An Oriented Semantic Reasoning Framework for End-to-End Speech Topic Classification

Shanthala Tarikere Nagaraja
Department of Information Science and Engineering, Global Academy of Technology, Visvesvaraya Technological University, Belagavi, Karnataka, India
shanthala_12@rediffmail.com (corresponding author)

Kiran Y. Chandrappa
Department of Information Science and Engineering, Global Academy of Technology, Visvesvaraya Technological University, Belagavi, Karnataka, India
kiranchandrappa@gmail.com

Received: 27 March 2026 | Revised: 30 April 2026 and 18 May 2026 | Accepted: 19 May 2026

 DOI: https://doi.org/10.48084/etasr.18964

ABSTRACT

Speech topic classification aims to identify the dominant thematic category of spoken content and plays a key role in applications such as speech analytics, content indexing, and information retrieval. Despite recent progress in speech representation learning, accurately inferring topics from raw speech remains challenging due to semantic variability, long-duration dependencies, and the absence of explicit alignment between speech and topic-level semantics. Existing approaches often rely on cascading automatic speech recognition with text-based models or focus on local acoustic representations, which limits their effectiveness in end-to-end settings. This study presents a Topic-Oriented Semantic Reasoning Framework (TOSR-Framework) for end-to-end speech topic classification. The proposed framework integrates topic-oriented speech encoding, semantic alignment between speech and language representations, and global topic reasoning within a unified architecture. By emphasizing topic-relevant semantic information and enabling structured aggregation of distributed cues over time, the framework improves robustness under conversational and long-form speech conditions. Experimental evaluations on the Fisher, Switchboard, and TED Speech Topic datasets demonstrate that the proposed approach consistently outperforms existing methods, confirming its effectiveness for speech topic classification in diverse scenarios.

Keywords-speech topic classification; end-to-end speech understanding; semantic reasoning; speech-language representation alignment; long-form conversational speech analysis

I. INTRODUCTION

The increasing volume of spoken content generated through lectures, meetings, broadcasts, and online media has created a growing demand for automated techniques that can organize and interpret speech at a semantic level. Among these techniques, speech topic classification has emerged as an important research problem, as it aims to identify the underlying thematic category of a spoken segment based on its semantic content. Accurate topic classification enables efficient indexing, retrieval, summarization, and analysis of speech data across a wide range of real-world applications. Speech topic classification is inherently more challenging than topic classification in written text. Spoken language is continuous, temporally extended, and often unstructured, with semantic information distributed over long durations rather than expressed in concise sentences. In addition, speech data frequently exhibit variability in pronunciation, speaking style, and linguistic expression [1]. These challenges are further amplified in realistic scenarios where speech may contain informal language usage, multilingual expressions, or code-mixed content, leading to instability in semantic interpretation. A fundamental difficulty in speech topic classification lies in the robust representation of semantic meaning [2]. Topic inference relies on capturing high-level semantic cues rather than isolated acoustic patterns, yet these cues may be weak, ambiguous, or inconsistently expressed across a speech segment. Linguistic variability, such as the mixing of languages or the use of non-standard lexical forms, can further distort semantic signals, making direct topic inference unreliable without mechanisms that stabilize and contextualize meaning.

Existing approaches predominantly rely on cascading automatic speech recognition with text-based topic classification or adopt end-to-end speech models that primarily focus on local acoustic representations [3]. However, ASR-based pipelines suffer from error propagation, while many end-to-end models struggle to capture global semantic structure and long-range discourse information, particularly in conversational and long-form speech [4]. Moreover, most existing speech-based topic classification frameworks do not explicitly distinguish between task-relevant semantic information and

task-irrelevant acoustic variations, leading to suboptimal topic discrimination [5]. Another critical limitation lies in the absence of effective semantic alignment between speech-derived representations and topic-level language semantics [6]. Without such alignment, models fail to consistently aggregate distributed semantic evidence over time, resulting in degraded performance under topic drift, linguistic variability, and diverse speech conditions [7]. These challenges are evident across widely used benchmark datasets, including conversational corpora and structured monologic speech.

