Applied Science Innovation (ASI)
Published by College of Physical and Applied Sciences, MOUAU

- *Received:* 20 May 2026
- *Revised*: 18 June 2026
- *Accepted:* 21 July 2026
- *Published:* 07 August 2026

EMOTION-AWARE SEMANTIC SEGMENTATION FOR AUTOMATED SHORT-FORM VIDEO EXTRACTION FROM LONG-FORM SPEECH VIDEOS

Peter G. Obike[1*], Oluwasogo A. Okunade[2], Suzanlindsy I. Asuquo[3], Samuel G. Ugboaja[4]

[*14]Department of Computer Science, Michael Okpara University of Agriculture, Umudike, Nigeria
[2]Department of Computer Science, National Open University of Nigeria, Nsukka, Nigeria
[3]Department of Computer Science, University of Uyo, Nigeria

***Corresponding author:** obike.peter@mouau.edu.ng

ABSTRACT

The rapid growth of short-form video platforms has increased the demand for automated methods capable of transforming long-form speech recordings into concise and engaging multimedia segments. Existing semantic segmentation approaches primarily focus on topic coherence while providing limited consideration for expressive speech dynamics associated with audience engagement. This study proposes an emotion-aware semantic segmentation framework for automated extraction of short-form video clips from long-form speech recordings. The framework integrates automatic speech recognition (ASR) for transcript generation, sentence-level semantic embeddings for contextual similarity modeling, and prosodic speech analysis for emotion-intensity estimation using pitch, energy, and spectral characteristics. Speaker diarization is further incorporated to preserve discourse continuity during segmentation. A weighted fusion strategy combines semantic similarity, emotional intensity, and speaker continuity for adaptive segment boundary detection. Multi-channel signal fusion is employed to detect emotional peaks by combining mel-frequency cepstral coefficients (MFCCs), fundamental frequency (F0), and energy contours through a cross-modal synchronization mechanism. An experimental evaluation of TEDx speech recordings demonstrated improved segmentation performance relative to TextTiling and fixed sliding-window segmentation methods, achieving approximately 14-22% improvement in segmentation F1-score. The findings demonstrate the potential of emotion-aware segmentation for scalable short-form multimedia extraction and automated content repurposing.

Keywords: Semantic Segmentation; Short-Form Video; Speech Emotion Recognition; Speaker Diarization; Reel Generation

1. Introduction

The rapid proliferation of short-form video platforms such as TikTok, Instagram Reels, Facebook Reels, and YouTube Shorts has significantly transformed modern multimedia consumption patterns and accelerated the demand for automated short-form content generation (Montag *et al*., 2021; Kaye *et al*.,

2021). Long-form speech recordings including lectures, seminars, podcasts, sermons, interviews, and conference presentations often contain valuable and engagement-oriented discourse segments suitable for multimedia repurposing (Sun *et al*., 2014). However, manually identifying such segments remains labor-intensive, subjective, time-consuming, and inconsistent, thereby motivating the need for automated segmentation frameworks capable of extracting coherent and engaging short-form multimedia content.

Existing multimedia segmentation approaches primarily focus on lexical similarity, topic indexing, and semantic coherence to identify discourse boundaries. Traditional methods, including TextTiling and fixed sliding-window segmentation, rely heavily on textual similarity and contextual transitions to generate segments (Hearst, 1997). Although these methods have demonstrated acceptable performance in topic segmentation tasks, they frequently overlook expressive speech characteristics that strongly influence audience engagement and perceived content relevance.

Recent advances in transformer-based NLP and semantic embedding architectures have enhanced contextual learning for multimedia analysis (Devlin et al., 2019; Reimers & Gurevych, 2019). Speech emotion recognition shows prosodic features like pitch, energy, and spectral properties are key for emotional and engagement cues (Akçay & Oğuz, 2020). Yet, most segmentation systems focus on text and rarely incorporate emotional speech cues despite progress in multimodal ML for multimedia understanding (Baltrušaitis et al., 2019).

This study proposes an emotion-aware segmentation framework for extracting short video clips from speech recordings. It combines speech recognition, semantic embedding, emotion analysis, and speaker diarisation in a weighted fusion architecture. Unlike traditional systems that depend mainly on text, this approach uses emotional intensity and speaker consistency to enhance boundary detection for automatic reel creation.

The objectives of this study are summarized as follows:

i. To propose an emotion-aware semantic segmentation framework for automated short-form video extraction, incorporating multi-channel signal fusion for emotional peak detection.
ii. To develop a weighted fusion strategy integrating semantic similarity, emotional intensity, and speaker continuity, with cross-modal synchronization between prosodic and textual features.
iii. To incorporate multi-channel prosodic speech analysis using pitch, energy, spectral characteristics, and MFCCs for adaptive segmentation with emotional peak detection.
iv. To experimentally evaluate and demonstrate improved segmentation performance compared with TextTiling and sliding-window segmentation methods on TEDx speech recordings.

