

Article

Not peer-reviewed version

Speech-Adaptive Detection of Unnatural Intra-Sentential Pauses Using Contextual Anomaly Modeling for Interpreter Training

[Hyoeun Kang](#)[†], [Jin-Dong Kim](#)[†], [Juriae Lee](#)[†], [Hee-Jo Nam](#), [Kon Woo Kim](#), [Joowon Lim](#), [Hyun-Seok Park](#)^{*}

Posted Date: 18 March 2026

doi: 10.20944/preprints202603.1454.v1

Keywords: interpreting fluency; unnatural pause detection; TalkTrack; isolation forest; sliding window; intra-sentential pause; speech-rate adaptive detection; Computer-Assisted Interpreter Training (CAIT)



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Speech-Adaptive Detection of Unnatural Intra-Sentential Pauses Using Contextual Anomaly Modeling for Interpreter Training

Hyoeun Kang ^{1,†}, Jin-Dong Kim ^{2,†}, Juriae Lee ^{3,†}, Hee-Jo Nam ⁴, Kon Woo Kim ⁵, Joowon Lim ⁶ and Hyun-Seok Park ^{4,*}

- ¹ Department of Statistics, Ewha Womans University, Seoul 03760, Republic of Korea
- ² Database Center for Life Science, Research Organization of Information and Systems, 178-4-4 Wakashiba, Kashiwa, Chiba 277-0871, Japan
- ³ Graduate School of Translation and Interpretation, Ewha Womans University, Seoul 03760, Republic of Korea
- ⁴ School of Artificial Intelligence, Ewha Womans University, Seoul, Republic of Korea
- ⁵ National Institute of Informatics, 2 Chome-1-2 Hitotsubashi, Chiyoda City, Tokyo 101-8430, Japan
- ⁶ Department of Computer Science and Information Engineering, The Catholic University of Korea, Gyeonggi-do 14662, Republic of Korea
- * Correspondence: neo@ewha.ac.kr
- † These authors contributed equally to this work.

Abstract

Detecting unnatural pauses is a critical component of automated quality assessment (AQA) in interpreter training, as pause patterns directly reflect an interpreter’s cognitive load and fluency. Traditional pause detection methods rely on static temporal thresholds (e.g., 1.0 second), which often fail to account for segment-specific speech rate variability and individual speaking styles. This study proposes a context-adaptive pause detection framework that integrates unsupervised anomaly detection using Isolation Forest (iForest) with a sliding window technique. To enhance pedagogical validity, we specifically focused on intra-sentential pauses by delineating sentence boundaries using a specialized segmentation model. The proposed model was evaluated against ground-truth labels annotated by professional interpreting experts. Our results demonstrate that the sliding window-based contextual anomaly detection model significantly outperforms the conventional static baseline, particularly in terms of recall and Cohen’s kappa. Furthermore, by applying a weighted F3-score and the “Recognition-over-Recall” principle, we confirmed that the proposed model substantially reduces the instructor’s total operational burden by shifting the workload from de novo annotation creation to more efficient corrective pruning. These findings suggest that speech-adaptive modeling provides a more reliable and labor-saving framework for automated interpreting assessment and feedback.

Keywords: interpreting fluency; unnatural pause detection; TalkTrack; isolation forest; sliding window; intra-sentential pause; speech-rate adaptive detection; Computer-Assisted Interpreter Training (CAIT)

1. Introduction

Fluency is a fundamental dimension of interpreting quality, reflecting the extent to which an interpreted message is delivered smoothly, naturally, and with minimal disruption. It directly influences listener comprehension, perceived professionalism, and overall communicative effectiveness. In interpreter training, fluency is therefore regarded as a critical performance indicator and is frequently assessed alongside accuracy and completeness [1–3].

Previous research has evaluated interpreting fluency using a range of observable indicators, including silent pauses, filled pauses, backtracking, prosodic variation, and speech rate [4–6]. Among these factors, silent pauses have attracted particular attention because they provide measurable evidence of cognitive processing difficulty and production disruption. Excessive or poorly positioned pauses may interrupt speech continuity, increase listener processing effort, and weaken discourse coherence, thereby negatively affecting perceived interpreting quality [7,8]. Therefore, accurately detecting and analyzing these pauses is an essential task in establishing an Automated Quality Assessment (AQA) system for interpreting [9,10].

A common approach to identifying unnatural pauses has been to apply fixed temporal thresholds, whereby pauses exceeding a predefined duration are treated as indicators of disfluency. However, pause duration alone does not reliably reflect fluency. Longer pauses may serve linguistically meaningful functions such as syntactic boundary marking, discourse planning, or strategic reformulation. Consequently, threshold-based detection approaches risk overestimating disfluency in slower speech while under-detecting problematic pauses in faster speech contexts. This limitation highlights the need for pause detection methods that incorporate contextual and speech-rate-dependent characteristics.

To support automated fluency assessment in interpreter training, our research team developed TalkTrack, an interpreting feedback system that detects pause patterns and provides learners with visual fluency analytics [11]. TalkTrack incorporates the TexAE annotation system [12]. The system enables trainees to monitor pause frequency and duration through graphical feedback, facilitating self-reflection and targeted fluency improvement. Nevertheless, the current implementation relies primarily on duration-based criteria, which may inaccurately classify cognitively appropriate pauses as disfluencies. This is because TalkTrack relies on previous studies that define a pause as one second or longer [13,14]. Such overgeneralization can impose unnecessary cognitive pressure on trainees and distort their perception of effective interpreting strategies.

Another important consideration is the distinction between inter-sentential and intra-sentential pauses. Inter-sentential pauses often reflect natural segmentation and breathing patterns, whereas intra-sentential pauses are more likely to signal disruptions in speech planning or lexical retrieval. Treating these pause types uniformly may therefore reduce detection validity and pedagogical usefulness. A more refined detection framework should focus on identifying abnormal intra-sentential pauses while accounting for variability in local speech dynamics. Previous research has highlighted that grammatical pauses at sentence or syntactic boundaries facilitate meaning transmission and are not necessarily indicative of disfluency [15,16].

Building on these insights, this study proposes a context-sensitive pause detection model that incorporates the dynamic characteristics of speech. Specifically, the model adjusts detection criteria based on speech rate and acoustic properties, integrating statistical techniques with machine learning methods such as Isolation Forest and sliding window-based anomaly detection [17]. Experimental results demonstrate that the proposed model outperforms conventional static approaches in terms of F1-score and agreement with expert annotations, thereby enhancing the reliability of AQA systems and providing interpreters with more accurate, actionable feedback.

The proposed framework is implemented within the TalkTrack platform and evaluated using expert-annotated interpreting data. By improving agreement with expert judgments and enhancing the interpretability of automated feedback, the study aims to strengthen the pedagogical reliability of automated quality assessment in interpreter training. Ultimately, this research contributes to the development of context-aware fluency assessment methods that move beyond static pause definitions and provide more accurate, actionable feedback for interpreting learners.

2. Literature Review

This section reviews prior research on interpreting fluency and pause analysis, focusing on limitations of threshold-based detection approaches and recent advances in automated, context-aware pause modeling relevant to the proposed framework.

2.1. *The Role of Pauses in Interpreting Fluency and the Evolution of Technical Analysis*

In the field of interpreting education and assessment, pauses are widely recognized as a cardinal indicator of fluency and a crucial variable reflecting an interpreter's proficiency, cognitive load, and overall interpreting quality [8,18,19]. Research on interpreting pauses has evolved in tandem with technological advancements. Early studies focused on systematically classifying the pause patterns of professional interpreters, demonstrating that both pause frequency and duration differ significantly based on proficiency levels; this established pauses as a primary metric for fluency [20,21]. Subsequently, in the mid-2000s, Ahrens [22] observed in English-German simultaneous interpreting that the duration, frequency, and distribution of pauses reflect the specific prosodic and speaking styles of individual interpreters.

Since 2015, research has increasingly conceptualized pauses as by-products of cognitive processing rather than mere acoustic phenomena, shifting the focus toward quantitative modeling and automated evaluation. Wang and Li [8] demonstrated through Chinese-English simultaneous interpreting research that pauses in the target speech are typically longer and less frequent than those in the source text, exhibiting a hierarchical distribution governed by syntactic complexity. Furthermore, Christodoulides [23] identified distinct pause patterns across different interpreting modes, noting that interpreters tend to produce fewer but longer pauses compared to native speakers, alongside greater variability in speech rate.

2.2. *Debates on Pause Thresholds and the Refinement of Quantitative Methodologies*

Establishing an objective threshold to define pauses has remained a subject of ongoing debate within interpreting studies. While the literature reflects a broad spectrum of criteria—ranging from 0.2 to 3 seconds [24]—recent scholarly efforts have sought to resolve this inconsistency through empirical validation. Han and An [7], for instance, evaluated 16 distinct thresholds (from 0.25 to 1.0 seconds) using English-to-Chinese consecutive interpreting data. Their findings revealed that higher thresholds tended to show weaker correlations with subjective fluency ratings, whereas the speech-to-pause ratio exhibited the strongest alignment with human evaluations.

