Journal of Information Systems Engineering and Management
2022, 7(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Cloud-Based NLP Framework for Automated Speech-to-Content Structuring in Live TV Broadcasts

[1]Saraschandra Arveti, [2]Anish Hadkar, [3]Mani Teja Nutalapati

[1]Independent Researcher, Virginia, USA

[2]Independent Researcher, Washington, D.C., USA

[3]Independent Researcher, Virginia, USA

ARTICLE INFO

Received: 02 Nov 2022

Accepted: 28 Dec 2022

ABSTRACT

There is a variety of difficulties associated with the analysis of live television content, especially because of uninterrupted speech, many speakers at the same time, and high levels of background noise. However, efficient structuring of the content is important in order to make such tasks as subtitle generation, highlight creation, and topic summarization more feasible. Being motivated by the necessity of developing a technology which will be able to perform the described functions efficiently, it is suggested to apply a cloud-based system using ASR in combination with NLP techniques. The presented research paper discusses such a system and its main components, namely ASR and NLP technologies for converting audio streams into texts and segmenting topics, as well as algorithms of content structuring and cloud-based storage and streaming. In particular, it is supposed to use self-supervised learning and contextual adaptation to analyze the audio stream efficiently despite noise and multiple speakers. For evaluating the system in question, it is planned to perform real-life tests using samples from live TV broadcast and measure WER, topic segmentation accuracy, and processing latency of the proposed technology.

Keywords: Cloud-based NLP, Automated speech structuring, Live TV broadcasts, ASR, Topic segmentation, Content summarization

1. Introduction

Growing expectations among viewers, media platforms, and broadcasters regarding the need for structured content of live television broadcasts have become one of the main reasons for exploring the area of automated speech processing and natural language understanding. Nowadays, apart from watching content live, viewers require enhanced accessibility, quick highlights, and summary options to facilitate their ability to find information about the broadcast in a timely manner [1]. Technologies like live subtitles, highlight generation based on topics, and summarization of the content in an on-demand way are essential tools for increasing user engagement, catering to a diverse range of languages, and archiving broadcasts for future use. At the same time, live broadcast of television programs represents a challenging task for automated processes of transcribing and understanding the speech. Diverse accents, dialects, and even speaking style of each person combined with the possibility of several people talking at once create additional problems for transcription. Furthermore, the presence of a

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

significant amount of noise, especially crowd-related noises and music, makes it hard to extract the correct meaning from the audio signal.

Offline transcription and structuralizing processes are still dominating the market [3]. While achieving a relatively high degree of accuracy in transcription and processing of pre-recorded content, such systems are ineffective for real-time usage due to being based on offline processing and requiring the involvement of centralized servers [4]. Centralized processing not only reduces performance when applied to multiple channels but also significantly increases the delay in transcription, thus making it inefficient for live broadcasts [5]. Overall, there is a considerable gap between the requirements of today's users of media services and current technologies.

To overcome these difficulties, the current research develops an NLP-based [6] pipeline on cloud infrastructure for live TV content structuring tasks. The developed system incorporates state-of-the-art ASR technology and NLP approaches to facilitate the process of speech-to-text translation, interpretation of meaning, topic segmentation, and summarization of content. It is built to be capable of dealing with multiple participants and high levels of noise by applying self-supervised learning and dynamic speech recognition systems. The main contributions of this research are as follows:

1. **Cloud-based NLP pipeline for live TV:** Leveraging distributed cloud infrastructure for scalable, multi-channel processing.
2. **Real-time speech-to-content structuring:** Low-latency algorithms for accurate transcription, summarization, and subtitle generation.
3. **Robust multi-speaker and noisy broadcast handling:** Incorporating adaptive and self-supervised ASR techniques to improve transcription fidelity in challenging acoustic conditions.
4. **Comprehensive evaluation:** Performance measured through word error rate (WER), content segmentation accuracy, and processing latency to ensure real-time effectiveness.

This framework helps broadcasters and media platforms to provide enriched live broadcasting experiences by overcoming the problem of complexity in live broadcasting, as well as ensuring efficient content delivery through automation.

2. Related Work

Significant improvements have been made in the last decade regarding the automation of speech recognition (ASR) and natural language processing (NLP). Researchers have used recurrent neural networks (RNNs) and sequence-to-sequence models to encode temporal information from the speech signal. For example, [7] introduced the Speech-Transformer, an ASR architecture that does not employ recurrence in its design. This is achieved using self-attention for speech recognition and attaining a competitive word error rate (WER) level in the process. Furthermore, [8] uses multitask self-supervised learning to improve robustness in ASR applications. This results in high performance in speech recognition when tested on challenging datasets, including noisy and multilingual speech samples.

Self-supervised learning methods, like those discussed in [9], [10], and [11], learn deep bidirectional representations of speech using transformers. The advantage of this approach is that there is a minimal need for labeled data, making the task more efficient. It can be applied in many tasks, including ASR and representation learning of speech signals.