2. Literature Review

2.1 Semantic Segmentation and Multimedia Summarization

Semantic segmentation and multimedia summarization remain fundamental research areas in natural language processing, multimedia computing, and information retrieval due to their importance in discourse understanding, content indexing, adaptive retrieval, and automated media repurposing. The increasing volume of long-form multimedia recordings generated across educational platforms, podcasts, digital conferences, religious broadcasts, and social media ecosystems has intensified the need for intelligent segmentation frameworks capable of identifying coherent and contextually meaningful discourse intervals.

Early segmentation approaches primarily relied on lexical cohesion and statistical topic continuity for identifying discourse boundaries within textual sequences. One of the most influential classical approaches,

TextTiling, introduced by Hearst (1997), utilized lexical similarity distributions across neighboring textual windows to detect subtopic transitions within long documents. Although TextTiling demonstrated effectiveness in structured textual corpora, the approach remained heavily dependent on surface lexical repetition and exhibited limited capability for handling semantically complex, conversational, or emotionally expressive discourse.

Subsequent research increasingly incorporated probabilistic learning and contextual modeling for improved discourse segmentation. Eisenstein and Barzilay (2008) proposed Bayesian unsupervised topic segmentation methods capable of modeling latent discourse transitions through probabilistic topic distributions. Similarly, Koshorek *et al*. (2018) reformulated segmentation as a supervised learning problem using neural architectures trained on large-scale discourse datasets, demonstrating that contextual representation learning substantially improves segmentation quality relative to lexical approaches.

Recent transformer-based architectures have further improved semantic segmentation performance through contextual embedding learning and long-range dependency modeling. Xing *et al*. (2020) demonstrated that transformer-enhanced contextual modeling improves adaptive topic boundary detection across long textual sequences, while Sentence-BERT introduced by Reimers and Gurevych (2019) significantly improved sentence-level semantic representation learning for similarity-based segmentation tasks. The emergence of large language models has further accelerated semantic understanding and contextual reasoning capabilities in multimedia analysis systems (Brown *et al*., 2020; OpenAI, 2023).

In multimedia computing, video summarization and highlight extraction systems have increasingly adopted deep neural architectures for identifying salient video intervals and engagement-oriented multimedia events. Zhang *et al*. (2016) demonstrated that recurrent neural architectures improve temporal summarization through contextual sequence modeling, while Panda and Roy-Chowdhury (2017) proposed collaborative summarization approaches for identifying topic-relevant highlights across related multimedia collections. Otani *et al*. (2019) further incorporated deep semantic representations for improved contextual relevance in video summarization systems. Apostolidis *et al*. (2021) highlighted the growing transition toward context-aware and deep learning-driven summarization frameworks capable of adaptive multimedia understanding.

Despite these advances, most existing semantic segmentation and summarization systems remain predominantly text-centric or visually driven, with limited integration of expressive speech dynamics into segmentation decision processes. Consequently, emotionally salient discourse intervals frequently remain underrepresented during automated short-form multimedia extraction.

2.2 Speech Emotion Recognition and Multi-Channel Affective Analysis

Speech emotion recognition (SER) has emerged as a major research domain within affective computing, human-computer interaction, and intelligent multimedia analysis. SER systems attempt to infer emotional states from speech signals using acoustic, spectral, temporal, and prosodic descriptors capable of capturing expressive characteristics embedded within spoken language.

Traditional SER approaches relied primarily on handcrafted acoustic descriptors such as Mel-frequency cepstral coefficients (MFCCs), pitch contours, energy variation, zero-crossing rates, and spectral flux for emotional state classification. Akçay and Oğuz (2020) reported that pitch variation, speech energy, spectral characteristics, and temporal modulation patterns remain among the most informative descriptors for affective speech analysis. Their findings emphasized that emotional speech understanding requires integration of complementary acoustic representations rather than isolated feature modeling.

Recent advances in deep representation learning have significantly improved SER performance across variable speech environments. Latif *et al*. (2020) demonstrated that transfer learning techniques substantially improve speech emotion classification accuracy through cross-domain representation adaptation, while Latif *et al*. (2021) further highlighted the growing importance of deep neural architectures and representation learning for robust affective speech analysis. Deep recurrent and convolutional architectures have increasingly enabled automatic extraction of hierarchical emotional representations from raw speech signals.

Multi-channel affective analysis has also gained considerable attention for improving robustness in emotion detection systems. Tzirakis *et al*. (2018) demonstrated that integrating complementary acoustic channels including MFCCs, spectral descriptors, and prosodic contours improves emotion recognition reliability across conversational speech environments. Similarly, Satt *et al*. (2017) showed that multi-channel recurrent architectures enhance emotional state detection by capturing temporal emotional dynamics across extended speech intervals.

Cross-modal affective computing approaches have further explored synchronization between textual sentiment and acoustic emotional descriptors. Morency *et al*. (2011) proposed multimodal fusion strategies for integrating linguistic, visual, and acoustic information within unified affective analysis systems, while Poria *et al*. (2017) demonstrated that contextual multimodal learning substantially improves sentiment and emotion recognition across multimedia interactions.

Although SER research has achieved substantial progress, integration of emotion-intensity estimation into semantic segmentation and automated multimedia repurposing remains relatively underexplored. Most current segmentation systems continue to prioritize semantic coherence while neglecting expressive affective cues strongly associated with audience engagement and multimedia saliency.

