the optimal percentage of the maximum frame energy in $\text{Reference}_3$: The threshold was varied between 0.01% and 1% in increments of 0.01%, and the resulting segmentations were evaluated on the READLET development subset using CLMR as a joint measure of transcription and boundary accuracy. CLMR gradually increased, up 0.1%, and declined at higher values, indicating an optimal trade-off between over- and under-segmentation. Lower thresholds (<0.05%) tended to merge adjacent speech units, reducing boundary precision, whereas higher thresholds (>0.3%) fragmented continuous speech into excessively short segments.

To further investigate the behaviour of our system in disfluent contexts, we analysed token boundary corrections across three prevalent disfluency types annotated in the READLET corpus: (i) short pauses (less than 0.5 s), (ii) long pauses (greater than 1 s), and (iii) word repetitions or truncations. The results demonstrated that our pipeline effectively addressed boundary misalignment across all categories. The greatest improvement in boundary detection was observed for short pauses, with a 75% relative improvement over WhisperX. This suggests that our energy-based segmentation technique successfully detected low-energy hesitation regions. Instead, a 20% relative improvement was observed in the detection of long-pause boundaries, because background noise can create artificial energy islands in long silent segments. As for repetitions and truncated words, the dual-alignment strategy achieved a 33% relative improvement in boundary detection, effectively managing irregular phoneme durations introduced by self-corrections. These findings confirm that the proposed approach enhanced token boundary detection, specifically in segments containing disfluent speech.

These findings underscore that the newly integrated modules significantly enhanced token boundary detection performance compared to the off-the-shelf WhisperX ASR.

### 3.3. Analysis of marker detection

As a final analysis, we compared the statistical distributions of token onset and offset deltas across the test CLIPS corpus. We specifically focused on the basic Baseline and Reference pipelines (v1) in contrast to their enhanced versions that incorporated all modules (v3) (Fig. 4).

The distributions exhibited a leftward shift when passing from the off-the-shelf WhisperX pipeline ($\text{Baseline}_1$) to its alternative implementation that embedded the dual aligner and token-marker post-processing ($\text{Baseline}_3$). Notably, the $\text{Baseline}_3$ distribution displayed a median of approximately zero milliseconds, resulting from the combined effects of the two alternative modules. Although the token-marker offset was developed using the READLET corpus, it also demonstrated effective performance on the CLIPS corpus. A similar leftward shift was observed between the $\text{Reference}_1$ and $\text{Reference}_3$ distributions.

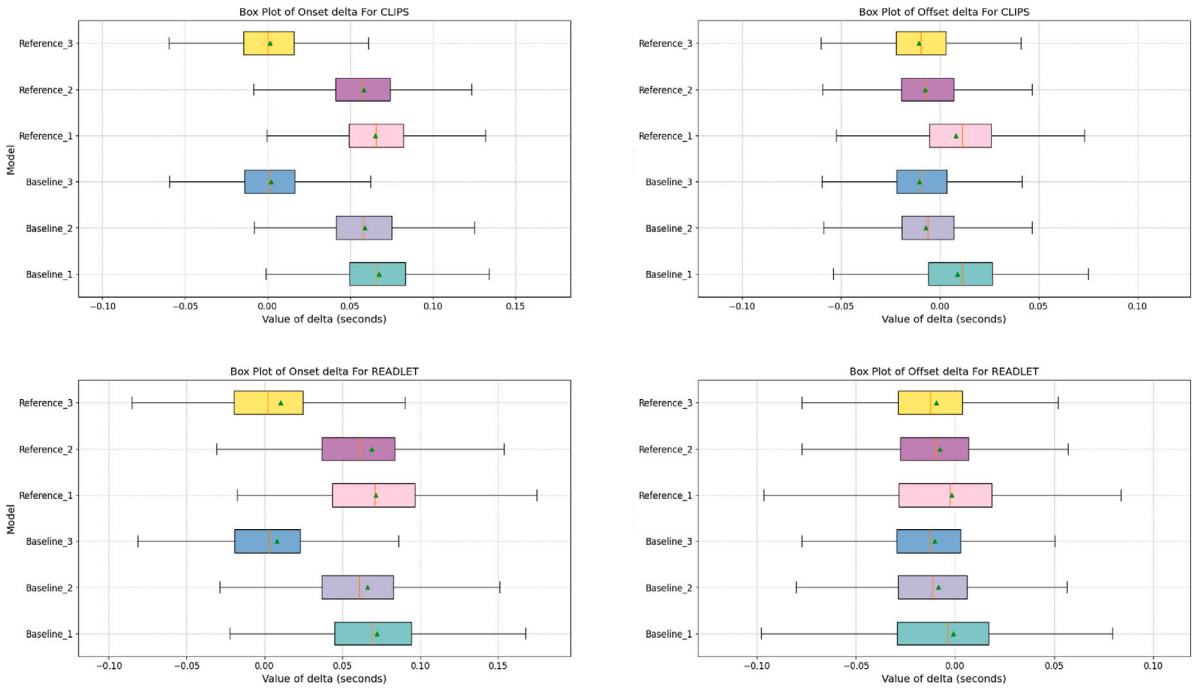


**Fig. 3.** The distributions of the onset and offset deltas for all benchmarked pipelines on the READLET and CLIPS corpora. Each distribution shows the temporal difference (in seconds) between the predicted and manually annotated token boundaries. Negative values indicate that the automatic boundary occurs before the reference annotation, while positive values indicate a delay. These signed deltas highlight systematic biases in boundary estimation, specifically whether a model tends to predict token onsets and offsets earlier or later than manual annotation.