However, while thresholds can be set as low as 0.25 or 0.3 seconds, these intervals are often hardly perceptible to listeners as actual pauses during live interpreting. Consequently, a significant body of research adopts a 1.0-second threshold to identify cognitively and perceptually salient pauses. For instance, Choi [13] employed a 1.0-second threshold in a study of graduate translation and interpreting students, reporting an average pause length of 1.9 seconds. Similarly, Lee [14] utilized the same 1.0-second criterion for Japanese-to-Korean simultaneous interpreting. Empirical data from other studies further support this range; Wang and Li [8] observed mean pause lengths of 1.06 seconds for professionals and 1.15 seconds for trainees, while Shreve et al. [25] identified 1.0 second as a critical threshold in sight translation, with observed pauses ranging from 1.021 to 11 seconds.

Building upon this discourse, Wang and Wang [10] enhanced statistical rigor by adopting a median threshold (0.25 seconds) to mitigate the impact of outliers. They further introduced a more nuanced taxonomy of temporal metrics, employing the interquartile range (IQR) to differentiate between 'relatively long pauses' (NRLUP) and 'particularly long pauses' (NPLUP). Their comprehensive set of temporal parameters includes: (1) number of silent pauses (NUP), (2) mean length of silent pauses (MLUP), (3) number of relatively long silent pauses (NRLUP), (4) number of particularly long silent pauses (NPLUP), (5) number of filled pauses (NFP), (6) mean length of filled pauses (MLFP), (7) number of relatively slow articulations (NRSA), (8) number of particularly slow articulations (NPSA), (9) number of relatively quick articulations (NRQA), and (10) number of particularly quick articulations (NPQA).

Significantly, these outlier-based parameters demonstrate superior explanatory power regarding subjective fluency assessments compared to traditional, simplistic indicators. Regression analyses indicated that models incorporating these refined parameters achieved substantial

explanatory power, ranging from 48% to 83.6%, whereas models relying solely on conventional metrics performed notably lower [10].

2.3. Cognitive Load and Pause Patterns

Recent studies interpret pauses not as mere vocal breaks but as dynamic signals reflecting an interpreter's real-time cognitive processing. Zhang and Jing [19], grounded in the Information Packaging Hypothesis (IPH), demonstrated that restricting gesture use significantly increases the frequency and duration of disfluencies among interpreting trainees. Specifically, when gestures were restricted, the frequency of disfluencies rose significantly from 3.92 to 5.22 compared to the unrestricted condition ($F=8.030$, $p=0.006$), while the duration of disfluencies increased from 12.31 to 16.9 seconds ($F=7.779$, $p=0.007$).

These findings suggest that higher cognitive effort is required to reorganize information into spoken language when multimodal support (i.e., gestures) is absent, leading to more frequent self-corrections, repetitions, and prolonged disfluencies. Statistically, it has been confirmed that pause patterns are highly sensitive to fluctuations in cognitive load, particularly in tasks involving spatial or complex content.

2.4. Limitations of Fixed Thresholds

The fixed-threshold approach adopted by most conventional studies faces several fundamental limitations. First, it fails to adequately account for individual speaking styles or linguistic idiosyncrasies [26]. Research by Toivola et al. [26] indicates that native Finnish speakers may produce longer pauses more frequently than non-native speakers, suggesting that interpreting experience, personal speech patterns, and L1 characteristics are critical variables. Second, static models often overlook the prosodic-syntactic interface, such as the syntactic position of a pause or the acoustic properties of preceding vowels [27,28].

To overcome these limitations, recent research has shifted toward speech rate-adaptive thresholding [3,10] and data-driven modeling approaches. In broader speech processing research, anomaly detection techniques and distribution-based outlier modeling have been widely applied to detect rare temporal events in non-stationary signals [3,8,10,29]. However, their application to interpreting fluency assessment remains largely unexplored.

Unlike prior studies that rely on predefined pause duration thresholds, this study adopts an unsupervised anomaly detection framework that models pause events relative to both global and local distributional characteristics of the speech stream. This approach enables the detection of contextually abnormal pauses rather than merely long silences.

3. TalkTrack System

This section describes the methodological framework for context-sensitive pause detection, covering silence extraction, speech-rate normalization, and sliding window-based anomaly detection.

3.1. TalkTrack Platform Architecture

The interpreting audio data analyzed in this study was collected via TalkTrack, an automated feedback system for interpreting training developed by the research team. TalkTrack is a Learning Management System (LMS)-based platform specifically designed to satisfy the pedagogical requirements of Computer-Assisted Interpreter Training (CAIT). The platform integrates task creation, execution, and feedback processes into a unified ecosystem (see Appendix A).

In traditional interpreting pedagogy, tasks and feedback are often fragmented across multiple tools, which complicates the systematic recording and analysis of learner performance. Furthermore, fluency assessment has historically relied on the subjective, experiential judgments of instructors, leading to ongoing concerns regarding the objectivity and consistency of feedback.

TalkTrack was developed to address these structural limitations. By integrating Speech-to-Text (STT) technology and automatic annotation features within an LMS specialized for interpreting, the system automates the collection of performance data and systematizes fluency-related feedback. This design aims to alleviate the evaluative burden on instructors while providing trainees with a data-driven environment for self-monitoring and reflection.

Technically, TalkTrack is a web-based system accessible across both PC and mobile platforms. Its architecture consists of a web server for user authentication and task management, a database for storing audio files and analysis results, and a specialized analysis module for speech processing and automatic annotation. When an instructor assigns a task, learners can listen to the source audio and record their interpretation directly within the system, ensuring a seamless workflow without the need for external software(Figure 1).

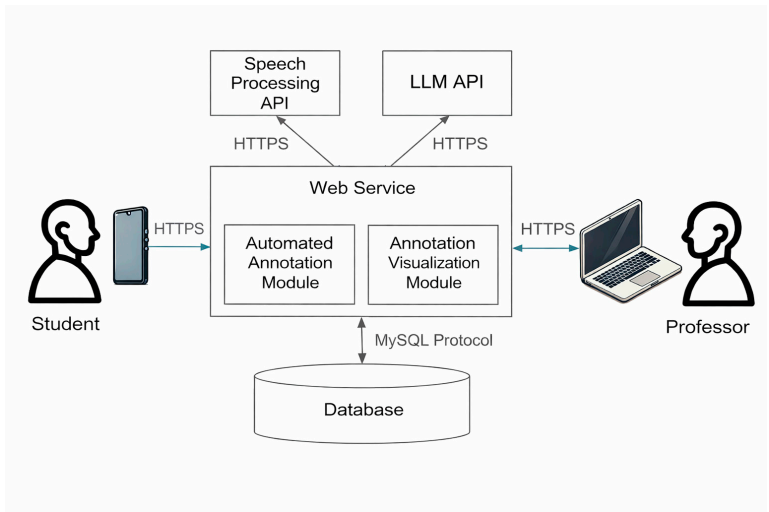


Figure 1. Overall architecture of the TalkTrack-based pause detection system. The framework consists of speech input acquisition, forced alignment-based silence extraction, feature normalization, anomaly detection, and feedback visualization modules integrated within the LMS environment (adapted from Lee et al. [11]).

3.2. Automatic Annotation and Fluency Feedback Functionality

A core feature of TalkTrack is the automatic annotation of disfluency markers within the interpreting output. The system detects and visually highlights features such as pauses, filler words, and backtracking within the transcribed text, enabling an intuitive review of the learner’s speech characteristics. Among these elements, pauses are identified based on silent intervals exceeding a predefined duration. Following conventions in existing literature, the current iteration of the system defines a pause as a silence lasting one second or longer.

All interpreting tasks conducted within TalkTrack are archived alongside their corresponding audio recordings, transcripts, annotations, and feedback. Automatic annotation results are presented via a text-based interface, allowing instructors to manually refine annotations or append qualitative, content-related feedback. As illustrated in [Figure 2], performance annotations are automatically labeled as ‘PAUSE’ (Unfilled pause), and ‘FILLER’ (Filled pause).

The interpretation audio submitted by learners is asynchronously converted to text via API integration, ensuring that the analysis process does not interrupt the user workflow. These generated transcripts and annotation results are stored in a structured database, providing a robust foundation for both immediate pedagogical feedback and long-term statistical analysis.

A distinctive advantage of TalkTrack is its capacity to decouple fluency feedback from content-based evaluation, thereby enabling an independent assessment of delivery-related performance and semantic accuracy. These feedback results are statistically aggregated by category and presented through graphical visualizations (Figure 3), allowing trainees to intuitively recognize their speech patterns and performance trends.