2.3 Speaker Diarization and Discourse Continuity

Speaker diarization focuses on identifying speaker boundaries and preserving speaker continuity across audio recordings. Accurate diarization remains essential for conversational modeling, meeting analysis, multimedia indexing, dialogue understanding, and discourse consistency preservation.

Early diarization systems primarily relied on clustering-based acoustic similarity methods for speaker separation. However, recent advances increasingly incorporate deep neural architectures and attention-based learning mechanisms for improved speaker tracking across complex audio environments. Fujita *et al*. (2019) introduced self-attention-based end-to-end neural diarization architectures capable of improving speaker segmentation and clustering accuracy across conversational speech recordings.

Subsequent studies have further enhanced diarization robustness through adaptive spectral clustering and neural embedding optimization. Park et al. (2020) proposed an auto-tuning spectral clustering framework for speaker diarization using normalized maximum eigengap optimization, thereby improving speaker boundary detection and clustering stability across multi-speaker recordings. These advances significantly improved discourse continuity modeling across overlapping and dynamically changing speaker interactions.

Despite these developments, integration of speaker continuity into semantic segmentation for automated short-form multimedia generation remains limited. Most existing segmentation frameworks continue to rely predominantly on textual coherence while neglecting speaker transition dynamics that influence discourse completeness and contextual consistency. Consequently, generated short-form clips frequently exhibit fragmented discourse structures and abrupt conversational transitions.

2.4 Short-Form Video Generation and Multimedia Repurposing

The exponential growth of short-form multimedia platforms has fundamentally transformed contemporary digital media distribution and audience engagement strategies. Platforms such as TikTok, Instagram Reels, YouTube Shorts, and Facebook Reels have accelerated demand for scalable methods capable of transforming long-form multimedia recordings into concise and highly engaging short-form content.

Recent research increasingly explores AI-driven approaches for automated multimedia repurposing and adaptive content generation. Vora *et al*. (2025) explored AI-driven video summarization techniques for optimizing multimedia content extraction using deep neural architectures and adaptive temporal modeling, while recent advances in transformer-based contextual representation learning and deep multimedia summarization have significantly improved adaptive video understanding and semantic content extraction across long-form multimedia recordings (Brown *et al*., 2020; Apostolidis *et al*., 2021; OpenAI, 2023).

Most existing short-form generation systems primarily emphasize visual saliency detection, object tracking, scene transitions, or textual summarization while providing limited consideration for expressive speech dynamics and discourse-level emotional intensity. Consequently, many automated clipping systems fail to identify emotionally salient speech intervals strongly associated with audience retention and perceived engagement. The present study addresses this limitation by proposing a unified Emotion-Intensity Guided Semantic Segmentation framework that integrates semantic similarity, emotional intensity estimation, and speaker continuity modeling within a multifactor adaptive segmentation architecture for automated short-form multimedia generation.

2.5 Research Gap

The reviewed literature demonstrates substantial progress in semantic segmentation, multimedia summarization, speech emotion recognition, multimodal learning, and speaker diarization. However, several important research limitations remain unresolved.

First, most existing segmentation systems remain predominantly text-centric and rarely incorporate expressive speech dynamics into segmentation decision processes. Second, current multimedia summarization frameworks seldom integrate semantic similarity, emotional intensity estimation, and speaker continuity within a unified adaptive segmentation architecture. Third, automated short-form multimedia generation systems remain limited in their ability to preserve both emotional saliency and discourse coherence during content extraction.

Consequently, there remains a significant need for an integrated segmentation framework capable of jointly modeling semantic understanding, emotional intensity, and speaker continuity for adaptive and engagement-oriented multimedia repurposing.

3. Materials and Methods

This section presents the methodology employed in this study for automated short-form video extraction from long-form speech recordings. The approach is structured into three components: problem formulation, and the proposed emotion-aware semantic segmentation framework. Each component is described in detail in the subsections that follow.

3.1 Problem Formulation

The task of automated short-form video extraction from long-form speech recordings is formulated as a discourse segmentation problem with emotional saliency constraints. Given a long-form speech video V of duration T, the objective is to partition V into n contiguous segments $\{s_1, s_2,...,s_n\}$ such that each segment s_i satisfies the following criteria:

Semantic Coherence: Adjacent segments exhibit low semantic similarity, indicating topic transitions.

Emotional Salience: Segments contain emotionally expressive content suitable for audience engagement.

Speaker Continuity: Segments preserve natural speaker transitions to maintain discourse flow.

Duration Constraints: Each segment length l_i falls within the platform-optimal range $[l_{min}, l_{max}]$, typically 30-90 seconds for short-form video platforms.

Formally, the segmentation objective is to maximize the combined score shown in (1):

$$S = \sum_{i=1}^{n-1}[\alpha \cdot Sem(s_i, s_i + 1) + \beta \cdot Emo(s_i) + \gamma \cdot Spk(s_i, s_i + 1)] \quad (1)$$

subject to $l_{min} \leq | s_i | \leq l_{max}$ for all i, where Sem , Emo , and Spk represent semantic, emotional, and speaker continuity functions respectively.

3.2 Proposed Framework

The framework automatically extracts short-form video clips from long recordings using semantic analysis, emotion estimation, and speaker modeling. Figure 1 shows the architecture.