These findings underscore the practical advantages conferred by the dual aligner and token-marker post-processing for token boundary detection. Furthermore, they suggest that these benefits are likely independent of the underlying segmentation process.

## 4. Discussion and conclusions

Our results demonstrate that an end-to-end speech transcription pipeline, such as WhisperX, can benefit from employing energy-based audio segmentation as an alternative to neural network-based VAD. Our proposed segmentation method is language-agnostic and rooted in observed human speech patterns — specifically, Inter-Pausal Units and Tone Units — thus enabling the processing to focus on more natural and coherent audio segments. This approach is particularly effective for disfluent speech, as it circumvents the common pitfalls of VADs, which may prematurely or inappropriately segment audio during disfluencies, ultimately leading to higher omission errors (Goldwater et al., 2010; Liscombe et al., 2021). Our primary motivation for replacing the default VAD of WhisperX arose from the observation that it often introduced omissions by segmenting audio at junctures with no semantic or syntactic significance. In contrast, our segmentation method offers greater control over the audio segmentation process, enabling users to tailor it to the nuances of disfluent speech, such as hesitations, repetitions, and prolonged pauses. The effectiveness of this solution is evidenced by a reduction in deletion errors compared to WhisperX (−11% relative on READLET and -18% relative on CLIPS).

Furthermore, our results highlight the advantages of incorporating additional alternative modules into the WhisperX pipeline for token boundary detection. The principal benefit lies in the improved detection of disfluent speech, as indicated by the higher token-boundary detection accuracy on the READLET corpus (+48% absolute CLMR compared to the off-the-shelf baseline), which frequently contained disfluencies. This enhancement is further corroborated by the analysis of deletions, insertions, and substitutions detailed in the results. The implementation of TU-oriented segmentation prior to audio transcription directly facilitated grammar-based reconstruction. Alongside dual aligner and token-marker post-processing, this approach improved token transcription accuracy, especially in the presence of pauses and repetitions. The results demonstrated the technique's potential, which can be further improved with fine-tuning. The boundaries closely approximated human transcription efforts designed to emphasise disfluent speech, as also demonstrated by the high Collared Matching Rate. This precision in token-boundary prediction is particularly significant for psycholinguistic research (Mandera et al., 2015), which frequently uses this information to investigate language planning in individuals with dyslexia and other speech-production disorders (Cappelli & Noccetti, 2022; Engelhardt et al., 2021; Harm & Seidenberg, 1999; Swets et al., 2021). In particular, since disfluencies are inherently temporal events marked by abnormal pauses, extended syllables, or abrupt restarts, accurate onset and offset markers are indispensable for quantifying their occurrences and durations (Barrett et al., 2024; Mahesh et al., 2025). By reducing systematic biases in boundary estimation, the dual-alignment strategy thus enables a finer-grained correspondence between acoustic and lexical representations by producing boundary annotations that better reflect the utterance's true temporal structure. This correspondence enables more robust identification of disfluent segments, supports the extraction of timing-based biomarkers for speech planning, and enhances the interpretability of ASR output in psycholinguistic and clinical applications. In this sense, boundary detection serves as the bridge between automatic transcription and the temporal analysis of disfluency patterns.

Our findings suggest a potential trade-off between omission (deletion) error rates and insertion error rates when selecting between the default WhisperX pipeline and our proposed pipeline. While deletions generally decrease in our pipeline, insertions correspondingly increase. In particular, IER and DER exhibited a negative correlation (Spearman's $r = -0.65$), confirming that improvements in one are typically accompanied by proportional increases in the other. This phenomenon is likely attributed to the WhisperX ASR internal language model, attempting to complete unfinished words, false starts, and repetitions. This aspect could serve as an indicator of disfluent speech, as it highlights insertions that may correspond to instances of disfluency.