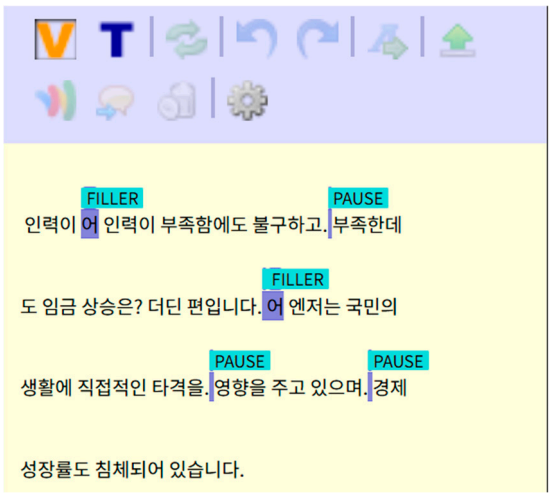


Figure 2. Visual interface of the automatic pause annotation functionality in TalkTrack. Silence segments extracted via forced alignment are automatically labeled and visualized, enabling expert verification and efficient construction of annotated training data.

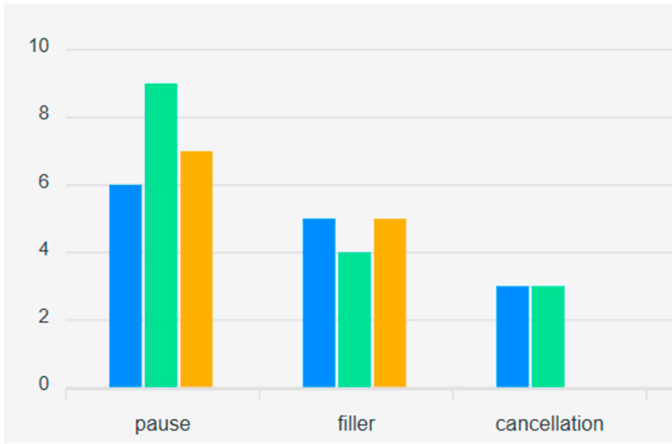


Figure 3. Statistical feedback dashboard presenting graphical fluency metrics. The dashboard visualizes the frequency of unnatural pauses, fillers, and cancellations as metrics of interpretative disfluency across multiple iterations. In this example, data from three attempts are displayed, allowing learners to track the progression of their speech patterns and pinpoint specific fluency deficits.

Figure 3 illustrates the quantitative results for disfluency markers, including pauses, fillers, and cancellations (backtracking or false starts). The stacked bar chart uses blue to represent the 1st assignment, green for the 2nd assignment, and yellow for the 3rd assignment, which allows trainees to track changes in each performance metric over time. Through these cumulative statistics, trainees can monitor their developmental progress over time. Furthermore, Figure 3 shows that no cancellations appeared in the 3rd assignment. This longitudinal tracking capability is invaluable not only for cross-sectional assessments of individual performance but also for analyzing the evolution of a learner’s fluency throughout the training period. Consequently, the system serves as a robust tool for empirical data collection, enabling a granular analysis of the frequency and distribution of fluency-related deficiencies.

However, despite these advantages, the current configuration of TalkTrack possesses a critical limitation: it employs a uniform threshold for pause detection and aggregation. Currently, any silent interval exceeding one second is categorized as a pause. When applied indiscriminately, this criterion conflates natural, functional pauses—which are inherent to fluent speech—with those that genuinely indicate a loss of fluency. In particular, silences occurring at semantic transitions or grammatical

boundaries are statistically aggregated alongside disfluent pauses, potentially skewing the overall assessment.

From a pedagogical perspective, such undifferentiated data can mislead trainees, obscuring the distinction between pauses that are communicative and those that require remedial attention. To enhance the precision of fluency feedback, it is imperative to move beyond a singular definition of pauses. Accordingly, this study proposes a refined approach that distinguishes between intra-sentential and inter-sentential pauses, rather than treating all silent intervals as homogeneous errors.

4. Methodology

This section presents the experimental design and evaluation results of the proposed pause detection framework, including dataset configuration, annotation reliability, and comparative performance analysis.

4.1. Data Collection

The dataset for this study, comprises Japanese–Korean interpreting data collected during the development and pilot testing of the TalkTrack platform (see Figure 4)



Figure 4. User interface of the TalkTrack interpreting training platform. The interface facilitates the recording and submission of interpretations by learners, followed by automatic annotation and manual corrections by instructors. By integrating these visual components, the system supports iterative interpreting practice and enables fluency self-assessment.

In this project, for the purpose of analyzing interpreting performance, a total of 532 video data files were collected, including 390 cases of English ST (Source Text) with TT (Target Text) in Korean, 74 cases of Chinese ST, 56 cases of Japanese ST, and 2 cases of Russian ST, all from publicly available international conference interpreting videos on YouTube. Additionally, 38 cases of performance data from 12 graduate students in interpreting, collected via TalkTrack, were also gathered, covering Chinese-Korean consecutive interpreting, Chinese-Korean simultaneous interpreting, Japanese-Korean consecutive interpreting, Korean-Japanese consecutive interpreting, Japanese-Korean simultaneous interpreting, and Korean-Japanese simultaneous interpreting. Among these, this study focuses on pause measurement using the Japanese-to-Korean simultaneous interpreting data from 7 students collected to enhance the performance of TalkTrack.

To examine the usability of TalkTrack, seven graduate students from a Graduate School of Translation and Interpretation were recruited. Among them, six were Korean and one was Japanese. Three participants were in their fourth semester and had received three semesters of simultaneous interpreting training, while the remaining four were in their second semester and had completed one semester of training. All participants performed bidirectional Japanese–Korean and Korean–Japanese interpreting tasks, including both consecutive and simultaneous interpreting, on the TalkTrack

platform. The quality of interpreting was evaluated by a panel of four professional interpreters based on the interpreting outputs of interpreting trainees.

From the collected dataset, the present study analyzes pauses in simultaneous interpreting. Pauses, fillers, and cancellations (backtracking) were automatically detected and annotated by the TalkTrack system. The annotations were subsequently reviewed by three professional annotators with expertise in translation and interpreting, who corrected detection errors and added missing annotations where necessary.

During this review process, the combined analysis by human experts confirmed that an interpreter's pause is a significant disfluency indicator, reflecting not merely a break in speech but rather a lack of confidence or cognitive effort to find a better expression. Specifically, the annotators identified as disfluent those pauses that frequently occurred immediately before elements like numbers, fillers, hesitation, sighs, or repetitions, suggesting the pause serves as a preliminary indicator of speech planning issues. This expert insight highlights a critical nuance: not all pauses are equal.

A major limitation of existing pause detection methodologies is that they treat Intra-sentential and inter-sentence pauses identically. Previous studies have demonstrated that these two types of pauses exert different effects on speech fluency. Intra-sentential pauses, occurring unnaturally within an utterance, tend to disrupt fluency and thus have a critical impact on interpreting training. In contrast, inter-sentence pauses are often intentionally utilized by speakers to enhance content delivery and are not necessarily regarded as disfluencies by professional interpreters. However, current detection systems do not sufficiently account for this distinction, leading to potential distortions in the assessment of speech flow.

To address this problem, the present study focuses specifically on Intra-sentential pauses. This requires an accurate determination of whether an utterance has concluded. An preliminary approach relied on punctuation marks (e.g., periods); however, conventional Speech-to-Text (STT) systems often fail to reliably identify sentence boundaries, as they do not always detect utterance endings with precision. To overcome this limitation, we removed all punctuation marks from the STT output and employed a specialized sentence segmentation model, KIWI [30], to delineate sentence boundaries. This approach enabled a more precise extraction of Intra-sentential pauses by minimizing reliance on inconsistent STT-generated punctuation.

As demonstrated in Figure 5, the KIWI model utilized in this study exhibits superior performance compared to other benchmarked sentence segmentation models, justifying its selection for the precise delineation of sentence boundaries. The actual segmentation results obtained through the KIWI model are presented in Figure 6, illustrating how the system effectively identifies boundaries in continuous speech streams. By employing this robust segmentation approach, we were able to filter out inter-sentential pauses and focus exclusively on intra-sentential silences for more granular disfluency analysis. To construct the dataset for pause classification, all silent intervals were first extracted from the interpreting audio data. The audio was subsequently transcribed using STT, and the resulting text was segmented via the KIWI model. Through this process, only silent intervals occurring within sentence boundaries were retained. The final dataset comprised these Intra-sentential silence durations, each paired with a binary label (0 or 1) representing the expert interpreters' judgment on whether the interval constituted a disfluent pause, ordered sequentially by time.

The datasets were categorized into consecutive interpreting (Sources 1–7) and simultaneous interpreting (Sources 8–14). As described above, the consecutive interpreting data (Sources 1–7) were excluded from the present analysis, and the pauses in the simultaneous interpreting data (Sources 8–14) were examined. Accordingly, the evaluation was conducted primarily using the latter group.