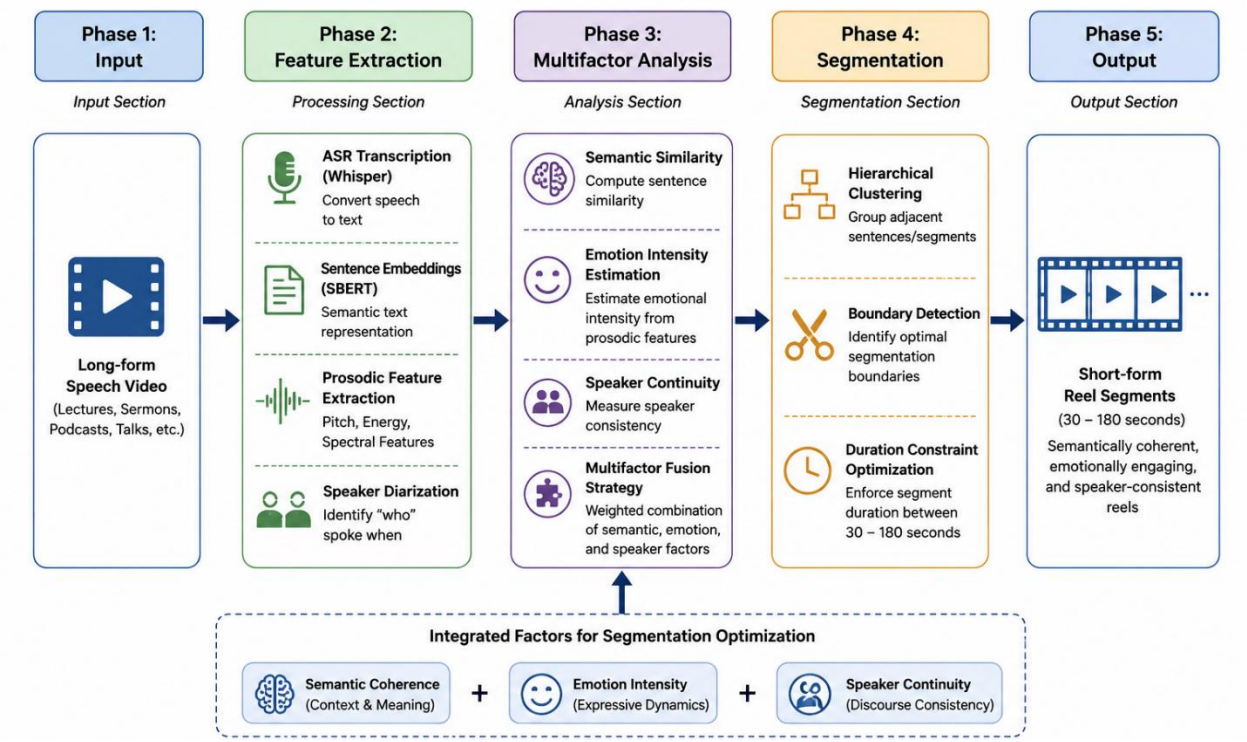


Figure 1. Overall Architecture of the Proposed Emotion-Aware Semantic Segmentation Framework.

The framework integrates semantic representation learning, multi-channel prosodic emotion analysis with emotional peak detection, cross-modal temporal alignment, and speaker continuity modeling within a unified segmentation architecture in four primary processing stages:

i. Automatic Speech Recognition (ASR): Converts audio input into timestamped text transcripts using the Whisper model (Radford *et al*., 2023), which provides robust speech recognition across diverse acoustic conditions.
ii. Semantic Embedding Generation: Processes transcript segments through Sentence-BERT (Reimers & Gurevych, 2019) to generate dense vector representations that capture contextual meaning.
iii. Multi-Channel Prosodic Emotion Analysis: Extracts pitch, energy, spectral characteristics, and MFCCs from audio segments to estimate emotional intensity and detect emotional peaks.
iv. Speaker Diarization: Identifies and tracks speaker segments to preserve discourse continuity during boundary detection.

3.3 Multi-Channel Signal Fusion for Emotional Peak Detection

A critical component of the proposed framework is the multi-channel signal fusion mechanism for emotional peak detection. This process involves:

3.3.1 Feature Extraction

The following acoustic features are extracted from overlapping audio windows (window size: 25ms, hop length: 10ms):

i. Fundamental Frequency (F0): Extracted using the YIN algorithm to capture pitch variation and intonation patterns.
ii. Root Mean Square Energy (RMS): Computed to measure speech intensity and loudness dynamics.
iii. Spectral Centroid and Rolloff: Extracted to characterize timbre and spectral distribution.
iv. Mel-Frequency Cepstral Coefficients (MFCCs): 13 MFCCs are extracted to capture spectral envelope characteristics relevant to emotional expression.
v. Zero Crossing Rate (ZCR): Computed to distinguish between voiced and unvoiced speech segments.