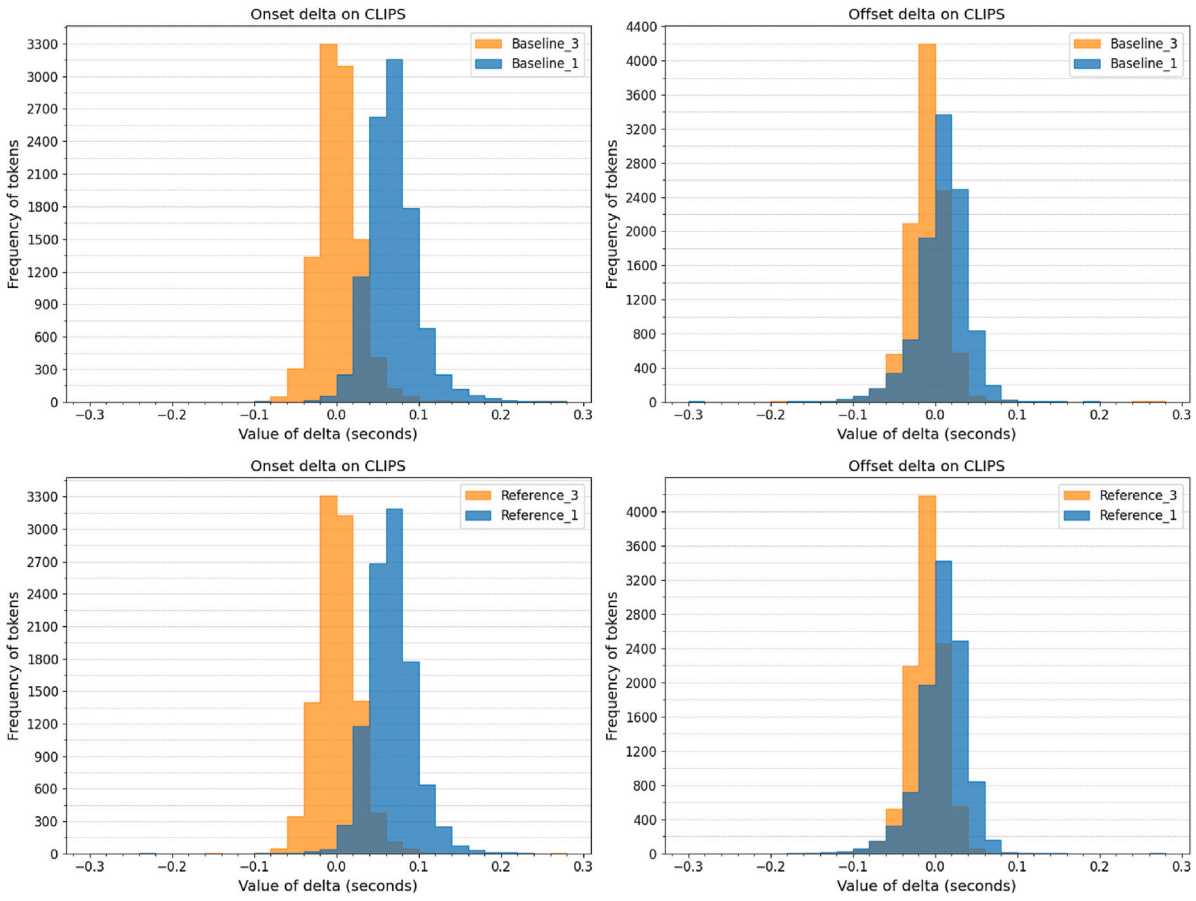


**Fig. 4.** Pairwise comparison between the distributions of the onset and offset deltas of $Baseline_1$, $Baseline_3$, $Reference_1$, and $Reference_3$ on the CLIPS corpus. Each distribution represents the temporal difference (in seconds) between the predicted and manually annotated token boundaries; negative values indicate that the automatic boundary precedes the reference annotation, while positive values indicate a temporal lag. These signed deltas reveal systematic biases in boundary estimation, specifically whether a model tends to anticipate or delay token onsets and offsets.

However, this raises concerns regarding hallucinations in the speech transcription, which may ultimately degrade the performance of the ASR. Consequently, this trade-off should be considered when employing our pipeline for standard transcription tasks. In scenarios where speech transcription quality is prioritised over highlighting disfluency, our pipeline may not be the optimal solution.

Despite the modifications to the off-the-shelf WhisperX pipeline, our approach maintains the modularity of this system while demonstrating that robust transcription (including disfluent speech) can be achieved without fine-tuning the ASR models. Nonetheless, further enhancements in transcription accuracy and token boundary precision could potentially be realised through a specific fine-tuning of the Whisper ASR for disfluent speech. The modular architecture of our pipeline would facilitate substituting the core ASR with a new model tailored to specific use cases.

Finally, although tested on Italian, the proposed segmentation and alignment strategy is language-agnostic, relying on universal acoustic cues such as energy continuity and temporal regularity. Adapting it to other languages would primarily require substituting the ASR model and lexicon, while the overall pipeline and evaluation methodology remain unchanged.

### *4.1. Limitations and future work*

The principal limitations and future enhancements of our methodology can be summarised as follows:

**Dual aligner overhead**: The dual aligner contributes to increased computational overhead, as alignment is performed twice for each audio file. This overhead may constrain the applicability of our pipeline in near-real-time transcription. A potential solution is to parallelise the process across multiple cores, an approach we intend to explore in our future work.

**Benchmarking alternative ASRs**: Although the current phoneme-based Whisper ASR demonstrates superior performance for English, exploring alternative ASRs — such as XLSR (Conneau et al., 2020; Grosman, 2022a), Microsoft Phi-4 (Abdin et al., 2024), and the NVIDIA Transducer and Conformer-based ASRs (Noroozi et al., 2024; Nvidia, 2024) - may yield better results for Italian and disfluent speech (Hosom et al., 2004). We plan to benchmark these alternatives on disfluent speech to identify the optimal ASR for our pipeline.

**Improving pause detection**: The current approach for pause and TU detection is susceptible to errors in environments with elevated noise levels. This requires the user to fine-tune the parameters to optimise the pipeline for noisy audio, including frame duration, window shift, and silence threshold. In our future work, we will investigate adaptive unsupervised methods for high-noise conditions that automatically estimate the silence energy threshold via signal processing (Coro et al., 2023).

**Improving token-marker post-processing**: The automatic detection of the optimal token-marker shift is essential for rendering post-processing adaptable and generalisable across various ASRs. This adaptation should depend on the analysis of the distribution of discrepancies between automatic and manual markers over a reference dataset. Furthermore, the shift should be applied solely to markers exhibiting a higher probability of misalignment. In future work, we will explore modelling this probability based on the contextual acoustic features of each marker.

In summary, we regard the proposed methodology as a potential speech transcription and token boundary detection pipeline that may assist specialists in studying speech patterns, including disfluent reading by individuals with dyslexia. Our token boundary detection, in particular, can potentially enhance dyslexia detection in contemporary multimodal approaches that incorporate token boundary information (Barbiero et al., 2019; Nadalini et al., 2023). For instance, it may be integrated with finger tracking data in multimodal early-diagnosis systems (Marzi et al., 2023; Taxitari et al., 2021). In future investigations, we will examine the benefits of integrating our pipeline with such systems.