The core problem addressed in this research is the design of an end-to-end speech topic classification framework that can (i) extract topic-relevant semantic representations from raw speech, (ii) align speech semantics with topic-level language representations, and (iii) perform robust topic reasoning over long-duration speech without reliance on oracle transcripts. Addressing this problem is essential for developing scalable and reliable speech understanding systems suitable for real-world applications.

Recent advances in large-scale pre-trained speech models have significantly influenced downstream speech understanding tasks. Whisper adopts a weakly supervised training paradigm with prompt-based conditioning and has demonstrated strong robustness and generalization in automatic speech recognition [8]. Although its direct use for speech topic classification remains limited, several studies have explored its adaptability to related classification problems. For example, Whisper encoder representations have been combined with frequency-domain acoustic features using attention-based fusion, and a multi-task learning framework has been applied to jointly address emotion and gender classification [9]. In [10], fine-tuning strategies for Whisper were investigated in stuttering speech data by attaching a lightweight classification head and selectively freezing encoder layers to reduce training complexity. Whisper has also been extended to cognitive health assessment, processing long-duration speech by segmenting audio into shorter intervals and aggregating predictions to detect Alzheimer's disease, illustrating its effectiveness in transfer learning scenarios [11]. In parallel, transformer-based language models have been extensively studied for semantic classification tasks in the text domain. In [12], transformer encoders were fine-tuned for identifying misleading and inconsistent news articles by integrating cascade classification with masked language modelling and prompt-based learning.

Addressing the growing need for interpretability in large language models, an explanation framework [13] aligned contextual embeddings with feature importance scores through a Siamese-style neural network, eliminating the reliance on input perturbation during inference. Classical machine learning approaches continue to provide strong baselines for content categorization tasks, as demonstrated in the evaluation of several supervised classifiers for news topic classification, with optimized Support Vector Machines (SVM) offering competitive performance under reduced computational requirements [14, 15]. In the speech domain, a deep neural architecture [16] integrated convolutional operations with squeeze-excitation and residual connections to capture discriminative patterns in dysarthric speech. This study further adopted a cross-corpus fine-tuning strategy to improve severity classification, highlighting the role of transfer learning in speech-based analysis.

Despite recent progress in speech representation learning and transformer-based models, speech topic classification remains underexplored, particularly in end-to-end settings that do not rely on oracle transcripts [17]. Existing approaches either focus on improving acoustic representations or apply language models without explicitly modeling topic-oriented semantic abstraction across long speech durations [18]. Moreover, limited attention has been paid to the semantic alignment between speech-derived representations and topic-level language semantics, which is crucial for robust topic inference under conversational variability and topic drift [19]. These limitations highlight the need for a unified framework that integrates topic-aware speech encoding, semantic alignment, and global topic reasoning within an end-to-end speech-based classification system.

A. Motivation and Contribution

Traditional approaches to topic classification often assume that linguistic content is clean, well-structured, and homogeneous. However, such assumptions rarely hold in practical speech data. As a result, topic classification systems must be designed to handle semantic noise and contextual variability, ensuring that topic-level understanding is not overly sensitive to surface-level inconsistencies [17]. This perspective aligns with broader findings in language processing research, where semantic normalization and contextual integration have been shown to play a critical role in reliable topic inference [20]. Motivated by these challenges, this research views speech topic classification not merely as a final prediction task, but as a process that depends on the quality and stability of the underlying semantic representations [21]. By emphasizing semantic robustness and contextual coherence, the study aims to contribute toward more reliable topic classification in complex speech scenarios, particularly those involving linguistic diversity and long-form discourse. Such an approach supports the development of speech understanding systems that are better suited to real-world conditions, where variability and noise are intrinsic rather than exceptional.