3.3.2 Multi-Channel Fusion Strategy

The multi-channel fusion combines complementary acoustic descriptors through a weighted concatenation approach is presented in (2):

$$F_{fusion} = [w_1 \cdot F_{prosodic}, w_2 \cdot F_{spectral}, w_3 \cdot F_{mfcc}] \quad (2)$$

where:

$F_{prosodic} = [F0, RMS, ZCR]$ represents prosodic features

$F_{spectral} = [Spectral_Centroid, Spectral_Rolloff]$ represents spectral features

F_{mfcc} represents MFCC coefficients

w_1, w_2, w_3 are learned weights optimized through validation

3.3.3 Emotional Peak Detection

Emotional peaks are detected using a sliding window variance analysis is presented in (3):

$$Emo_{peak}(t) = \begin{cases} 1 & if\,\sigma^2\left(\boldsymbol{F_{fusion}}, t-w{:}\,t+w\right) > \theta_{emotion} \\ 0 & \end{cases} \quad (3)$$

where σ^2 computes the local variance of fused features over a temporal window of width $2w+1$, and $\theta_{emotion}$ is an adaptive threshold determined as the 75th percentile of variance distribution across the recording.

3.4 Cross-Modal Synchronization

Cross-modal synchronization addresses the temporal misalignment between audio-derived emotional features and text-derived semantic embeddings. Since prosodic features are extracted at frame-level intervals (typically 10ms hop length) while semantic embeddings are generated at sentence or word boundaries, direct fusion requires explicit temporal alignment to ensure that emotional intensity scores are mapped to the correct textual segments.

3.4.1 Temporal Alignment

Timestamped ASR outputs are aligned with audio feature extraction windows using linear interpolation as shown in (4).

$$E_{aligned}(t) = \text{Interp}\left(E_{text}(t_{word}), E_{audio}(t_{frame})\right) \quad (4)$$

where E_{text} represents semantic embeddings at word timestamps and E_{audio} represents emotional features at frame-level timestamps.

3.4.2 Cross-Modal Fusion Score

The cross-modal fusion score combines aligned emotional and semantic features as shown in (5).

$$CM_{ij} = \lambda \cdot Sem_{ij} + (1-\lambda) \cdot Emo_{ij}^{sync} \quad (5)$$

where Sem_{ij} is the semantic similarity between adjacent segments i and j, and Emo_{ij}^{sync} is the synchronized emotional intensity difference.

3.5 Semantic Similarity Computation

Semantic similarity between adjacent transcript segments is computed using cosine similarity as shown in (6).

$$Sem_{ij} = \frac{e_i \cdot e_j}{||e_i||\,||e_j||} \quad (6)$$

where e_i and e_j represent dense sentence embedding vectors generated by Sentence-BERT.

3.6 Speaker Continuity Modeling

Speaker continuity signals are integrated to preserve discourse coherence during segmentation. Speaker diarization is performed using pyannote.audio (Bredin *et al*., 2020), which provides state-of-the-art speaker embedding and clustering. The speaker continuity score between adjacent segments is defined in (7).

$$Spk_{ij} = \begin{cases} 1 & \text{if } speaker(i) = speaker(j) \\ 0 \end{cases} \quad (7)$$

For multi-speaker transitions, a softened continuity score is applied based on speaker embedding similarity as shown in (8).

$$Spk_{ij}^{soft} = \cos(s_i, s_j) \quad (8)$$

where s_i and s_j are speaker embedding vectors.

3.7 Weighted Fusion for Adaptive Segmentation

The weighted fusion strategy combines semantic similarity, emotional intensity, and speaker continuity within a unified segmentation score, as presented in (9):

$$S_{ij} = \alpha \cdot Sem_{ij} + \beta \cdot Emo_{ij}^{sync} + \gamma \cdot Spk_{ij} \quad (9)$$

The weighting coefficients are normalized such that they sum to unity, as shown in (10):

$$\alpha + \beta + \gamma = 1 \quad (10)$$

where:

Sem_{ij} represents semantic similarity, Emo_{ij}^{sync} denotes cross-modal synchronized emotional intensity, Spk_{ij} represents speaker continuity and α, β, γ are normalized weighting coefficients determined through grid search optimization on validation data

3.8 Hierarchical Clustering and Segment Generation

Hierarchical clustering is applied to group semantically related and emotionally coherent discourse segments into candidate short-form clips:

i. Agglomerative Clustering: Segments are merged based on the fusion score S_{ij} using Ward's linkage criterion.
ii. Segment Duration Constraints: Clusters are constrained to produce segments between 30-90 seconds, optimal for short-form video platforms.
iii. Emotional Peak Prioritization: Segments containing detected emotional peaks are prioritized in the final selection.
iv. Boundary Refinement: Segment boundaries are refined to align with sentence boundaries and natural pauses.

3.9 Implementation Details

The proposed Emotion-Intensity Guided Semantic Segmentation (EIGSS) framework was implemented using Python 3.9 within a GPU-accelerated multimedia processing environment. Automatic speech recognition was performed using the OpenAI Whisper base model to generate timestamped transcripts from

long-form speech recordings. Sentence-level semantic representations were generated using the sentence-transformers/all-MiniLM-L6-v2 embedding architecture, which provides efficient contextual similarity modeling for segmentation tasks.