## CRediT authorship contribution statement

**Manu Srivastava:** Conceptualisation, Methodology, Software, Resources, Validation, Visualisation, Writing – original draft. **Marcello Ferro:** Data collection, Conceptualisation, Validation, Writing – review & editing. **Vito Pirrelli:** Conceptualisation, Data collection, Validation, Resources, Funding acquisition, Supervision, Writing – review & editing. **Gianpaolo Coro:** Conceptualisation, Software, Validation, Supervision, Writing – review & editing.

## Software

The source code of the audio processing, the transcription post-processing, and all results for reproducibility are available at http://hdl.handle.net/20.500.11752/ILC-1039.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

The authors wish to thank the Phoné Consortium (https://phonegroup.github.io) for providing free access to the CLIPS corpus. We would also like to thank the READLET project, Italy (PRIN 2017 n. 2017W8HFRX) for sharing a subset of the original data used in this study. Special thanks goes to Andrea Nadalini and Alice Todesco, both contributors to the READLET project, for manually annotating and verifying the accuracy of transcribed data.

## Data availability

Data will be made available on request.

## References

Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., Harrison, M., Hewett, R. J., Javaheripi, M., Kauffmann, P., Lee, J. R., Lee, Y. T., Li, Y., Liu, W., Mendes, C. C. T., Nguyen, A., Price, E., Rosa, G. de., Saarikivi, O., .... Zhang, Y. (2024). *Phi-4 technical report*, https://arxiv.org/abs/2412.08905, arXiv:2412.08905.

Akkilic, A. N., Sabir, Z., Bhat, S. A., & Bulut, H. (2024). A radial basis deep neural network process using the bayesian regularization optimization for the monkeypox transmission model. *Expert Systems with Applications*, *235*, Article 121257.

Ananthakrishnan, S., & Narayanan, S. S. (2008). Automatic prosodic event detection using acoustic, lexical, and syntactic evidence. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*, 216–228.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., & Weber, G. (2020). Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th conference on language resources and evaluation*. 4211–4215.

Ash, S., Nevler, N., Irwin, D. J., Shellikeri, S., Rascovsky, K., Shaw, L., Lee, E. B., Trojanowski, J. Q., & Grossman, M. (2023). Apraxia of speech in the spontaneous speech of nonfluent/agrammatic primary progressive aphasia. *Journal of Alzheimer's Disease Reports*, *7*, 589–604.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, *33*, 12449–12460.

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). Whisperx: Time-accurate speech transcription of long-form audio. arXiv preprint arXiv:2303.00747.

Barbiero, C., Montico, M., Lonciari, I., Monasta, L., Penge, R., Vio, C., Tressoldi, P. E., Carrozzi, M., De Petris, A., De Cagno, A. G., et al. (2019). The lost children: The underdiagnosis of dyslexia in italy. a cross-sectional national study. *PloS One*, *14*, Article e0210448.

Barrett, L., Tang, K., & Howell, P. (2024). Comparison of performance of automatic recognizers for stutters in speech trained with event or interval markers. *Frontiers in Psychology*, *15*, Article 1155285.

Betz, S. (2020). *Hesitations in spoken dialogue systems* (Ph.D. thesis), Universität Bielefeld.

Biron, T., Baum, D., Freche, D., Matalon, N., Ehrmann, N., Weinreb, E., Biron, D., & Moses, E. (2021). Automatic detection of prosodic boundaries in spontaneous speech. *PloS One*, *16*, Article e0250969.

Boersma, P. (2011). Praat: doing phonetics by computer [computer program]. http://www.praat.org/.

Bredin, H., & Laurent, A. (2021). End-to-end speaker segmentation for overlap-aware resegmentation. https://arxiv.org/abs/2104.04045, arXiv:2104.04045.

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., & Gill, M. P. (2020). Pyannote.audio: neural building blocks for speaker diarization. In *ICASSP 2020, IEEE international conference on acoustics, speech, and signal processing* (pp. 7124–7128). Barcelona, Spain.

Brugnara, F., Falavigna, D., & Omologo, M. (1993). Automatic segmentation and labeling of speech based on hidden markov models. *Speech Communication*, *12*, 357–370.

Cappelli, G., & Noccetti, S. (2022). *A linguistic approach to the study of dyslexia.*: *vol. 20*, Channel View Publications.

Collard, P. (2009). *Disfluency and listeners' attention: An investigation of the immediate and lasting e ects of hesitations in speech.* (Ph.D. thesis), Psychology collection.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2020). Unsupervised cross-lingual representation learning for speech recognition. https://arxiv.org/abs/2006.13979, arXiv:2006.13979.

Corley, M., & Stewart, O. W. (2008). Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass*, *2*, 589–602.

Coro, G., Bardelli, S., Cuttano, A., & Fossati, N. (2022). Automatic detection of potentially ineffective verbal communication for training through simulation in neonatology. *Education and Information Technologies*, *27*, 9181–9203.

Coro, G., Bardelli, S., Cuttano, A., Scaramuzzo, R. T., & Ciantelli, M. (2023). A self-training automatic infant-cry detector. *Neural Computing and Applications*, *35*, 8543–8559.

Coro, G., Cutugno, F., Caropreso, F., et al. (2007). Speech recognition with factorial-hmm syllabic acoustic models. *INTERSPEECH, International Speech Communication Association*, 870–873.

Coro, G., Cutugno, F., Schettino, L., Tanda, E., Vietti, A., & Vitale, V. N. (2025). Phoné: An initiative to develop a dataset for the automatic recognition of spoken italian. *Oral Archives Journal*, *1*, 89–107. http://dx.doi.org/10.36253/oar-3340, https://riviste.fupress.net/index.php/oarj/article/view/3340.