The proposed Topic-Oriented Semantic Reasoning Framework (TOSR-Framework) for end-to-end speech topic classification operates directly on raw speech without reliance on oracle transcripts. A topic-oriented speech encoding strategy is introduced, which selectively emphasizes semantic information relevant to topic inference while suppressing task-irrelevant acoustic variations [22]. A semantic alignment mechanism is developed, which explicitly aligns speech-derived representations with topic-level language semantics, enabling robust aggregation of distributed semantic cues across long speech durations. A topic reasoning module performs global semantic inference to improve topic discrimination in both conversational and long-form speech. Extensive experiments on Fisher, Switchboard, and TED Speech Topic datasets demonstrate that the proposed framework consistently outperforms existing state-of-the-art models.

II. PROPOSED METHOD

The proposed TOSR-Framework operates as an end-to-end speech topic classification system that progressively transforms raw speech into high-level topic predictions through structured semantic processing [23]. The framework first employs a Topic-Oriented Speech Encoder to extract hierarchical speech representations and selectively emphasize topic-relevant semantic information while suppressing task-irrelevant acoustic variations [24]. These representations are then passed to a Semantic Alignment Network, which aligns speech-derived semantics with topic-level language representations using attention-based interaction and latent semantic tokens, enabling robust aggregation of distributed semantic cues over time [25]. Finally, a Topic Reasoning Decoder performs global semantic reasoning on the aligned representations to infer the dominant topic of the input speech without reliance on oracle transcripts. By integrating topic-aware encoding, explicit semantic alignment, and discourse-level reasoning within a unified pipeline, the framework effectively addresses semantic variability and long-duration dependency challenges inherent in speech topic classification.

A. Topic-Oriented Speech Encoder (TOSE)

The Topic-Oriented Speech Encoder is designed to extract high-level semantic representations from raw speech that are informative for topic classification. The encoder first employs a Temporal–Spectral Feature Extraction Module (TSFEM) to capture local acoustic patterns in the time–frequency domain. These features are then processed through a stack of transformer layers to model long-range contextual dependencies in the speech signal. Finally, a Topic-Guided Layer Aggregation (TGLA) mechanism selectively combines representations from different transformer layers, emphasising those that contribute most to topic-level semantic understanding [26]. The resulting representation provides a stable and topic-relevant speech embedding for downstream processing.

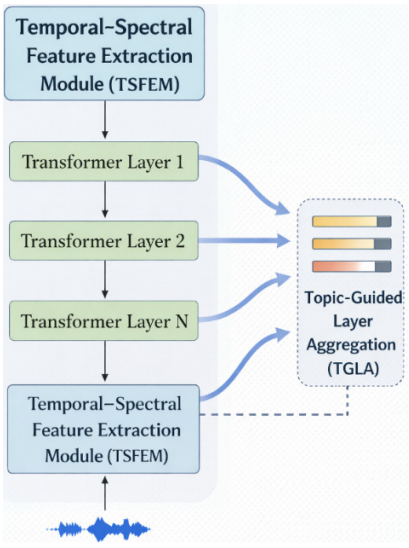

Fig. 1. Topic-oriented speech encoder.

B. Semantic Alignment Network (SAN)

The Semantic Alignment Network (SAN) aligns topic-relevant semantic information between speech and language representations through shared attention and cross-modal interaction. It consists of parallel Speech Semantic Transformer and Topic Semantic Transformer branches connected via topic-guided semantic queries. Semantic self-attention layers model intra-modality context and are shared across both branches to enforce semantic consistency [27]. Speech–Topic Cross-Attention enables explicit interaction between speech-derived features and topic text embeddings, allowing the network to extract aligned semantic cues. Semantic refinement layers further enhance discriminative topic-related features.

The network produces compact Speech Semantic Latent (SSL) and Topic Semantic Latent (TSL) representations and is trained using complementary objectives, including Speech–Topic Matching, Speech–Topic Contrastive Alignment, and Speech-to-Topic Semantic Generation. Together, these components ensure robust alignment between speech and topic semantics for downstream topic inference. Figure 2 shows the Speech–Topic Semantic Alignment and Reasoning architecture.