We used data sources 8–11 from the simultaneous interpreting dataset (sources 8–14) for parameter tuning, and data sources 12–14 for performance evaluation. As will be described later, the selected model is an unsupervised learning model that does not require labeled data for training. Moreover, in real-world interpreting quality assessment systems, the entire data source is typically

provided as input without labeled annotations. In this context, splitting a single data source into separate tuning and test subsets may yield results that do not accurately reflect real-world evaluation performance. Accordingly, we performed parameter tuning using the labeled data sources 8–11 and conducted performance evaluation using the separate data sources 12–14.

```
kiwi_list

[  
  Sentence(text='이 덩구 베틀을 맞추 예 대사님 말씀 감사합니다(pause)', start=0, end=33, tokens=None, subs=[]),  
  Sentence(text='잘 들었습니다', start=34, end=41, tokens=None, subs=[]),  
  Sentence(text='네, 이것으로 이번 세션 토론자본들의 말씀을 다 청해 들었습니다', start=42, end=77, tokens=None, subs=[]),  
  Sentence(text='아, 그런데 제가 사과 말씀을(pause) 드려야 될 것 같습니다', start=78, end=114, tokens=None, subs=[]),  
  Sentence(text='제가 좌장 역할을 제대로 하지 못했던 것 같습니다', start=115, end=142, tokens=None, subs=[]),  
  Sentence(text='오 본인아(pause) 초과된 상황인데요 주제 축에서 제가 전달 발기로는 본 세션을 열주시 이심오분까지 마쳐야 된다 라고 전달 받았습니다', start=143, end=219, tokens=None, subs=[]),  
  Sentence(text='그렇기 때문에 네 들었음당 시선아 충분히 않은 것이 사실인데요', start=220, end=254, tokens=None, subs=[]),  
  Sentence(text='하지만 그중에 또 불구하고 정말 하나 정도는 드리고자 합니다(pause)', start=255, end=295, tokens=None, subs=[]),  
  Sentence(text='해당 질문은 온라인 참석차로부터 받은 질문입니다', start=295, end=321, tokens=None, subs=[]),  
  Sentence(text='네 주치지기 전달해 주신 질문인데요', start=322, end=341, tokens=None, subs=[]),  
  Sentence(text='이 질문은 네 김성환 이사장님께 드리는 질문입니다', start=342, end=369, tokens=None, subs=[]),  
  Sentence(text='EUS 프라준은 이번 바이든 대통령에게 대한민국을 방문을 했었는데요', start=370, end=408, tokens=None, subs=[]),  
  Sentence(text='그리고 또 이후에 일본을 방문을 했었습니다', start=409, end=432, tokens=None, subs=[]),  
  Sentence(text='5월 경에 방문 또 방해를 해주셨는데요 그 이후에 2개 국가가 네 네토 그리고 IPF에 모두 함께 하겠다 라고 발표를 한 바 있습니다', start=433, end=509, tokens=None, subs=[]),  
  Sentence(text='이런 2개 국가가 IPF에 함께 하는 것이 역내 내 평화 안정에 어떠한 영향을 미칠까요?', start=510, end=560, tokens=None, subs=[])]
```

Figure 5. Example output of KIWI-based sentence segmentation for Japanese-to-Korean interpreting. The model identifies sentence boundaries in Japanese source-based Korean interpreted speech, enabling accurate extraction of intra-sentential silence segments used for pause analysis.

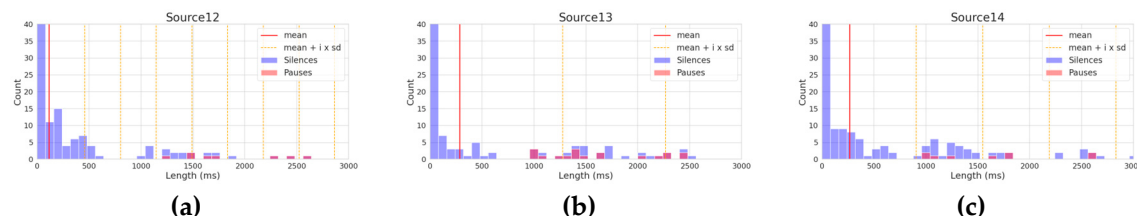


Figure 6. Comparative distribution of intra-sentential silence durations across interpreting sources. The distributions exhibit variability and heavy-tailed characteristics, motivating the use of anomaly detection rather than fixed duration thresholds. Panels (a-c) correspond to the three audio sources: (a) Source 12, (b) Source 13, and (c) Source 14.

4.2. Experimental Setting Model

For the experimental setting, data sources 8–11 from the available simultaneous interpreting dataset were used as the tuning set. Data sources 12, 13, and 14 were each designated as independent test sets, enabling a systematic evaluation of the model’s performance across multiple unseen data sources.

In Table 1, “Number of silences” refers to the total number of silence segments present within each source. In contrast, “Number of pauses” denotes the subset of those silence segments that were explicitly annotated as pauses by professional interpreters.

Table 1. Statistics of tuning and test datasets, including expert-annotated pause counts. The table summarizes the number of audio segments, silence candidates, and expert-labeled unnatural pauses used for model development and evaluation.

Type of Data	Source	Number of Silences	Number of Pauses
Tuning Data	8-11	1449	51
Test Data	12	448	8
	13	336	21
	14	337	11
	Total (12-14)	1133	40

In our experimental setup, all silence segments were used as input to the model, and a binary classification task was performed to determine whether each silence segment corresponds to a pause

or not. Accordingly, the “Number of silences” represents the total number of input samples, while the “Number of pauses” corresponds to the number of positive ground-truth instances in the dataset.

4.3. Model Design

4.3.1. Threshold Setting Based on Statistical Distribution

In this study, we implemented a dynamic thresholding mechanism for pause detection by utilizing the mean and standard deviation of silent intervals. performing a granular analysis of the pause distribution within each specific sound source, we aimed to move beyond the conventional “one-size-fits-all” approach, instead calibrating the detection criteria according to the unique statistical profile of the data.

Specifically, the pause threshold was calculated using the formula:

$$Threshold = \mu + i \times \sigma$$

where μ is the mean, i is the parameter, and σ is the standard deviation.

By varying the coefficient i , we conducted experiments across multiple datasets to identify the optimal sensitivity for each source. Our findings indicate that this statistical approach significantly improves pause detection performance compared to the traditional fixed 1-second threshold. This demonstrates that incorporating the specific distribution of silences—thereby reflecting individual variability and speech-rate fluctuations—can substantially enhance the precision of automated pause detection.

Figure 6 shows the statistical distributions of Intra-sentential silence durations observed from the three audio sources. While all the data exhibit natural, right-skewed distributions characterized by a dense concentration near zero and long tails, the mean and standard deviation are heterogeneous across audio sources. Such diversity suggests the necessity of adaptive criteria rather than a fixed threshold.

4.3.2. Anomaly Detection using Isolation Forest(iForest)

While the statistical method detailed in the previous section represents a top-down approach, characterizing the global distribution before analyzing parameters such as mean and standard deviation, we also employ a bottom-up strategy via anomaly detection. Unlike global models that assess whether a data point fits a predefined distribution, bottom-up techniques like Isolation Forests evaluate the local isolation of individual points. We specifically employed Isolation Forest (iForest) algorithm [31,32] due to its unsupervised nature and its documented efficacy in data-scarce environments. This makes it particularly suitable for our study, given the limited scale of our current dataset.

Anomaly Detection Using Isolation Forest(iForest)

The Isolation Forest(iForest) detects anomalies by isolating individual observations rather than estimating the full data density. Specifically, it constructs an ensemble of isolation trees(iTrees) using repeated random partitioning:

- 1) Random Feature Selection: At each node, the algorithm randomly selects a feature (in our case, the single feature interval length).
- 2) Random Splitting: It then randomly selects a split value within the feature’s range and partitions the data into two subsets.
- 3) Recursive Partitioning: This process repeats until points are isolated or a stopping criterion is met.

The fundamental intuition is that anomalies are significantly easier to isolate than normal observations: rare and extreme values —such as unusually long pauses— typically require fewer random splits to be separated from the dense cluster of near-zero intervals. Consequently, these outliers exhibit a shorter path length from the root to the leaf node in the iTree. Isolation Forest

aggregates this behavior across trees to compute an anomaly score, where samples with shorter average path lengths receive higher anomaly scores and are more likely to be classified as pauses (outliers).

Contamination Parameter

Isolation Forest provides additional flexibility through the contamination parameter, which represents the expected proportion of anomalies in the data (i.e., the assumed outlier rate). In practice, contamination determines the decision threshold on anomaly scores: a larger contamination value forces the model to label a larger fraction of samples as outliers.

Because this threshold directly controls the trade-off between false positives and false negatives, contamination functions as a tuning knob for precision–recall balance. In our experiments, we observed that changing contamination systematically shifts this balance; accordingly, we tuned this parameter to accurately reflect the pause prevalence in the target domain.

Domain-Knowledge-Based Contamination Selection

Importantly, the contamination parameter was not chosen arbitrarily. Instead, we determined it by leveraging domain knowledge: we statistically estimated the fraction of silent intervals labeled as pauses by human interpreters and used these estimates to set plausible outlier rates.