Coro, G., Massoli, F. V., Origlia, A., & Cutugno, F. (2021). Psycho-acoustics inspired automatic speech recognition. *Computers & Electrical Engineering*, *93*, Article 107238.

Cutugno, F., Coro, G., & Petrillo, M. (2005). Multigranular scale speech recognizers: technological and cognitive view. In *AI* IA 2005: advances in artificial intelligence: 9th congress of the Italian association for artificial intelligence, milan, Italy, September (2005) 21-32. proceedings: vol. 9*, (pp. 327–330). Springer.

Cutugno, F., D'Anna, L., Petrillo, M., & Zovato, E. (2002). Apa: Towards an automatic tool for prosodic analysis. In *Proc. speechProsody 2002* (pp. 231–234).

D'Anna, L., & Cutugno, F. (2003). Segmenting the speech chain into tone units: human behaviour vs automatic process. In *Proceedings of the xVth international congress of phonetic sciences* (pp. 1233–1236).

D'Anna, L., & Petrillo, M. (2003). Sistemi automatici per la segmentazione in unità tonali. In *Atti delle XIII giornate di studio del gruppo di fonetica sperimentale* (pp. 285–290).

Dawson, K., Antonenko, P., Lane, H., & Zhu, J. (2019). Assistive technologies to support students with dyslexia. *Teaching Exceptional Children*, *51*, 226–239.

DBDMG (2022). Wav2vec2-xls-r-1b-italian-robust. https://huggingface.co/dbdmg/wav2vec2-xls-r-1b-italian-robust.

Diwakar, G., & Karjigi, V. (2020). Improving speech to text alignment based on repetition detection for dysarthric speech. *Circuits, Systems, and Signal Processing*, *39*, 5543–5567.

Dovetto, F. M., Guida, A., Pagliaro, A., Guardasci, R., Raggio, L., Sorrentino, A., & Trillocco, S. (2022). Corpora di italiano parlato patologico dell'età adulta e senile. In *Cresti, E., Moneglia, M.(a cura di), Corpora e Studi Linguistici Proceedings of LIV congresso della società di linguistica Italiana* (pp. 165–177). Officinaventuno.

Dutta, S., Tao, S. A., Reyna, J. C., Hacker, R. E., Irvin, D. W., Buzhardt, J. F., & Hansen, J. H. 2022. Challenges remain in building asr for spontaneous preschool children speech in naturalistic educational environments. ISCA iNTERSPEECH-2022. https://par.nsf.gov/biblio/10362772. 10.21437/Interspeech.2022-555.

Engelhardt, P. E., Yuen, M. K., Kenning, E. A., & Filipovic, L. (2021). Are linguistic prediction deficits characteristic of adults with dyslexia? *Brain Sciences*, *11*, 59.

Evangelopoulos, G., & Maragos, P. (2006). Multiband modulation energy tracking for noisy speech detection. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*, 2024–2038.

Fan, R., Chu, W., Chang, P., & Alwan, A. (2023). A ctc alignment-based non-autoregressive transformer for end-to-end automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *31*, 1436–1448.

Ferro, M., Marzi, C., Nadalini, A., Taxitari, L., Lento, A., & Pirrelli, V. (2024). ReadLet: A dataset for oral, visual and tactile text reading data of early and mature readers. In N. Calzolari, M. Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation* (pp. 13595–13609). Torino, Italia: ELRA and ICCL, https://aclanthology.org/2024.lrec-main.1188/.

Fiorin, M., Ugarte, C. V. d, Capellini, S. A., & Oliveira, C. M. C. d. (2015). Oral reading cnd spontaneous speech fluency of students: comparative study between stutterers and non-stutterers. *Revista CEFAC*, *17*, 151–158.

Gala, N., & Ziegler, J. (2016). Reducing lexical complexity as a tool to increase text accessibility for children with dyslexia. In *Proceedings of the workshop on computational linguistics for linguistic complexity* (pp. 59–66).

Gales, M. J., Knill, K. M., Ragni, A., & Rath, S. P. (2014). Speech recognition and keyword spotting for low-resource languages: Babel project research at cued. In *Fourth international workshop on spoken language technologies for under-resourced languages* (pp. 16–23). International Speech Communication Association (ISCA.

Gerganov, G. (2023). Whisper.cpp. GitHub repository. https://github.com/ggerganov/whisper.cpp.

Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in r: the dtw package. *Journal of Statistical Software*, *31*, 1–24.

Gkoumas, D., Wang, B., Tsakalidis, A., Wolters, M., Purver, M., Zubiaga, A., & Liakata, M. (2024). A longitudinal multi-modal dataset for dementia monitoring and diagnosis. *Language Resources and Evaluation*, *58*, 883–902.

Goldwater, S., Jurafsky, D., & Manning, C. D. (2010). Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, *52*, 181–200.

Greenberg, S., & Kingsbury, B. E. (1997). The modulation spectrogram: In pursuit of an invariant representation of speech. In *1997 IEEE international conference on acoustics, speech, and signal processing* (pp. 1647–1650). IEEE.

Grosman, J. (2022a). Fine-tuned XLS-r 1B model for speech recognition in Italian. https://huggingface.co/jonatasgrosman/wav2vec2-xls-r-1b-italian.

Grosman, J. (2022b). Fine-tuned XLS-r 53 model for speech recognition in Italian. https://huggingface.co/jonatasgrosman/exp_w2v2t_it_xlsr-53_s387.

Hamzah, R., & Jamil, N. (2019). Investigation of speech disfluencies classification on different threshold selection techniques using energy feature extraction. *Malaysian Journal of Computing (MJoC)*, *4*, 178–192.

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: insights from connectionist models. *Psychological Review*, *106*, 491.