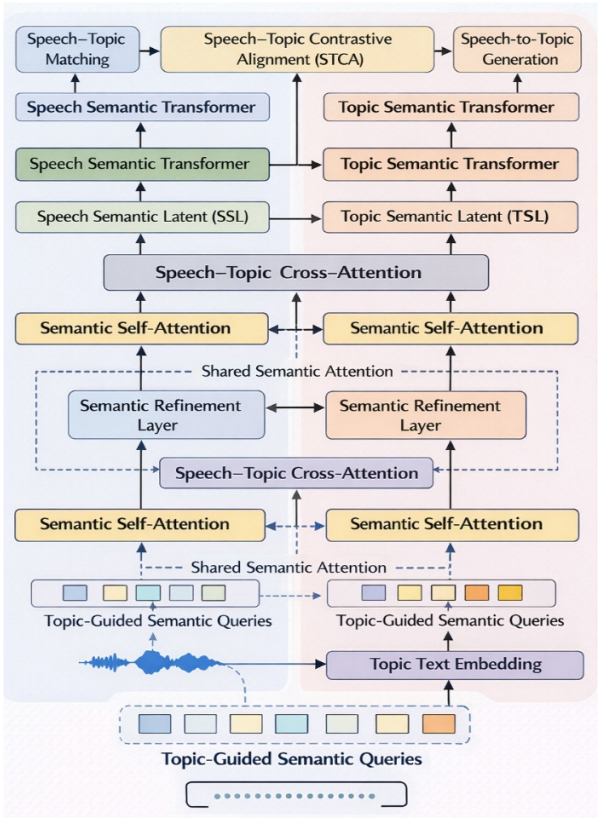

Fig. 2. Speech–Topic Semantic Alignment and Reasoning architecture.

C. Topic Reasoning Decoder

The Topic Reasoning Decoder receives topic-aligned semantic representations and performs high-level reasoning to infer the underlying topic of the speech [28]. Topic semantic

latents are first transformed into language-compatible representations through a Semantic Projection Layer. These projected representations are organized as Speech Semantic Tokens and Boundary Tokens, which define the semantic content and structural boundaries of the input sequence. A Topic-aware Tokeniser prepares the token sequence for the decoder, ensuring compatibility with the language model's input format. The Topic Reasoning Decoder, implemented using a large language model, processes these tokens to capture long-range semantic dependencies and generate coherent topic-level inferences. The decoder may optionally employ lightweight parameter adaptation to improve task-specific performance while preserving the general reasoning capabilities of the pre-trained model.

III. PERFORMANCE EVALUATION

A. *Dataset Details*

The Fisher, Switchboard, and TED Speech Topic datasets are widely used benchmarks for evaluating speech topic classification systems, as they collectively capture diverse speech characteristics and topic structures. Fisher [29] and Switchboard [30] are corpora of spontaneous telephone conversations. Their long duration and high topic variability make dominant-topic inference difficult because the relevant cues are spread thinly across informal and fragmented speech. In contrast, TedSpeechTopic is a large-scale, open-source dataset curated from TED talks, offering high-quality audio, structured discourse, and stable topic labels across formal monologic speech [31]. Together, these datasets provide a complementary evaluation framework, enabling assessment of topic classification models under both conversational and content-rich speech scenarios.

Figure 3 illustrates a performance comparison on the Fisher dataset, which consists of long, spontaneous telephone conversations with significant topic variability. Traditional text-based models such as TFIDF+SVM, TextCNN, and BERT achieve accuracies in the range of 95–96%, indicating strong performance when oracle transcripts are available. Speech-based models show relatively lower performance due to the challenges of conversational speech, with WavLM-based approaches achieving around 91%. MGANet improves performance to 94.25% by incorporating global attention mechanisms. WhisMultiNet significantly outperforms prior speech-based methods with an accuracy of 98.26%, demonstrating effective global semantic modelling. The Proposed System (PS) achieves the highest accuracy of 99.14%, indicating improved topic discrimination through enhanced topic-oriented speech encoding and semantic alignment.