Concretely, we derived the contamination value from the pause ratios of simultaneous–interpreting sources 8–11, which are not used for model evaluation. Additionally, we calculated the pause ratio for consecutive interpreting sources 1–7 and confirmed that they differ substantially from those of simultaneous interpreting. This suggests that if the model is later applied to consecutive interpreting data in the future, utilizing a specific consecutive pause ratio as the contamination value would likely be more appropriate. Furthermore, this finding also demonstrates that consecutive interpreting data should not be included in the tuning set for our study.

Overall, the pause rates of simultaneous and consecutive interpreting showed distinct characteristics, and we identified contamination rates of 0.035 and 0.05 to be reasonable values. Reflecting this, we utilized distinct contamination values for the two interpreting types to achieve optimal performance. The contamination values used in this study are summarized in Table 2 below.

Table 2. Contamination parameter settings for Isolation Forest across interpreting modes. Distinct contamination ratios were explored for simultaneous and consecutive interpreting to reflect differences in pause frequency and distribution.

	Simultaneous Interpretation	Consecutive Interpretation
Contamination	0.035	0.05

Preliminary Analysis of Contamination Parameter Tuning

Before conducting the main experiments, we performed preliminary analyses to evaluate whether tuning the contamination parameter improves performance and whether restricting the tuning data to simultaneous interpretation data is a valid approach.

As Table 3 shows, when the contamination parameter was set to “auto” without tuning, the method exhibited very poor overall performance. The algorithm’s tendency to over-classify natural silences results in a misleadingly high recall. However, this performance is undermined by an extremely low precision rate; with a false positive rate nearing 90%, the model lacks the discriminative power necessary for practical application.

Table 4 shows that, when the contamination value was tuned against the tuning dataset, including both the sequential and Simultaneous Interpreting data, the performance was significantly improved.

Table 3. Performance of the Isolation Forest model without parameter tuning. Baseline anomaly detection results highlight limitations of fixed contamination settings across heterogeneous interpreting data. (c = “auto”).

Source	Precision	Recall	F1 Score	Kappa
12	0.1000	1	0.1818	0.1536
13	0.3333	1	0.5000	0.4484
14	0.1209	1	0.2157	0.1688
Total	0.1709	1	0.2920	0.2465

Table 4. Isolation Forest performance tuned using combined interpreting data. Parameter tuning across both interpreting modes improves recall but reveals trade-offs in precision (c = 0.0417).

Source	Precision	Recall	F1 Score	Kappa
12	0.3684	0.8750	0.5185	0.5061
13	0.5714	0.3810	0.4571	0.4287
14	0.4286	0.5455	0.4800	0.4609
Total	0.4468	0.5250	0.4828	0.4622

Table 5 shows that, when the contamination value was tuned against only the Simultaneous Interpreting part of the tuning dataset, an improvement was observed, although it was only marginal. Based on these preliminary results, we elected to calibrate the contamination parameter using the Simultaneous Interpreting subset of the tuning data for all subsequent experiments.

Table 5. Isolation Forest performance tuned using simultaneous interpreting data only. Mode-specific tuning yields improved detection of short abnormal pauses typical in simultaneous interpreting. (c = 0.035).

Source	Precision	Recall	F1 Score	Kappa
12	0.5000	0.8750	0.6364	0.6279
13	0.6667	0.3810	0.4848	0.4604
14	0.3333	0.3636	0.3478	0.3256
Total	0.5000	0.4750	0.4872	0.4689

4.3.3. Anomaly Detection Using Sliding Window Technique

The application of the sliding-window technique [33,34] was motivated by the observation that it is not appropriate to apply the same parameters to all segments, given that different segments may exhibit distinct utterance characteristics such as speaking rate, rhythm, and local hesitation patterns. In time-series analysis, a sliding window is a standard approach that divides a long sequence into a series of overlapping, fixed-length subsequences (windows) that move along the timeline by a predefined stride. For each window, local summary statistics or features are computed, enabling the model to capture local, time-varying dynamics rather than relying on a single global distribution. This is particularly effective when the underlying process is non-stationary—meaning its statistical properties change over time—a condition that typically holds in real speech data where pace and cadence vary across segments.

In our setting, we utilized sliding windows to capture relative anomalousness. Instead of judging a pause solely by its absolute duration across the entire recording, we evaluated whether it appears anomalous relative to its local context. While the baseline anomaly detection model used only the pause duration as input, this approach ignores the fact that a “long” pause should be interpreted differently depending on the surrounding speaking behavior. To reflect the segment-level speaking characteristics—namely the overall speaking rate and rhythmic variability—we computed the mean and standard deviation of inter-word pauses within each window and incorporated them as additional features, while otherwise following the same procedure as the simple anomaly detection model. Using the sliding-window technique, we constructed the dataset as illustrated in

Figure 7 and applied the same anomaly detection model used in Section 4.3.2 setting to this transformed data.

We selected a window size of 120 based on a trade-off between statistical stability and contextual sensitivity. Empirically, smaller windows (<60) produced unstable local statistics due to sparse extreme values, while larger windows (>200) overly smoothed distributional variations, diminishing contextual differentiation. The selected size corresponds approximately to one-third of the average sequence length in our dataset, enabling sufficient local distribution modeling while preserving sensitivity to segment-level speech variability.

Although identifying a theoretically optimal window size is difficult, our data exhibit a characteristic pattern: numerous minute pause values with occasional extremely large outliers. If the window is too small, the estimated local statistics (mean and standard deviation) become overly volatile, providing information little different from raw pause duration itself. Conversely, if the window is too large, the window statistics become overly smoothed and the original motivation—capturing changes in local speaking flow—weakens. Therefore, we fixed the window size at 120, which corresponds to approximately one-third of an average sequence in our dataset.

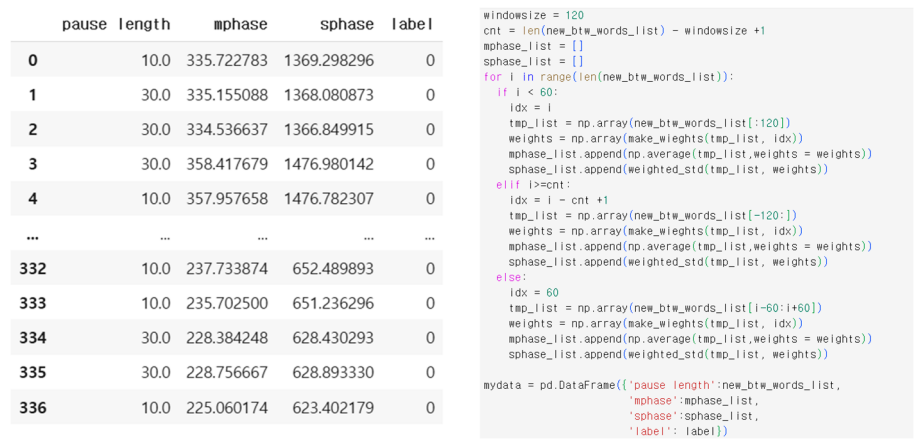


Figure 7. Data restructuring process for sliding window-based contextual anomaly detection. Pause candidates are reorganized into overlapping local contexts, allowing the model to detect abnormal pauses relative to nearby speech-rate dynamics rather than global duration statistics.

Although the number of participants is limited, the dataset contains over 1,100 intra-sentential silence segments, enabling statistically meaningful evaluation at the event level rather than solely at the speaker level. Furthermore, the primary objective of this study is methodological validation of pause detection mechanisms rather than population-level generalization. Nevertheless, future studies with larger and more diverse interpreter cohorts are necessary to confirm generalizability.

5. Results and Discussions

This section discusses the implications of the experimental findings, interpreting the observed performance improvements and examining their methodological and pedagogical significance for automated fluency assessment.

5.1. Comparative Analysis Across All Models

We aim to compare the performance of all models developed so far. For an overall comparison, we aggregated the samples from Sources 12, 13, and 14 and computed comprehensive performance metrics for each model. In total, there were 1,133 silence samples, of which 40 were identified as ground-truth pauses. We also evaluated the performance of the baseline model that had been used in the interpreting training platform prior to this study. The baseline model classifies a silence as a pause if it exceeds 1 second duration.

To evaluate model performance, we utilized standard Precision, Recall, and the F1-score, alongside Cohen’s Kappa coefficient. However, given our objective to minimize instructor workload, the practical implications of Precision and Recall differ significantly. Low precision necessitates the manual removal of erroneous annotations (commission errors), while low recall requires the de novo creation of missing annotations (omission errors). In line with the Recognition-over-Recall principle [35], the cognitive burden of the latter is substantially higher. Quantitative Human-Computer Interaction models, specifically the Keystroke-Level Model (KLM) [36], suggest that manual creation requires three to four times the temporal and mental resources of corrective pruning. Consequently, we adopted the F3-score as our primary performance metric, effectively weighting recall three times more heavily than precision to reflect this disparity in effort. Furthermore, we utilized Cohen’s Kappa as a proxy for inter-rater reliability; a higher coefficient indicates a stronger correlation with expert human ‘gold standard’ annotations, thereby ensuring the model’s output effectively mitigates the instructor’s total correction labor.