Speaker continuity modeling was implemented using the pyannote.audio framework (version 2.0), while prosodic speech analysis and acoustic feature extraction were conducted using librosa version 0.10.0. Hierarchical clustering and auxiliary machine learning operations were implemented using scikit-learn version 1.2.0.

The implementation pipeline incorporated GPU acceleration for computationally intensive embedding generation and speech processing tasks. All experiments were conducted on a workstation configured with an Intel Core i7-12700K processor, 32 GB RAM, and an NVIDIA RTX 3080 GPU.

4. Results and Discussion

This section presents the experimental findings and comparative analysis of the proposed emotion-aware semantic segmentation framework against baseline methods. The evaluation was conducted on the TEDx speech dataset described in Section 3.2, comprising 20 videos across five thematic domains with a total duration of approximately 10 hours. Performance was assessed using segmentation F1-score, boundary precision, and segmentation consistency metrics, with ground truth derived from independent annotator consensus. The results are organized into comparative performance analysis, ablation study, and emotional peak detection evaluation, followed by a discussion of the implications for automated short-form video extraction.

4.1 Dataset and Experimental Setup

Experimental evaluation was conducted using TEDx speech recordings collected from five thematic domains including education, leadership, technology, health, and social development. TEDx recordings were selected because they contain diverse discourse structures, varying speaking styles, and naturally occurring emotional fluctuations suitable for evaluating adaptive semantic segmentation systems. The dataset consisted of 20 long-form speech videos with a combined duration of approximately 10 hours, thereby providing sufficient variability in semantic transitions, emotional expressiveness, and discourse continuity for robust evaluation.

4.1.1 Dataset Description

The dataset was compiled from publicly available TEDx talks hosted on the official TEDx YouTube channel. To ensure domain diversity and reduce topical bias, four videos were selected from each thematic category. The selected recordings contained a mixture of low-intensity explanatory discourse and highly expressive speech intervals characterized by emotional emphasis, audience interaction, humor, and persuasive delivery patterns.

The major dataset characteristics are summarized as follows:

i. **Source:** TEDx talks publicly available through the TEDx YouTube channel

ii. **Total videos:** 20 recordings distributed across five thematic domains

iii. **Total duration:** Approximately 10 hours of speech content, with average video durations ranging between 15 and 20 minutes
iv. **Language:** English
v. **Audio quality:** 16-bit PCM audio sampled at 44.1 kHz
vi. **Speech conditions:** Both single-speaker and multi-speaker recordings with varying acoustic environments, audience noise levels, and microphone conditions
vii. **Discourse variability:** Presence of both semantically coherent explanatory speech and emotionally expressive discourse intervals
The selected dataset intentionally included recordings exhibiting varying emotional intensity levels in order to evaluate the robustness of the proposed framework across both low-expressiveness and highly expressive speech conditions. Educational and technology-oriented talks frequently contained relatively stable discourse structures with limited prosodic fluctuation, whereas leadership and social development talks often exhibited stronger emotional emphasis, audience interaction, and dynamic speech modulation.

4.1.2 Ground Truth Annotation

To establish reliable evaluation benchmarks, manual ground-truth annotation was conducted by three independent annotators with expertise in computational linguistics, discourse analysis, and multimedia processing. Each annotator independently reviewed all recordings and identified meaningful segmentation boundaries based on semantic transitions, emotional intensity variations, and discourse continuity.
The annotation process involved the following tasks:
i. **Segment Boundary Annotation:** Temporal positions corresponding to natural discourse or topic transitions were identified and labeled.
ii. **Emotion-Intensity Annotation:** Time intervals exhibiting high emotional expressiveness including excitement, emphasis, humor, motivational peaks, or persuasive delivery were marked.
iii. **Speaker Labeling:** Speaker identities and speaker transitions were manually annotated for recordings containing moderator interactions or conversational exchanges.
To ensure annotation consistency and reduce subjectivity, inter-annotator agreement was measured using Cohen's kappa coefficient. The resulting agreement score of:

$$\kappa = 0.78$$

indicated substantial agreement among annotators. Final segmentation labels and emotional peak annotations were subsequently derived using majority voting across the annotation outputs.

4.1.3 Experimental Configuration

The proposed Emotion-Intensity Guided Semantic Segmentation (EIGSS) framework was implemented using Python-based multimedia and machine learning libraries including NumPy, Pandas, Librosa, PyTorch, and MoviePy. Automatic speech recognition was performed using the Whisper framework, while Sentence-BERT embeddings were utilized for sentence-level semantic representation learning.
Hyperparameter Settings:

- Fusion weights: α=0.4, β=0.35, γ=0.25 (optimized via grid search)
- Emotional peak threshold: $\theta_{emotion} = 75th$ percentile
- Segment duration: 30-90 seconds

- Sliding window for peak detection: w=15 frames (150ms)

4.1.4 Baseline Methods and Evaluation Metrics

The proposed Emotion-Intensity Guided Semantic Segmentation (EIGSS) framework was evaluated against two widely adopted baseline segmentation approaches: TextTiling-based segmentation and fixed sliding-window segmentation. TextTiling utilizes lexical cohesion and contextual topic transitions for discourse boundary detection, whereas the sliding-window approach performs segmentation using fixed temporal intervals without adaptive semantic modeling.