Figure 4 presents the results on the Switchboard dataset, which is characterized by shorter but highly diverse conversational speech segments. Classical text-based models maintain strong performance, with BERT achieving 95.12% accuracy. However, speech-based methods exhibit a wider performance gap, with LMHA and WavLM-GCN showing lower accuracies due to the difficulty of modeling topic consistency in fragmented conversations. MGANet improves performance to 89.79% by leveraging multi-head attention. WhisMultiNet achieves a substantial improvement with an accuracy of 97.02%, reflecting its ability to aggregate semantic information across conversation segments. The proposed system further improves performance to 98.37%, confirming its robustness in handling conversational topic drift and semantic variability.

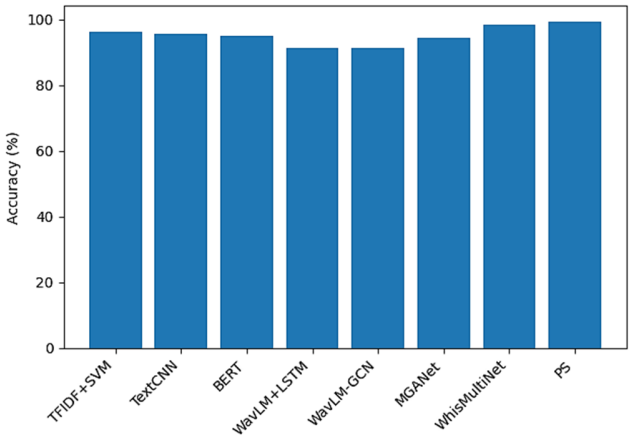

Fig. 3. Accuracy comparison on the Fisher dataset.

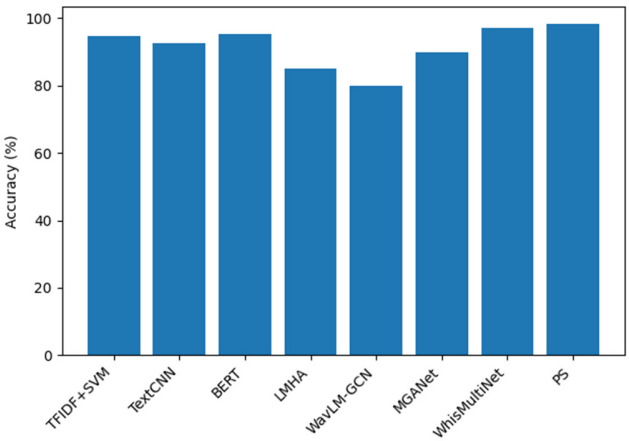

Fig. 4. Accuracy comparison on the Switchboard dataset.

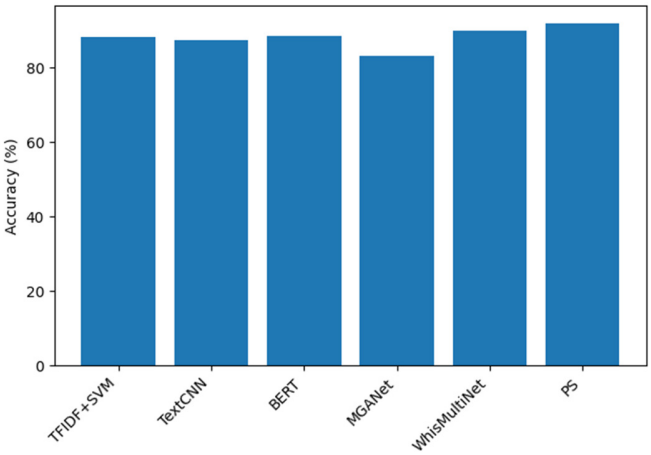

Fig. 5. Accuracy comparison on TED Speech Topic dataset.