Table 6 summarizes the overall performance of models. Baseline model classifies silence samples as pauses if their duration is 1 second or longer. This model represents the conventional criterion previously utilized in the interpreting training platform. Table 6 also presents the three core methodologies evaluated in this study.

Table 6. Overall performance comparison of pause detection models. Metrics include precision, recall, F1-score, F3-score and Cohen’s kappa, demonstrating the advantage of anomaly detection over rule-based baseline approaches.

Method	Precision	Recall	F1 Score	Kappa	Remark
Baseline	0.8750	0.3684	0.5185	0.3910	1 sec
Adaptive Threshold Based on Statistical Distribution	0.8000	0.3516	0.4885	0.3725	$i = 1$
Basic Isolation Forest	0.5000	0.4750	0.4872	0.4774	$c = 0.035$
Isolation Forest with sliding window	0.5366	0.5500	0.5432	0.5486	$c = 0.035$

The baseline approach, utilizing a fixed one-second threshold, exhibits a precision-skewed performance profile. However, the significantly low recall (0.3684) presents a substantial challenge; it necessitates extensive manual creation of missing annotations by instructors. As discussed above, in the context of computer-assisted assessment, this represents a higher cognitive load than the simple removal of false positives, thereby undermining the system’s utility as a labor-saving tool.

While the distribution-based adaptive thresholds and baseline Isolation Forest methods failed to yield meaningful gains, subsequent analysis demonstrated that the dynamic adaptation approach provides a viable pathway for improvement. Notably, the Isolation Forest with sliding window method achieved a meaningful breakthrough, showing consistent progress across most performance metrics except precision. Although precision decreased from 0.875 to 0.5366 (representing a 2.7 times of increase in the labor required for pruning false positives), recall improved substantially from 0.3684 to 0.5500 (a 28.8% reduction in the burden of manual annotation creation). The rise in the F1-score from 0.5185 to 0.5432 indicates a net gain from a balanced perspective. Furthermore, when applying a weighted F3 score to reflect the greater cognitive cost of omission errors over commission errors, the performance gain becomes even more pronounced (0.3910 → 0.5486). This is corroborated by the increase in the Cohen’s Kappa coefficient (0.4933 → 0.5263), confirming that the model’s output now aligns more closely with expert human annotation and effectively reduces overall correction labor.

5.2. Per-source Fine-grained Performance Analysis

This section provides a per-source performance comparison between the baseline and the other three methods, to discuss in-depth insights into the performance of each method.

5.2.1. Fixed-threshold Approach (Baseline)

Table 7 presents the performance of the baseline model for each individual test dataset. These results serve as a reference point, allowing the performance of other models to be evaluated relative to the baseline model. The following sections present detailed performance metrics for each model across each individual test dataset.

Table 7. Source-wise performance of the baseline threshold model. Performance variability across sources indicates limitations of fixed pause-duration criteria.

Source	Precision	Recall	F1 Score	Kappa
12	1	0.381	0.5517	0.5398
13	0.8571	0.5143	0.6429	0.6127
14	0.8182	0.2308	0.36	0.3268
Total	0.875	0.3684	0.5185	0.4933

5.2.2. Adaptive Threshold Based on Statistical Distribution

A statistical distribution-based approach was employed to determine the threshold. Specifically, the threshold was defined as

$$Threshold = \mu + i \times \sigma$$

where μ is the mean, i is the parameter, and σ is the standard deviation.

The parameter i was varied from 0 to 4 for each source to evaluate its impact on performance. The “total” row at the bottom of the table represents the aggregated performance metrics computed over the entire test dataset. If an optimal value of i could be identified, it was expected that this value could be adopted as a fixed parameter in the operational system.

However, as shown in Table 8, the optimal value of i varies across different audio sources. In a real-world deployment setting, ground-truth labels are not available, making it infeasible to experimentally determine the optimal i value for each individual case. Therefore, a static, globally applicable value of i would be required. Since no single value consistently yields optimal performance across all sources, this limitation makes it difficult to adopt this model in a practical system.

Table 8. Performance of statistical thresholding under varying duration multipliers. Optimal threshold values differ across sources, supporting the need for adaptive detection strategies.

Source	mean	SD*	i	Threshold	Precision	Recall	F1 Score	Kappa
12	162.5	431.2	0	162.5	1	0.1404	0.2462	0.2218
			1	593.7	1	0.3636	0.5333	0.5208
			2	1024.9	1	0.381	0.5517	0.5398
			3	1456.1	0.875	0.7	0.7778	0.7733
			4	1887.3	0.375	0.75	0.5	0.494
13	347.1	1037.7	0	347.1	1	0.4565	0.6269	0.592
			1	1384.8	0.7143	0.5	0.5882	0.5557
			2	2422.5	0.1429	0.5	0.2222	0.2001
			3	3460.2	0.1429	0.75	0.24	0.2245
			4	4497.9	0.0476	0.5	0.087	0.077
14	338.7	709.4	0	338.7	1	0.1897	0.3188	0.2806
			1	1048.1	0.8182	0.2308	0.36	0.3268
			2	1757.5	0.5455	0.4286	0.48	0.4609
			3	2466.9	0.3636	0.4	0.381	0.3617
			4	3176.3	0.1818	0.6667	0.2857	0.2759
Total	-	-	0	-	1	0.2484	0.398	0.3619

	1	-	0.8	0.3516	0.4885	0.4622
	2	-	0.425	0.4146	0.4198	0.3982
	3	-	0.35	0.5833	0.4375	0.4222
	4	-	0.15	0.6667	0.2449	0.235

* SD, standard deviation.

Nevertheless, when examined on a per-source basis, meaningful improvements can be observed. In particular, for source 12 and source 14, the performance achieved using the respective optimal i values were substantially higher than that of the baseline model, demonstrating the potential effectiveness of this statistical thresholding approach. Furthermore, it was observed that sources with smaller standard deviations tended to require relatively larger values of i to achieve optimal performance. Based on this observation, future work could explore estimating the optimal i value using predictive approaches such as linear regression, which may enable adaptive threshold selection and further improve overall system performance.

5.2.3. Isolation Forest

The rationale for adopting this methodology was to detect pauses by reflecting the flow of utterances within a sound source, rather than merely setting different thresholds for each sound source. In other words, we aimed to dynamically determine pauses based on the overall distribution of pauses and utterance patterns, rather than applying a constant threshold even within a sound source. However, when we used pause duration as the sole feature, we found that, contrary to our expectations, all pauses exceeding a certain length tended to be detected as pauses. This was due to the fact that the pause length alone was used to determine whether a silence was a pause or not, which did not fully reflect the flow of the sound source. These results suggest that effective pause detection must go beyond a simple duration criterion and reflect the context and speech patterns of the signal.

The detailed performance metrics for each sound source, obtained by applying the proposed anomaly detection model, are summarized in Table 9. The anomaly detection results obtained using the Isolation Forest model do not substantially outperform the baseline model in terms of overall performance. However, the model demonstrates notably high recall for source 12 and source 14. As discussed earlier, recall is a particularly important metric for the task at hand, since the primary objective is to reliably detect pause events. Therefore, these results can be considered meaningful, as they indicate that the Isolation Forest model is more effective in identifying true pause instances in certain sources.

Table 9. Performance of the Isolation Forest-based anomaly detection model. The anomaly detection approach achieves improved recall and balanced detection of contextually abnormal pauses.

Source	Precision	Recall	F1 Score	Kappa
12	0.5	0.875	0.6364	0.6279
13	0.6667	0.381	0.4848	0.4604
14	0.3333	0.3636	0.3478	0.3256
Total	0.5	0.475	0.4872	0.4689

5.2.4. Isolation Forest with Sliding Window

Table 10 presents the performance of pause detection, obtained by applying the Isolation Forest with Sliding Window (IF-sliding). While IF-sliding yields substantial improvements for sources 12 and 14, performance degrades on source 13. We attribute this discrepancy to the significantly higher pause density in source 13; as detailed in Table 1, this source contains double the pause rate of other sources. Because a uniform contamination parameter was applied across all sources, the model likely underestimated the outlier density for source 13, resulting in a recall rate lower than the baseline.

This suggests that per-source optimization of the contamination parameter is a promising avenue for further improvement.

Table 10. Performance of the sliding window–based contextual anomaly detection model. Incorporating local contextual modeling improves recall and overall F1-score, demonstrating the effectiveness of speech-adaptive pause detection.