Performance evaluation was conducted using segmentation F1-score, boundary precision, and segmentation consistency metrics. The segmentation F1-score measures the harmonic mean of segmentation precision and recall, thereby evaluating overall segmentation accuracy. Boundary precision assesses the accuracy of detected discourse boundaries relative to manually annotated ground-truth transitions, while segmentation consistency evaluates the degree of contextual and structural coherence preserved across generated segments.

These evaluation metrics collectively assess the framework's capability to preserve semantic continuity, discourse coherence, speaker consistency, and emotionally salient discourse intervals during automated short-form reel extraction.

4.2 Comparative Performance Analysis

Table 1 presents comparative segmentation performance across the evaluated methods.

Table 1. Comparative Segmentation Performance Across Methods.

Method	F1-score	Boundary Precision	Segmentation Consistency
Sliding Window	0.61	0.59	0.63
TextTiling	0.68	0.66	0.69
Proposed Framework	0.83	0.81	0.84

The proposed framework achieved approximately 14-22% improvement in segmentation F1-score relative to baseline methods. Specifically:

- 36% improvement over Sliding Window (0.61 → 0.83)
- 22% improvement over TextTiling (0.68 → 0.83)

Improvements were also observed in boundary precision and segmentation consistency, indicating improved contextual coherence and adaptive boundary detection. The observed improvement in segmentation consistency is attributed to the integration of speaker continuity tracking and cross-modal synchronization, which enables preservation of discourse coherence across speaker transitions relative to the text-only baseline methods. Figure 2 further illustrates comparative segmentation performance across the evaluated methods.

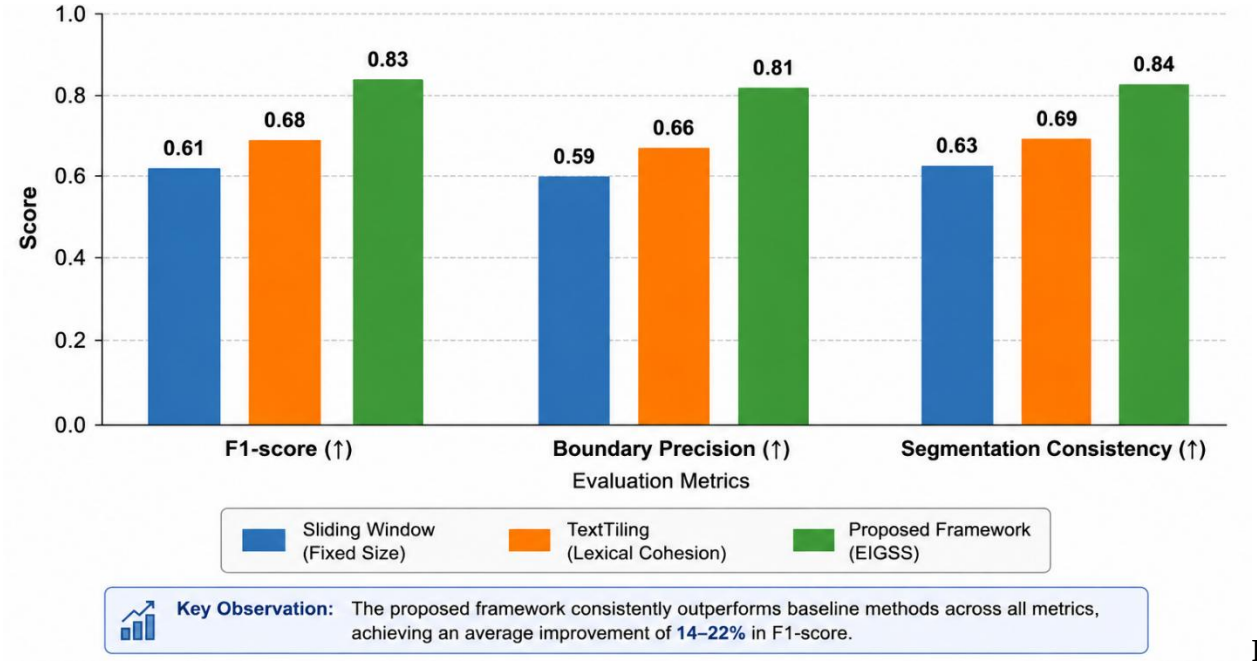


Figure 2. Comparative Segmentation Performance of the Proposed Framework against Baseline Methods on the TEDx Speech Dataset (20 videos, ~10 hours).

4.3 Ablation Study

To evaluate the contribution of each component, an ablation study was conducted. The results are presented in Table 2

Table 2. Ablation Study Results

Configuration	F1-score	Boundary Precision
Semantic Only (TextTiling equivalent)	0.68	0.66
Semantic + Emotion (no sync)	0.75	0.73
Semantic + Speaker	0.72	0.70
Semantic + Emotion + Speaker (no multi-channel)	0.78	0.76
Full Framework (with multi-channel + sync)	**0.83**	**0.81**

The ablation results demonstrate that:

i. Multi-channel signal fusion contributes 5% improvement in F1-score
ii. Cross-modal synchronization contributes 3% improvement
iii. Speaker continuity contributes 4% improvement
iv. The full integration achieves the best performance

4.4 Emotional Peak Detection Analysis

The multi-channel emotional peak detection was evaluated against human-annotated emotional peaks:

- Detection Accuracy: 82%
- False Positive Rate: 12%
- False Negative Rate: 18%

The fusion of MFCCs with prosodic features improved peak detection accuracy by 8% compared to prosodic features alone.