Hasija, T., Kadyan, V., Guleria, K., Alharbi, A., Alyami, H., & Goyal, N. (2022). Prosodic feature-based discriminatively trained low resource speech recognition system. *Sustainability*, *14*(614).

Hau, D. (2014). *Learning hierarchical speech representations using deep convolutional neural networks* Master's thesis, The University of Manchester (United Kingdom.

Hosom, J. P., Shriberg, L., & Green, J. R. (2004). Diagnostic assessment of childhood apraxia of speech using automatic speech recognition (asr) methods. *Journal of Medical Speech-Language Pathology*, *12*(167).

Hughes, T., & Mierle, K. (2013). Recurrent neural networks for voice activity detection. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 7378–7382). IEEE.

Jian (2023). Stabilizing timestamps for whisper. GitHub repository. https://github.com/jianfch/stable-ts.

Jiang, P. P., Tobin, J., Tomanek, K., MacDonald, R. L., Seaver, K., Cave, R., Ladewig, M., Heywood, R., & Green, J. R. (2024). Learnings from curating a trustworthy, well-annotated, and useful dataset of disordered english speech. arXiv preprint arXiv:2409.09190.

Kane, J., Yanushevskaya, I., Looze, C. De., Vaughan, B., & Chasaide, A. N. (2014). Analysing the prosodic characteristics of speech-chunks preceding silences in task-based interactions. In *INTERSPEECH* (pp. 333–337).

Kearns, D. M., & Whaley, V. M. (2019). Helping students with dyslexia read long words: Using syllables and morphemes. *Teaching Exceptional Children*, *51*, 212–225.

Kingsbury, B. E., Morgan, N., & Greenberg, S. (1998). Robust speech recognition using the modulation spectrogram. *Speech Communication*, *25*, 117–132.

Klein, G. (2023). Faster whisper transcription with ctranslate2. GitHub repository. https://github.com/guillaumekln/faster-whisper.

Knowles, T., Clayards, M., Sonderegger, M., Wagner, M., Nadig, A., & Onishi, K. H. (2015). Automatic forced alignment on child speech: Directions for improvement. In *Proceedings of meetings on acoustics, acoustical society of america* (p. 060001).

Kocharov, D., Kachkovskaia, T., & Skrelin, P. (2019). Prosodic boundary detection using syntactic and acoustic information. *Computer Speech & Language*, *53*, 231–241.

Koenecke, A., Choi, A. S. G., Mei, K. X., Schellmann, H., & Sloane, M. (2024). Careless whisper: Speech-to-text hallucination harms. In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency* (pp. 1672–1681).

Kolář, J., & Liu, Y (2010). Automatic sentence boundary detection in conversational speech: A cross-lingual evaluation on english and czech. In *2010 IEEE international conference on acoustics, speech and signal processing* (pp. 5258–5261). IEEE.

Kumari, R., Dev, A., & Kumar, A. (2022). An efficient syllable-based speech segmentation model using fuzzy and threshold-based boundary detection. *International Journal of Computational Intelligence and Applications*, *21*, Article 2250007.

Kürzinger, L., Winkelbauer, D., Li, L., Watzel, T., & Rigoll, G. (2020). Ctc-segmentation of large corpora for german end-to-end speech recognition. In *International conference on speech and computer* (pp. 267–278). Springer.

Lea, C., Huang, Z., Narain, J., Tooley, L., Yee, D., Tran, D. T., Georgiou, P., Bigham, J. P., & Findlater, L. (2023). From user perceptions to technical improvement: Enabling people who stutter to better use speech recognition. In *Proceedings of the 2023 CHI conference on human factors in computing systems* (pp. 1–16).

Leoni, F. A. (2014). CLIPS : corpora e lessici di italiano parlato e scritto. In *LINDAT/CLARIAH-CZ digital library at the institute of formal and applied linguistics (úFAL)*. Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11372/LRT-865.

Leoni, F. A., Sobrero, A. A., & Paoloni, A. (2007). Corpora e lessici di italiano parlato e scritto (clips). *Bollettino Di Italianistica*, *4*, 0–121.

Lindström, A., Villing, J., Larsson, S., Seward, A., & Åberg, C. (2008). The effect of cognitive load on disfluencies during in-vehicle spoken dialogue. In *INTERSPEECH* (pp. 1196–1199).

Liscombe, J., Kothare, H., Neumann, M., Ocampo, A., Roesler, O., Habberstad, D., Cornish, A., Pautler, D., Suendermann-Oeft, D., & Ramanarayanan, V. (2021). Voice activity detection considerations in a dialog agent for dysarthric speakers. In *International workshop on spoken dialog systems* (pp. 1–15).

Liu, Y., Chawla, N. V., Harper, M. P., Shriberg, E., & Stolcke, A. (2006). A study in machine learning from imbalanced data for sentence boundary detection in speech. *Computer Speech & Language*, *20*, 468–494.

Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M., & Harper, M. (2006). Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*, 1526–1540.

Liu, J., Wumaier, A., Wei, D., & Guo, S. (2023). Automatic speech disfluency detection using wav2vec2. 0 for different languages with variable lengths. *Applied Sciences*, *13*, 7579.

Louradour, J. (2023). Whisper-timestamped. https://github.com/linto-ai/whisper-timestamped.

Ludusan, B., Origlia, A., Cutugno, F., et al. (2011). On the use of the rhythmogram for automatic syllabic prominence detection. In *INTERSPEECH* (pp. 2413–2416).

Ludusan, B., Ziegler, S., & Gravier, G. (2014). Is syllable stress information robust for asr in adverse conditions? In *International conference on speech prosody* (pp. 939–943).