Source	Precision	Recall	F1 Score	Kappa
12	0.5000	1	0.6667	0.6585
13	0.6667	0.3810	0.4848	0.4604
14	0.4615	0.5455	0.5000	0.4823
Total	0.5366	0.5500	0.5432	0.5263

The results for source 12 are particularly demonstrative of the practical trade-offs involved. According to Table 7, the baseline exhibits a high-precision, low-recall bias ($P=1.0$, $R=0.381$), correctly identifying only 3 out of 8 pauses. Consequently, an instructor would be required to manually identify and create annotations for the 5 remaining pauses from scratch. In contrast, according to Table 10, the IF-sliding model produces a recall-skewed profile ($P=0.5$, $R=1.0$), capturing all 8 pauses but generating 8 false positives.

From a workload perspective, this presents a choice between manually creating 5 new annotations (the baseline) versus recognizing and removing 8 erroneous annotations (IF-sliding). The preference for recognition over recall-based production is well-supported by psychological studies [35]. Given the 3–4x higher cognitive cost of de novo creation compared to pruning [36], the IF-sliding approach represents a significant reduction in the instructor’s total operational burden despite the higher raw number of errors. This aligns with empirical research in translation post-editing, which demonstrates that ‘repairing’ machine-generated output is often more efficient than manual translation [37,38].

Furthermore, the observed sensitivity to pause density suggests that the model could be significantly enhanced by calibrating the contamination parameter according to the learner’s proficiency level. In an educational context, a learner’s ‘progressive level’ acts as a reliable heuristic for disfluency rates. For novice interpreters, who exhibit a higher frequency of hesitation pauses, a more aggressive (higher) contamination value would be appropriate to maintain high recall. Conversely, as the learner progresses and their delivery becomes more fluent, the parameter can be dynamically reduced. This approach effectively transforms the Isolation Forest from a rigid algorithm into an adaptive scaffolding tool that evolves alongside the learner’s developing competence.

Despite its encouraging results, the proposed framework has limitations. First, only pause duration and window-level statistics were used as features. Second, cross-linguistic generalizability remains untested beyond the Japanese–Korean language pair. Third, real-time deployment latency was not evaluated. Addressing these limitations will strengthen practical applicability.

While error analysis by human experts confirm that Intra-sentential pauses serve as critical indicators of cognitive load and speech planning issues, an error analysis of the automated model revealed specific limitations. The model occasionally misclassified syntactically permissible short pauses at clause boundaries as false positives, while failing to detect acoustically brief mid-phrase micro-pauses that experts identified as disfluent (false negatives). These findings suggest that to improve contextual understanding and classification accuracy, future iterations should integrate syntactic boundary features and prosodic cues, such as pitch contour.

6. Conclusions

Research on interpreting pauses has progressed from descriptive analyses to quantitative modeling and cognitively informed approaches [7,19,20,35]. Nevertheless, most existing methods still rely on static duration thresholds, which inadequately capture variability arising from speech rate, syntactic structure, and individual speaking styles.

To overcome these limitations, this study proposed a speech-adaptive pause detection framework that integrates Isolation Forest with sliding window-based contextual anomaly detection within the TalkTrack system. By replacing rigid threshold-based classification with context-sensitive modeling, the proposed approach improved agreement with expert annotations and enhanced the reliability of automated fluency assessment.

The findings demonstrate that pause perception in interpreting is inherently contextual rather than absolute. Listeners evaluate fluency relative to surrounding speech rhythm and cognitive flow, and the proposed framework reflects this mechanism by modeling local statistical pause characteristics. The resulting improvement in recall is particularly meaningful in educational settings, where failure to detect genuine disfluencies may reduce the effectiveness of feedback and hinder learner development.

Overall, this study confirms the effectiveness of contextual anomaly detection for identifying unnatural intra-sentential pauses and highlights the pedagogical value of speech-adaptive fluency assessment. The proposed framework contributes to the advancement of Computer-Assisted Interpreter Training by providing more accurate and actionable feedback, and it offers broader potential for application in speech analysis tasks that require context-aware modeling of temporal disfluencies.

7. Future Works

Although the proposed speech-adaptive pause detection framework improves the reliability of automated interpreting assessment, several limitations remain. First, the performance of the sliding window-based anomaly detection model depends on the selection of window size and step interval. Future studies should investigate data-driven optimization strategies to determine context lengths that best capture local speech dynamics.

Second, the current framework applies uniform model parameters across interpreters, which may not fully reflect individual differences in proficiency, speaking style, and cognitive processing strategies. Incorporating interpreter-specific adaptive parameter tuning represents an important direction for enhancing personalization and improving detection robustness.

Third, this study focused primarily on inter-word pauses and speech rate. However, pause perception is also influenced by additional prosodic and acoustic cues, including pitch variation, stress patterns, and emotional or cognitive load indicators. Integrating multimodal acoustic features could enable more comprehensive modeling of fluency and improve discrimination between functional pauses and genuine disfluencies.

Finally, future research may explore real-time deployment of context-aware pause detection within interactive training environments, allowing adaptive feedback that dynamically reflects learners' evolving speech patterns. Such developments would further strengthen the pedagogical impact of automated fluency assessment and support more personalized interpreter training.

Author Contributions: Conceptualization, Ju.L.; methodology, H.K., J.-D.K.; software, Jo.L., K.W.K.; validation, J.-D.K.; formal analysis, H.K.; data curation, H.K., K.W.K.; writing—original draft preparation H.K., Ju.L.; writing—review and editing, H.-S.P., H.-J.N.

All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Ministry of Education of the Republic of Korea 875 and the National Research Foundation of Korea (NRF-2021S1A5A2A03062819).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Korean Institutional Review Board (IRB, ewha-202104-0019-06) on 21 January 2024.

Informed Consent Statement: Not applicable.

Data Availability Statement: The interpreting audio data analyzed in this study were collected through TalkTrack, an LMS-based automated feedback system developed by the research team to support Computer-

Assisted Interpreter Training (CAIT). The dataset consists of bidirectional Korean–Japanese and Japanese–Korean interpreting performances produced by seven graduate students enrolled in a Graduate School of Translation and Interpretation, including both simultaneous and consecutive interpreting tasks. Speech disfluency phenomena, including pauses, fillers, and cancellations, were automatically detected and annotated by the TalkTrack system using integrated Speech-to-Text (STT) and annotation functionalities, and subsequently reviewed and corrected by professional annotators to ensure annotation accuracy. The dataset generated and analyzed during the current study is publicly available at: https://drive.google.com/drive/folders/1rv9rsz604n96zuvKfY830cesVGyAE8t1?usp=drive_link.

Acknowledgments: The authors would like to express their sincere gratitude to the researchers and annotators who participated in this study. Their dedicated efforts in data collection and precise annotation were fundamental to the development and validation of the TalkTrack system.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Glossary

Anomaly Detection: A machine learning approach for identifying rare or unusual patterns in data that deviate from the majority distribution. In this study, anomaly detection is used to identify unusually long intra-sentential pauses in interpreting speech.

Automated Quality Assessment (AQA): A computational framework designed to evaluate interpreting performance using objective, data-driven metrics rather than solely relying on human judgment.

Cohen’s Kappa (κ): A statistical coefficient that measures inter-rater agreement beyond chance. In this study, it quantifies the agreement between model predictions and expert-annotated pause labels.

Computer-Assisted Interpreter Training (CAIT): A pedagogical framework integrating digital technologies and automated feedback systems to enhance interpreter training efficiency and objectivity.

Contamination (Isolation Forest Parameter): A hyperparameter in the Isolation Forest algorithm representing the expected proportion of anomalies (outliers) in the dataset. It determines the classification threshold for anomaly detection.

Disfluency: Interruptions in the natural flow of speech, including pauses, fillers, repetitions, and backtracking, which may negatively affect perceived fluency.

Dynamic Thresholding: A pause detection strategy in which the decision boundary is adjusted based on statistical characteristics of the data (e.g., mean and standard deviation), rather than using a fixed duration threshold.

Filler: A vocalized hesitation such as “uh,” “um,” or equivalent expressions in other languages, indicating temporary speech planning difficulty.

F1-Score: The harmonic mean of precision and recall, used to evaluate the balance between false positives and false negatives in pause detection.

Intra-Sentential Pause: A silent interval occurring within a sentence boundary. In this study, such pauses are considered more likely to disrupt speech fluency than inter-sentential pauses.

Inter-Sentential Pause: A pause occurring at sentence boundaries, often serving functional or rhetorical purposes and not necessarily indicating disfluency.

Isolation Forest (iForest): An unsupervised anomaly detection algorithm that isolates rare observations using random partitioning. It identifies unusually long pauses by measuring average path lengths in isolation trees.

Forced alignment: Forced alignment is a speech processing technique that temporally aligns an audio recording with its corresponding transcript by mapping linguistic units (e.g., words or phonemes) onto precise time intervals in the speech signal. Using acoustic models, the method determines the start and end timestamps of each unit, enabling accurate identification of speech and non-speech segments such as pauses.

Pause Duration: The temporal length of a silent interval between words, typically measured in milliseconds. It serves as the primary feature for pause detection in this study.