Figure 5 shows an accuracy comparison on the TED Speech Topic dataset, which consists of structured, monologic TED talks with stable topic distributions. Text-based models perform competitively, achieving accuracies close to 88%, as the speech content is well-organized and semantically coherent. MGANet attains 83.29%, indicating limited effectiveness in capturing long-range discourse semantics. WhisMultiNet improves accuracy to 90.12% by effectively modeling global semantic dependencies across long-form speech. The proposed system achieves the highest accuracy of 92.05%, demonstrating its ability to leverage topic-oriented semantic representations and perform reliable topic reasoning on content-rich speech.

TABLE I. PERFORMANCE EVALUATION

Model	Fisher (T = 10) Accuracy (%)	Switchboard (T = 30) Accuracy (%)	TED Speech Topic (T = 10) Accuracy (%)
TFIDF + SVM	96.12	94.51	88.18
TextCNN	95.6	92.67	87.32
BERT	95.09	95.12	88.56
LMHA [19]	–	84.88	–
WavLM+LSTM+ Attention [17]	91.3	–	–
WavLM-GCN [20]	91.3	79.97	–
MGANet [21]	94.25	89.79	83.29
WhisMultiNet (Existing)	**98.26**	**97.02**	**90.12**
PS	**99.14**	**98.37**	**92.05**

B. Comparative Analysis

The proposed system consistently outperforms WhisMultiNet across all evaluated datasets, achieving improvements of 0.88%, 1.35%, and 1.93% on Fisher, Switchboard, and TED Speech Topic, respectively. These gains can be attributed to the integration of topic-oriented speech encoding, semantic alignment mechanisms, and enhanced topic-level reasoning. The improvements are particularly notable on TED Speech Topic, where long-form and content-rich speech requires stable semantic representation and effective discourse-level aggregation. Overall, the comparative results indicate that while existing models perform well under constrained or transcript-dependent settings, the proposed system offers superior robustness and generalization in end-to-end speech topic classification. By explicitly modeling topic-relevant semantics and improving alignment between speech and topic representations, PS establishes a clear advancement over existing state-of-the-art methods.

IV. CONCLUSION

This paper introduced a Topic-Oriented Semantic Reasoning Framework (TOSR-Framework) for end-to-end speech topic classification. The framework targets three problems that have limited progress in this area: semantic variability of spoken content, long-range dependencies that scatter topic cues across an utterance, and missing alignment between speech-derived features and topic-level language semantics. The architecture consists of three parts: A Topic-Oriented Speech Encoder builds a hierarchical representation of the audio and biases it toward topic-relevant content, suppressing acoustic variation that does not help classification; A Semantic Alignment Network then ties those speech features to language-level topic semantics through attention and latent semantic tokens, which lets the model pool evidence across the utterance; The Topic Reasoning Decoder reasons over the aligned features and predicts the topic directly from raw speech, with no oracle transcript required. Across three benchmarks, the framework outperformed the strongest published baselines. The framework also transfers from spontaneous telephone speech to formal monologic TED talks, an indication that topic-centric semantic modeling carries across acoustic and discursive settings.

However, several limitations of this study are worth noting. The evaluation covers only English speech corpora, so behavior on multilingual or code-mixed audio, which is common in real deployments, is still unknown. The alignment module depends on pre-trained language representations and may carry over their biases. The decoder works on fixed-length topic segments, which is awkward for streaming or low-latency use. The framework also treats topic labels as a flat categorical set and ignores hierarchical relationships between topics.

Four lines of follow-up work are planned. The first is multilingual and code-mixed speech, paired with cross-lingual alignment of the semantic encoder. The second is hierarchical topic labels, so that the decoder can return predictions at multiple granularities. The third is a streaming variant for long recordings such as broadcasts and meetings. The fourth is self-supervised pre-training of the alignment module on large unlabeled speech-text corpora, which should reduce reliance on weakly supervised topic labels. Taken together, these directions should make the TOSR-Framework more broadly applicable as an architecture for semantic understanding of long-form speech.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to all those who have supported and contributed to this research project. No funding was raised for this research.

DATA AVAILABILITY

Fisher is distributed by the Linguistic Data Consortium and described in [29]. Switchboard is described in [30]. The TED Speech Topic benchmark is described in [31]. All three corpora can be obtained from their original distributors.