Mahesh, S., Manasa, A., Sreedevi, N., & Veda, P. (2025). Acoustic analysis of speech of children who stutter. *Journal of All India Institute of Speech and Hearing*, 10–4103.

Mandera, P., Keuleers, E., & Brysbaert, M. (2015). How useful are corpus-based methods for extrapolating psycholinguistic variables? *Quarterly Journal of Experimental Psychology*, *68*, 1623–1642.

Marzi, C., Melloni, C., & Vender, M. (2023). Finger-tracking reading profiles in monolingual and bilingual early graders. *Lingue E Linguaggio*, *22*, 327–361.

Marzi, C., Narzisi, A., Milone, A., Masi, G., & Pirrelli, V. (2022). Reading behaviors through patterns of finger-tracking in italian children with autism spectrum disorder. *Brain Sciences*, *12*(1316).

Marzinzik, M., & Kollmeier, B. (2002). Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. *IEEE Transactions on Speech and Audio Processing*, *10*, 109–118.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech* (pp. 498–502).

Nadalini, A., Marzi, C., Ferro, M., Taxitari, L., Lento, A., Crepaldi, D., & Pirrelli, V. (2023). Eye-voice and finger-voice spans in adults' oral reading of connected texts: Implications for reading research and assessment. *The Mental Lexicon*, *18*, 366–400.

Nahar, R., & Kai, A. (2020). Effect of data augmentation on dnn-based vad for automatic speech recognition in noisy environment. In *2020 IEEE 9th global conference on consumer electronics* (pp. 368–372). IEEE.

Noroozi, V., Majumdar, S., Kumar, A., Balam, J., & Ginsburg, B. (2024). Stateful conformer with cache-based inference for streaming automatic speech recognition. In *ICASSP 2024-2024 IEEE international conference on acoustics, speech and signal processing* (pp. 12041–12045). IEEE.

Nvidia (2024). NVIDIA FastConformer-CTC. https://huggingface.co/nvidia/stt_en_fastconformer_ctc_large.

Pakhomov, S. V., Marino, S. E., & Birnbaum, A. K. (2013). Quantification of speech disfluency as a marker of medication-induced cognitive impairment: An application of computerized speech analysis in neuropharmacology. *Computer Speech & Language*, *27*, 116–134.

Patil, R. M., & Patil, C. (2024). Unveiling the state-of-the-art: A comprehensive survey on voice activity detection techniques. In *2024 Asia Pacific conference on innovation in technology* (pp. 1–5). IEEE.

Pedersen, J. S., & Larsen, L. B. (2010). A speech corpus for dyslexic reading training. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (Eds.), *Proceedings of the seventh international conference on language resources and evaluation* (pp. 2820–2823). Valletta, Malta: European Language Resources Association (ELRA, https://aclanthology.org/L10-1025/.

Prakash, A., & Murthy, H. A. (2024). Exploring an inter-pausal unit (ipu) based approach for indic end-to-end tts systems. arXiv preprint arXiv:2409.11915.

Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., & Collobert, R. (2020). Mls: A large-scale multilingual dataset for speech research. ArXiv abs/2012.03411.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *International conference on machine learning* (pp. 28492–28518). PMLR.

Ramlan, S., Isa, I., Harron, N., Saod, A., Azid, M., & Lepas, B. (2023). Reading assistive tool (readys) for dyslexic children: Speech recognition performance. *J. Creat. Pract. Lang. Learn. Teach.(CPLT)*, *11*, 57–73.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., et al. (2021). Speechbrain: A general-purpose speech toolkit. arXiv preprint arXiv:2106.04624.

Rehbein, I., Ruppenhofer, J., & Schmidt, T. (2020). Improving sentence boundary detection for spoken language transcripts. In *Proceedings of the twelfth language resources and evaluation conference* (pp. 7102–7111).

Romana, A. (2024). *Transforming disfluency detection: integrating large language and acoustic models* (Ph.D. thesis), University of Michigan.

Romana, A., Koishida, K., & Provost, E. M. (2024). Automatic disfluency detection from untranscribed speech. In *IEEE/ACM transactions on audio, speech, and language processing*.

Sabir, Z., Abdelkawy, M., Baghdady, A., & Berro, B. (2025). A dual-layered neural network for the cancer system with stem cells and chemotherapy. *The European Physical Journal Plus*, *140*(885).

Sabir, Z., Ali, M. R., Raja, M. A. Z., Shoaib, M., & Núñez, R. (2022). Computational intelligence approach using levenberg–marquardt backpropagation neural networks to solve the fourth-order nonlinear system of emden–fowler model. *Engineering with Computers*, *38*, 2975–2991.

Sabir, Z., Assaad, A. A., Alkak, A., & Bayram, M. (2025). A radial basis bayesian regularization procedure for the lassa virus mathematical model. *The European Physical Journal Plus*, *140*(938).

Sabir, Z., Said, S. B., & Al-Mdallal, Q. (2023). A fractional order numerical study for the influenza disease mathematical model. *Alexandria Engineering Journal*, *65*, 615–626.

Sabir, Z., Wahab, H. A., Javeed, S., & Baskonus, H. M. (2021). An efficient stochastic numerical computing framework for the nonlinear higher order singular models. *Fractal and Fractional*, *5*(176).

Sabu, K., Vaidya, M., & Rao, P. (2021). Cnn encoding of acoustic parameters for prominence detection. arXiv preprint arXiv:2104.05488.

Salesky, E., Wiesner, M., Bremerman, J., Cattoni, R., Negri, M., Turchi, M., Oard, D. W., & Post, M. (2021). The multilingual tedx corpus for speech recognition and translation. https://arxiv.org/abs/2102.01757, arXiv:2102.01757.