Sliding Window Technique: A time-series analysis method that divides sequential data into overlapping segments (windows) to compute local statistics. In this study, it captures context-sensitive speaking dynamics for adaptive pause detection.

Speech Rate: The speed of spoken language, typically measured in syllables or words per second. It influences the interpretation of pause duration.

Speech-to-Text (STT): An automatic speech recognition technology that converts spoken language into written text for further analysis.

TalkTrack: TalkTrack is an AI-based interpreting and foreign language speaking training platform under development at Ewha Womans University. The system is designed to support interpreter training, performance assessment, and learner-centered speaking practice in both online and hybrid educational environments. By integrating speech recognition, natural language processing, and automated analytics, TalkTrack enables real-time, interactive evaluation that overcomes delays commonly found in conventional learning management systems. A central function of TalkTrack is delivery analysis, which automatically detects and visualizes features such as fillers, pauses, repetitions, and speech rate to provide objective and quantitative feedback immediately after learner performance. The platform also includes speaker detection and audio classification to identify interpreter turn-taking and distinguish source speech from interpreted output, allowing speaking-rate comparison and performance monitoring. Additionally, TalkTrack supports shadowing-based similarity measurement for simultaneous interpreting assessment. The platform aims to enhance feedback immediacy, evaluation objectivity, and personalized learning in AI-supported interpreter education.

Threshold (Pause Threshold): A predefined duration value used to determine whether a silent interval should be classified as a pause. Traditional systems often use a fixed 1-second threshold.

References

1. Da, L. Establishing interpreting fluency evaluation criteria based on correlational analysis of speech rate, articulation rate, and mean pause length. *Adv. Soc. Behav. Res.* 2025, 16, 162–166.
2. Dayter, D. Variation in non-fluencies in a corpus of simultaneous interpreting vs. non-interpreted English. *Perspectives* 2021, 29, 489–506.
3. Han, C.; Chen, S.; Fu, R.; Fan, Q. Modeling the relationship between utterance fluency and raters' perceived fluency of consecutive interpreting. *Interpreting* 2020, 22, 211–237.
4. Han, C.; Yang, L. Relating utterance fluency to perceived fluency of interpreting: A partial replication and a mini meta-analysis. *Transl. Interpret. Stud.* 2023, 18, 421–447.
5. Gósy, M. Occurrences and durations of filled pauses in relation to words and silent pauses in spontaneous speech. *Languages* 2023, 8, 79.
6. Kajzer-Wietrzny, M.; Ivaska, I.; Ferraresi, A. Fluency in rendering numbers in simultaneous interpreting. *Interpreting* 2024, 26, 1–23.
7. Han, C.; An, K. Using unfilled pauses to measure (dis) fluency in English-Chinese consecutive interpreting: In search of an optimal pause threshold (s). *Perspectives* 2021, 29, 917–933.
8. Wang, B.; Li, T. An empirical study of pauses in Chinese-English simultaneous interpreting. *Perspectives* 2015, 23, 124–142.
9. Han, C.; Lu, X. Interpreting quality assessment re-imagined: The synergy between human and machine scoring. *Interpret. Soc.* 2021, 1, 70–90.
10. Wang, X.; Wang, B. Identifying fluency parameters for a machine-learning-based automated interpreting assessment system. *Perspectives* 2024, 32, 278–294.
11. Lee, J.R.A.; Kim, J.D.; Park, H.S.; Park, H.K.; Son, J.B.; Oh, U.; Sang, W.Y.; Kim, S.; Lim, J.W.; Cho, H.S.; et al. Development of an LMS-based automatic annotation authoring tool for computer-assisted interpreter training (CAIT). *Interpret. Transl. Stud.* 2025, 29, 79–113.
12. Kim, J.D.; Wang, Y.; Fujiwara, T.; Okuda, S.; Callahan, T.J.; Cohen, K.B. Open Agile text mining for bioinformatics: the PubAnnotation ecosystem. *Bioinformatics* 2019, 35, 4372–4380.

13. Choi, M.S. A comparative analysis of disfluency in consecutive and simultaneous interpreting by trainee interpreters. *Interpret. Transl.* 2015, 17, 177–207.
14. Lee, S. Analysis of Factors of disfluency in Japanese Simultaneous Interpretation: Focused on Pause and Filler. *J. Transl. Stud.* 2021, 22, 205–230.
15. Fang, J.; Zhang, X. Pause in sight translation: A longitudinal study focusing on training effect. In *Diverse Voices in Chinese Translation and Interpreting: Theory and Practice*; Springer: Singapore, 2021; pp. 157–189.
16. Park, S. Measuring Fluency: Temporal Variables and Pausing Patterns in L2 English Speech. Ph.D. Thesis, Purdue University, West Lafayette, IN, USA, 2016.
17. Leveni, F. Structure-based Anomaly Detection and Clustering. arXiv 2025, arXiv:2505.12751.
18. Rennert, S. The impact of fluency on the subjective assessment of interpreting quality. *Interpret. Newsl.* 2010, 15, 101–115.
19. Zhang, Q.; Jing, Y. The impact of interpreting students' gestures and speech content on speech fluency of consecutive interpreting. *Front. Psychol.* 2025, 16, 1568341.
20. Cecot, M. Pauses in simultaneous interpretation: A corpus-based study of professional interpreters' performance. *Interpret. Newsl.* 2001, 11, 63–85.
21. Tissi, B. Silent pauses and disfluencies in simultaneous interpretation: A descriptive analysis. *Interpret. Newsl.* 2000, 10, 103–127.
22. Ahrens, B. Pauses (and other prosodic features) in simultaneous interpreting. *FORUM Rev. Int. d'interprétation et de traduction/Int. J. Interpret. Transl.* 2007, 5, 1–18.
23. Christodoulides, G. Prosodic features of simultaneous interpreting. In *Proceedings of the Prosody-Discourse Interface Conference*, Leuven, Belgium, 11–13 September 2013; pp. 33–37.
24. Mead, P. Methodological issues in the study of interpreters' fluency. *Interpret. Newsl.* 2005, 13, 39–63.
25. Shreve, G.M.; Lacruz, I.; Angelone, E. Sight translation and speech disfluency. In *Methods and Strategies of Process Research*; John Benjamins: Amsterdam, The Netherlands, 2011; pp. 93–120.
26. Toivola, M.; Lennes, M.; Aho, E. Speech rate and pauses in non-native Finnish. In *Proceedings of the INTERSPEECH 2009*, Brighton, UK, 6–10 September 2009; pp. 1707–1710.
27. Duez, D. Perception of silent pauses in continuous speech. *Lang. Speech* 1985, 28, 377–389.
28. Ruder, K.F.; Rupp, J.A. Pause Detection Thresholds in a Stop-Phoneme Environment. *J. Acoust. Soc. Am.* 1970, 48, 94–95.
29. Christodoulides, G.; Lenglet, C. Prosodic correlates of perceived quality and fluency in simultaneous interpreting. In *Proceedings of the Speech Prosody*, Dublin, Ireland, 20–23 May 2014; pp. 1002–1006.
30. Lee, M. Kiwi: Developing a Korean Morphological Analyzer Based on Statistical Language Models and Skip-Bigram. *KJDH* 2024, 1, 109–136.
31. Liu, F.T.; Ting, K.M.; Zhou, Z.H. Isolation Forest. In *Proceedings of the 2008 IEEE International Conference on Data Mining*, Pisa, Italy, 15–19 December 2008; pp. 413–422.
32. Xiang, H.; Zhang, X.; Dras, M.; Beheshti, A.; Dou, W.; Xu, X. Deep optimal isolation forest with genetic algorithm for anomaly detection. In *Proceedings of the 2023 IEEE International Conference on Data Mining (ICDM)*, Shanghai, China, 1–4 December 2023; pp. 678–687.
33. Guo, L.; Wei, Y. A Time Series Segment Finding Motifs Based On Sliding Window Algorithm. In *Proceedings of the 2024 IEEE International Conference on Industrial Technology (ICIT)*, Bristol, UK, 25–27 March 2024; pp. 1–6.
34. Norwawi, N.M. Sliding window time series forecasting with multilayer perceptron and multiregression of COVID-19 outbreak in Malaysia. In *Data Science for COVID-19*; Elsevier: Amsterdam, The Netherlands, 2021; pp. 547–564.
35. Haist, F.; Shimamura, A.P.; Squire, L.R. On the relationship between recall and recognition memory. *J. Exp. Psychol. Learn. Mem. Cogn.* 1992, 18, 691–702. <https://doi.org/10.1037/0278-7393.18.4.691>
36. Card, S.K.; Moran, T.P.; Newell, A. *The Psychology of Human-Computer Interaction*; Lawrence Erlbaum Associates: Hillsdale, NJ, USA, 1983.

37. Krings, H.P. *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes*; Kent State University Press: Kent, OH, USA, 2001.
38. Koponen, M. Comparing human perceptions of post-editing effort with machine translation quality metrics. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, QC, Canada, 7–8 June 2012; pp. 134–143.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.