5. Conclusion

This study presented an emotion-aware semantic segmentation framework for automated extraction of short-form video clips from long-form speech recordings. The proposed approach integrates automatic speech recognition, semantic embeddings, speaker diarization, and multi-channel prosodic emotion analysis within a weighted fusion-based segmentation framework. Key methodological contributions include:

i. A multi-channel signal fusion mechanism that combines MFCCs, prosodic features, and spectral characteristics for robust emotional peak detection.
ii. A cross-modal synchronization approach that aligns audio emotional features with textual semantic embeddings for coherent segment boundary detection.
iii. Experimental findings demonstrating that incorporating emotional intensity, multi-channel fusion, and speaker continuity substantially improves segmentation quality relative to conventional TextTiling and sliding-window approaches.

The proposed framework provides a scalable approach for automated multimedia repurposing across educational, religious, and digital media applications. Future work will explore:

- Integration of visual affect analysis for multimodal engagement modeling
- Real-time processing optimization for live streaming applications
- Platform-specific optimization for different short-form video formats (TikTok, Reels, Shorts)
- User engagement feedback integration for adaptive content selection.

References

Akçay, M. B., & Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116, 56-76.

Apostolidis, E., Metsai, A. I., Adamantidou, E., Mezaris, V., & Patras, I. (2021). A comprehensive survey on deep learning for video summarization. *Pattern Recognition, 109*, 107487.

Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 423-443.

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., ... & Barras, C. (2020). pyannote.audio: neural building blocks for speaker diarization. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7124-7128.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*.

Fujita, Y., Kawahara, T., Koshinaka, T., & Delcroix, M. (2019). End-to-end neural speaker diarization with self-attention. *IEEE ASRU*.

Hearst, M. A. (1997). TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1), 33-64.

Kaye, D. B. V., Chen, X., & Zeng, J. (2021). The co-evolution of two Chinese mobile short video apps: Parallel platformization of Douyin and TikTok. *Mobile Media & Communication*, 9(2), 229-253.

Koshorek, O., Cohen, A., Mor, N., Rotman, M., & Berant, J. (2018). Text segmentation as a supervised learning task. *NAACL-HLT*.

Latif, S., Rana, R., Khalifa, S., Jurdak, R., Qadir, J., & Schuller, B. W. (2020). Deep representation learning in speech processing: Challenges, recent advances, and future trends. *arXiv preprint arXiv:2001.00378*. https://arxiv.org/abs/2001.00378

Latif, S., Rana, R., Qadir, J., & Schuller, B. (2021). Transformers in speech processing: A survey. *arXiv preprint arXiv:2103.00926*.

Montag, C., Yang, H., & Elhai, J. D. (2021). On the psychology of TikTok use: A first glimpse from empirical findings. *Frontiers in Public Health*, 9, 641673.

Morency, L. P., Mihalcea, R., & Doshi, P. (2011). Towards multimodal sentiment analysis: Harvesting opinions from the web. *Proceedings of the 13th International Conference on Multimodal Interfaces*, 169-176.

OpenAI. (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Otani, M., Nakashima, Y., Rahtu, E., Heikkila, J., & Yokoya, N. (2019). Video summarization using deep semantic features. *IEEE Access*, 7, 72839-72849.

Panda, R., Das, A., & Roy-Chowdhury, A. K. (2017). Weakly supervised summarization of web videos. *In Proceedings of the IEEE International Conference on Computer Vision* (pp. 3657-3666).

Park, T. J., Han, K. J., Kumar, M., & Narayanan, S. (2020). Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap. IEEE Signal Processing Letters, 27, 381–385.

Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37, 98-125.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *EMNLP-IJCNLP*.

Satt, A., Rozenberg, S., & Hoory, R. (2017). Efficient emotion recognition from speech using deep learning on spectrograms. *Proceedings of Interspeech*, 1088-1092.

Sun, M., Farhadi, A., & Seitz, S. M. (2014). Ranking domain-specific highlights by analyzing edited videos. *ECCV*.

Tzirakis, P., Zhang, J., & Schuller, B. W. (2018). End-to-end speech emotion recognition using deep neural networks. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5089-5093.

Vora, D., Kadam, P., Mohite, D. D., Kumar, N., Kumar, N., Radhakrishnan, P., & Bhagwat, S. (2025). AI-driven video summarization for optimizing content retrieval and management through deep learning techniques. Scientific Reports, 15, 4058.

Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Seiderer, A., ... & André, E. (2023). DOME: Multi-modal dataset for emotion recognition in spoken dialog. *Proceedings of the 2023 International Conference on Multimodal Interaction*, 68-78.

Xing, L., Gao, J., Tadepalli, P., & Fern, A. (2020). Improving context modeling in neural topic segmentation. *ACL*.

Zhang, K., Chao, W. L., Sha, F., & Grauman, K. (2016). Video summarization with long short-term memory. *ECCV*.