REFERENCES

[1] V. Blaschke, M. Winkler, and B. Plank, "Standard-to-Dialect Transfer Trends Differ across Text and Speech: A Case Study on Intent and Topic Classification in German Dialects." arXiv, Apr. 16, 2026, https://doi.org/10.48550/arXiv.2510.07890.

[2] N. Kazanci, "Extended topic classification utilizing LDA and BERTopic: A call center case study on robot agents and human agents," *Applied Intelligence*, vol. 55, no. 5, Jan. 2025, Art. no. 360, https://doi.org/10.1007/s10489-024-06106-5.

[3] J. Fillies and A. Paschke, "Improving Hate Speech Classification with Cross-Taxonomy Dataset Integration," in Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL

2025), Feb. 2025, pp. 148–159, https://doi.org/10.18653/v1/2025.latechclfl-1.14.

[4] M. Morchid, R. Dufour, M. Bouallegue, and G. Linares, "Author-topic based representation of call-center conversations," in *2014 IEEE Spoken Language Technology Workshop (SLT)*, Dec. 2014, pp. 218–223, https://doi.org/10.1109/SLT.2014.7078577.

[5] A. S. Luna, A. Machado-Lima, and F. L. S. Nunes, "Identification and classification of speech disfluencies: A systematic review on methods, databases, tools, evaluation and challenges," *Journal of the Brazilian Computer Society*, vol. 31, no. 1, pp. 154–173, Feb. 2025, https://doi.org/10.5753/jbcs.2025.4443.

[6] J. Sun, W. Guo, Z. Chen, and Y. Song, "Topic Detection in Conversational Telephone Speech Using CNN with Multi-stream Inputs," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 7285–7289, https://doi.org/10.1109/ICASSP.2019.8682201.

[7] T. Liu and W. Guo, "Topic Classification on Spoken Documents Using Deep Acoustic and Linguistic Features," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec. 2021, pp. 427–432, https://doi.org/10.1109/ASRU51503.2021.9687969.

[8] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," in *Proceedings of the 40th International Conference on Machine Learning*, July 2023, pp. 28492–28518.

[9] A. Goel, M. Hira, and A. Gupta, "Exploring Multilingual Unseen Speaker Emotion Recognition: Leveraging Co-Attention Cues in Multitask Learning." arXiv, June 20, 2024, https://doi.org/10.48550/arXiv.2406.08931.

[10] H. Ameer, S. Latif, R. Latif, and S. Mukhtar, "Whisper in Focus: Enhancing Stuttered Speech Classification with Encoder Layer Optimization." arXiv, Nov. 09, 2023, https://doi.org/10.48550/arXiv.2311.05203.

[11] J. Li and W. Q. Zhang, "Whisper-Based Transfer Learning for Alzheimer Disease Classification: Leveraging Speech Segments with Full Transcripts as Prompts," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2024, pp. 11211–11215, https://doi.org/10.1109/ICASSP48485.2024.10448004.

[12] S. Kumar and S. R. Singh, "From Headlines and News to Truth: Leveraging LLMs With Cascade Classification and Prompt Tuning for Incongruent News Detection," *IEEE Transactions on Computational Social Systems*, vol. 13, no. 2, pp. 2438–2449, Apr. 2026, https://doi.org/10.1109/TCSS.2025.3623667.

[13] Y. Rahulamathavan, M. Farooq, and V. De Silva, "PLEX: Perturbation-Free Local Explanations for LLM-Based Text Classification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 19, no. 7, pp. 1266–1278, Oct. 2025, https://doi.org/10.1109/JSTSP.2025.3633593.

[14] S. Daud, M. Ullah, A. Rehman, T. Saba, R. Damaševičius, and A. Sattar, "Topic Classification of Online News Articles Using Optimized Machine Learning Models," *Computers*, vol. 12, no. 1, Jan. 2023, https://doi.org/10.3390/computers12010016.