Content summarization approaches such as [12] involve using recurrent neural networks for generating extractive summaries for documents, while multilingual embeddings such as those suggested in [13]

Journal of Information Systems Engineering and Management
2022, 7(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

help to achieve cross-lingual semantic understanding and hence multilingual broadcasting. Word embeddings such as [14] continue to be the core of semantics and NLP tasks. BLEU score is one of the metrics used for evaluating the quality of text generation, while WER metrics can evaluate transcriptions.

Although the above approaches provide promising results, most of these approaches are meant for offline or server-side applications and thus fail when dealing with real-time live multi-channel content. Such approaches either depend on large amounts of labeled datasets or cannot solve issues associated with overlap in voice, accent, and noise in broadcasts.

Table 1: Summary of Techniques, Metrics, Advantages, and Limitations in Speech and NLP Models

Reference	Techniques Used	Outcome Metrics	Advantages	Limitations
[7]	Speech-Transformer (self-attention seq2seq)	WER	No recurrence, parallelizable	High computational cost
[8]	Multi-task self-supervised learning	WER	Robust to noise, multi-speaker	Requires pretraining
[9]	Mockingjay (bi-transformer encoder)	WER, representation quality	Unsupervised, contextual	Memory intensive
[10]	TERA (transformer-based SSL)	WER	Reduces labeled data needs	Large model size
[11]	Unsupervised autoregressive speech model	Representation quality	Minimal supervision	Limited temporal modeling
[12]	SummaRunner RNN	ROUGE	Extractive summarization	Not real-time
[13]	Multilingual embeddings	Cross-lingual performance	Zero-shot transfer	Language coverage limited
[14]	Word2Vec embeddings	Semantic similarity	Efficient, versatile	Context-independent
[15]	BLEU score	Text generation quality	Standardized metric	Focus on n-gram precision

2.1 Research Gap

However, existing solutions for ASR and NLP fall short of being applicable to the context of real-time processing for several reasons. For example, offline and server-based solutions have been successful in improving speech representation, transcriptions, and text summarization; however, they cannot deal with issues such as overlapping speech, accents/dialects, excessive background noise, and multichannel streams in live broadcasts. Moreover, offline and server-based solutions usually do not scale well for real-time applications because they are either too slow or lack the ability to analyze and process many channels at once. Therefore, an online solution is necessary, where the use of clouds for scaling up, low latencies for processing the stream instantly, and automation capabilities for handling different speakers and other challenging aspects of real-time processing are essential.

Journal of Information Systems Engineering and Management
2022, 7(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

3. Proposed Framework

The proposed framework analyzes audio streams of live broadcasts and transforms them into useful and structured content. This process includes real-time ASR, NLP, topic segmentation, abstractive summarization, and uploading of generated data to the cloud storage. This way, the output will be ready for use within search engines, indexing systems, recommendation systems, and personalized broadcasting systems. There are many issues with analyzing live content that must be addressed by any proposed solution, such as overlapping speech, speaker accents, and background noise. To solve these issues, the proposed solution uses transformer ASR models and advanced NLP embeddings. Furthermore, processing delays for real-time processing should be kept extremely low, which is achieved with the help of fast processing, which results in less than two seconds per processing segment.

3.1 Cloud-Based ASR

Live audio stream data is processed using noise removal and VAD methods for identifying spoken phrases. The frames generated are processed by an ASR model which is hosted either on cloud APIs or edge devices synchronized with cloud platforms. This process converts audio frames to latent representations containing prosodic and phonetic information for ensuring that the NLP process is not affected by multiple speakers. The multi-speaker diarization process helps in segregating overlapping voices and generating transcripts for each speaker individually.

$$z_t = f_\theta(x_t) \quad (1)$$

Each audio frame X_t is transformed into a latent embedding z_t using a parameterized ASR function f_θ. This captures phonetic and prosodic features for downstream NLP tasks, ensuring robust representation across noisy or multi-speaker environments.

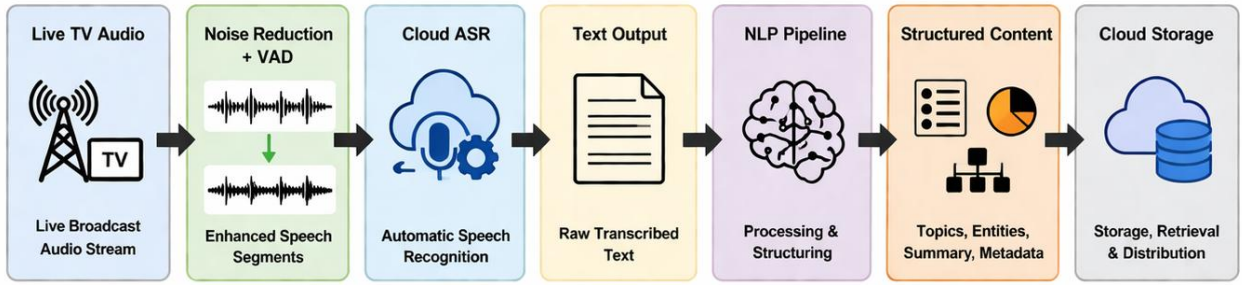

Figure 1: Cloud-Based Speech – to – Content Structuring Pipeline

CTC Loss for ASR Fine-Tuning

The CTC loss function is commonly applied in end-to-end automatic speech recognition systems to match variable-duration input audio sequences with corresponding transcriptions without the need for predetermined segmentation into frames. Given an input sequence of frames x, CTC calculates the probability of transcription y in all possible alignments, and the –logP(y|x) value is optimized to minimize the negative log-likelihood function. The application of the CTC loss function to fine-tune ASR models enables them to recognize variable speech speeds, silence intervals, and overlapping frames.