Sanchez, Y. G., Umar, M., Sabir, Z., Guirao, J., & Raja, M. A. Z. (2018). Solving a class of biological hiv infection model of latently infected cells using heuristic approach. *Discrete. Contin. Dyn. Syst., S*, 14.

Segbroeck, M. Van., Travadi, R., Vaz, C., Kim, J., Black, M. P., Potamianos, A., & Narayanan, S. S. (2014). Classification of cognitive load from speech using an i-vector framework. In *Interspeech* (pp. 751–755). Citeseer.

Sheikh, S. A., & Kodrasi, I. (2024). Impact of speech mode in automatic pathological speech detection. In *2024 32nd European signal processing conference* (pp. 81–85). IEEE.

Shezi, N., & Reddy, S. (2020). Word boundary estimation of isizulu continuous speech. In *2020 international conference on power, instrumentation, control and computing* (pp. 1–6). IEEE.

Shriberg, E. (2001). To 'errrr'is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, *31*, 153–169.

Silero Team (2024). Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. https://github.com/snakers4/silero-vad.

Siohan, O. (2017). Ctc training of multi-phone acoustic models for speech recognition. In *INTERSPEECH* (pp. 709–713).

Southwell, R., Pugh, S., Perkoff, M., Clevenger, C., Bush, J., Lieber, R., Ward, W., Foltz, P., & D'Mello, S. (2022). Challenges and feasibility of automatic speech recognition for modeling student collaborative discourse in classrooms. *International Educational Data Mining Society*.

Stehwien, S., Schweitzer, A., & Vu, N. T. (2020). Acoustic and temporal representations in convolutional neural network models of prosodic events. *Speech Communication*, *125*, 128–141.

Stolcke, A., Shriberg, E., Bates, R. A., Ostendorf, M., Hakkani, D. Z., Plauche, M., Tür, G., & Lu, Y. (1998). Automatic detection of sentence boundaries and disfluencies based on recognized words. In *ICSLP* (pp. 2247–2250). Citeseer.

Swets, B., Fuchs, S., Krivokapić, J., & Petrone, C. (2021). A cross-linguistic study of individual differences in speech planning. *Frontiers in Psychology*, *12*, Article 655516.

Systran (2025). Fasterwhisper asr. https://huggingface.co/Systran/faster-whisper-large-v3.

Taxitari, L., Cappa, C., Ferro, M., Marzi, C., Nadalini, A., & Pirrelli, V. (2021). Using mobile technology for reading assessment. In *2020 6th IEEE congress on information science and technology* (pp. 302–307). IEEE.

Treviso, M., Shulby, C., & Aluísio, S. (2017). Evaluating word embeddings for sentence boundary detection in speech transcripts. In *Proceedings of the 11th Brazilian symposium in information and human language technology* (pp. 151–160).

Università di Napoli Federico II (2024). CLIPS corpus. Website. http://www.clips.unina.it/home.

Valk, J., & Alumäe, T. (2021). Voxlingua107: a dataset for spoken language recognition. In *2021 IEEE spoken language technology workshop* (pp. 652–658). IEEE.

Vitale, V. N., Cutugno, F., Origlia, A., & Coro, G. (2024). Exploring emergent syllables in end-to-end automatic speech recognizers through model explainability technique. *Neural Computing and Applications*, *36*, 6875–6901.

Wang, C., Riviere, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., & Dupoux, E. (2021). VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)* (pp. 993–1003). Association for Computational Linguistics, Online., http://dx.doi.org/10.18653/v1/2021.acl-long.80, https://aclanthology.org/2021.acl-long.80/.

Weise, T., Demir, K. C., Pérez-Toro, P. A., Arias-Vergara, T., Maier, A., Nöth, E., Schuster, M., Heismann, B., & Yang, S. H. (2025). Towards end-to-end speech articulation and spoken language analysis using deep learning. *Human-Centric Intelligent Systems*, 1–20.

Widiaputri, R., Purwarianti, A., Lestari, D. P., Azizah, K., Tanaya, D., & Sakti, S. (2023). Speech recognition and meaning interpretation: Towards disambiguation of structurally ambiguous spoken utterances in indonesian. In *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 16813–16824).

Wu, S. L., Kingsbury, E., Morgan, N., & Greenberg, S. (1998). Incorporating information from syllable-length time scales into automatic speech recognition. In *Proceedings of the 1998 IEEE international conference on acoustics, speech and signal processing, iCASSP'98 (cat. no. 98CH36181)* (pp. 721–724). IEEE.

Yamasaki, H., Louradour, J., Hunter, J., & Prévot, L. (2023). Transcribing and aligning conversational speech: A hybrid pipeline applied to french conversations. In *2023 IEEE automatic speech recognition and understanding workshop* (pp. 1–6). IEEE.

Yang, L., Achard, C., & Pelachaud, C. (2022). Multimodal analysis of interruptions. In *International conference on human-computer interaction* (pp. 306–325). Springer.

Yildirim, S., & Narayanan, S. (2009). Automatic detection of disfluency boundaries in spontaneous speech of children using audio–visual information. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*, 2–12.

Zhang, X. L., & Wang, D. (2015). Boosting contextual information for deep neural network based voice activity detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*, 252–264.

Zhou, X., Kashyap, A., Li, S., Sharma, A., Morin, B., Baquirin, D., Vonk, J., Ezzes, Z., Miller, Z., Tempini, M. L. G., et al. (2024). Yolo-stutter: End-to-end region-wise speech dysfluency detection. arXiv preprint arXiv:2408.15297.