[15] N. Sureja, N. Chaudhari, P. Patel, J. Bhatt, T. Desai, and V. Parikh, "Hyper-tuned Swarm Intelligence Machine Learning-based Sentiment Analysis of Social Media," *Engineering, Technology & Applied Science Research*, vol. 14, no. 4, pp. 15415–15421, Aug. 2024, https://doi.org/10.48084/etasr.7818.

[16] A. K. Roy, H. K. Kathania, and P. Sapkota, "Enhancing Speaker-Independent Dysarthric Speech Severity Classification with DSSCNet and Cross-Corpus Adaptation." arXiv, Sept. 16, 2025, https://doi.org/10.48550/arXiv.2509.13442.

[17] T. Cao, L. He, and F. Niu, "End-to-end speech topic classification based on pre-trained model Wavlm," in *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Dec. 2022, pp. 369–373, https://doi.org/10.1109/ISCSLP57327.2022.10037815.

[18] X. Qi, X. Zhao, Z. Li, and L. He, "WhisMultiNet: Advancing End-to-End Speech Topic Classification With Whisper and MultiGateGNN," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 4697–4711, 2025, https://doi.org/10.1109/TASLPRO.2025.3622978.

[19] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 12449–12460.

[20] F. Niu, T. Cao, Y. Hu, H. Huang, and L. He, "Speech Topic Classification Based on Pre-trained and Graph Networks," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*, July 2023, pp. 1721–1726, https://doi.org/10.1109/ICME55011.2023.00296.

[21] F. Niu, X. Qi, X. Chen, and L. He, "Speech Topic Classification Based on Multi-Scale and Graph Attention Networks," in *Proceedings Interspeech 2024*, 2024, pp. 4313–4317, https://doi.org/10.21437/Interspeech.2024-1934.

[22] W. N. Hsu, B. Bolte, Y. H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021, https://doi.org/10.1109/TASLP.2021.3122291.

[23] S. Bansal, H. Kamper, A. Lopez, and S. Goldwater, "Cross-Lingual Topic Prediction For Speech Using Translations," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 8164–8168, https://doi.org/10.1109/ICASSP40776.2020.9054169.

[24] S. M. Chu and L. Mangu, "Improving arabic broadcast transcription using automatic topic clustering," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2012, pp. 4449–4452, https://doi.org/10.1109/ICASSP.2012.6288907.

[25] J. Ao *et al.*, "SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 5723–5738, https://doi.org/10.18653/v1/2022.acl-long.393.

[26] I. Lane, T. Kawahara, T. Matsui, and S. Nakamura, "Out-of-Domain Utterance Detection Using Classification Confidences of Multiple Topics," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 150–161, Jan. 2007, https://doi.org/10.1109/TASL.2006.876727.

[27] H. Li, W. Ding, Y. Kang, T. Liu, Z. Wu, and Z. Liu, "CTAL: Pre-training Cross-modal Transformer for Audio-and-Language Representations," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 3966–3977, https://doi.org/10.18653/v1/2021.emnlp-main.323.

[28] T. Maekaku, J. Shi, X. Chang, Y. Fujita, and S. Watanabe, "Hubertopic: Enhancing Semantic Representation of Hubert Through Self-Supervision Utilizing Topic Model," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2024, pp. 11741–11745, https://doi.org/10.1109/ICASSP48485.2024.10448052.

[29] C. Cieri, D. Miller, and K. Walker, "The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text," in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Feb. 2004.

[30] D. Jurafsky, "Switchboard SWBD-DAMSL Shallow Discourse-Function Annotation Coders Manual," Institute of Cognitive Science - University of Colorado, USA, ICS Technical Report 97–02, 1997.

[31] F. Hernandez, V. Nguyen, S. Ghannay, N. Tomashenko, and Y. Estève, "TED-LIUM 3: Twice as Much Data and Corpus Repartition for Experiments on Speaker Adaptation," in *Speech and Computer*, vol. 11096, A. Karpov, O. Jokisch, and R. Potapova, Eds. Springer International Publishing, 2018, pp. 198–208.