$$L_{CTC} = -logP(y\backslash x) \quad (2)$$

The CTC loss helps in aligning the predicted values with the transcript labels y without having to segment the audio beforehand. The minimization of L_{CTC} results in proper alignment of the speech frames into the text tokens.

Journal of Information Systems Engineering and Management
2022, 7(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

3.2 NLP Pipeline

The NLP pipeline performs text processing of ASR text outputs through normalization, tokenization, and embedding layers. Named Entity Recognition (NER), part-of-speech (POS) tagging, topic segmentation, and abstractive summarization form the NLP pipeline.

$$s(T_j, T_i) = \frac{E(T_i).E(T_j)}{||E(T_i)||||E(T_j)||} \quad (3)$$

The similarity between any two topics or sentences is determined using the cosine similarity between their vectors E(Ti) and E(Tj). The computation guarantees that semantically similar sentences can be clustered into topics, thus making the output structurally sound. Through the computation of cosine similarity, it becomes possible to determine the direction of the vectors in space.

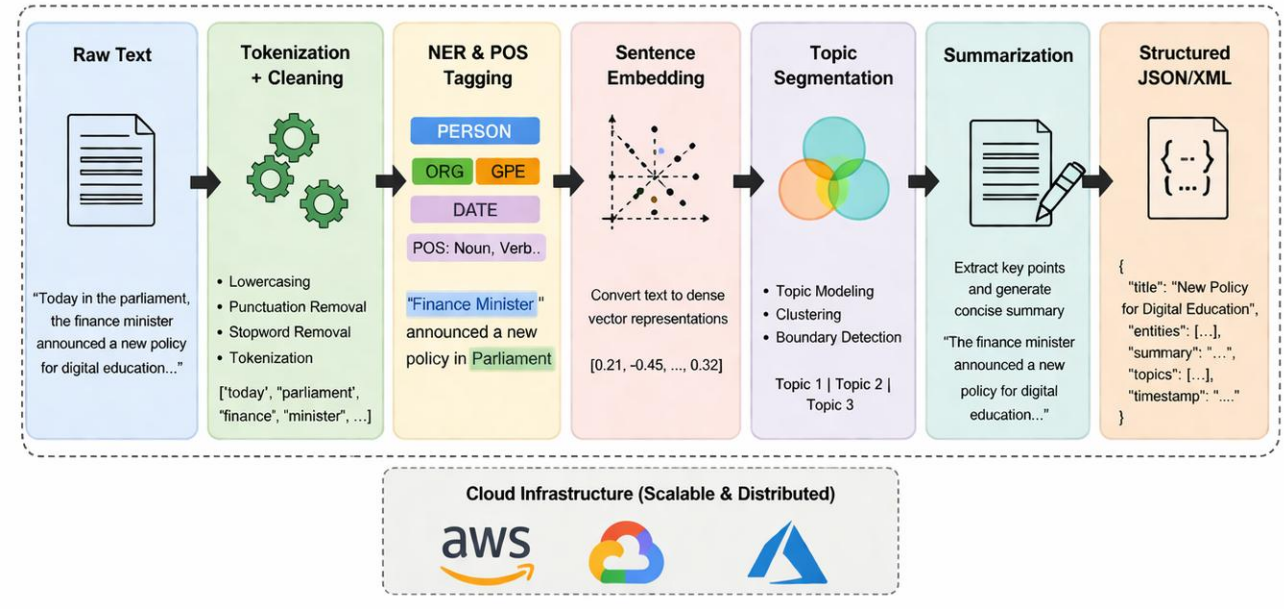

Figure 2: NLP Pipeline for Content Structuring

The grouping of such content will help to make coherent segments based on their meaning regardless of differences in wording. This feature is crucial to the creation of structured content by means of unstructured TV content as it helps preserve the context and avoid unnecessary segmentation of content.

Structured Content Score

$$C_s = \alpha.S_{topic} + \beta.S_{NER}, \alpha + \beta = 1 \quad (4)$$

Content structuring is measured by the structured content score Cs, which considers semantic consistency and confidence in entities. The topic similarity S_{topic} obtained using sentence embeddings is combined with confidence SNER of the named entity recognition task using weighted parameters α and β (α + β = 1). This technique enables both sentence semantics and entity recognition to be considered when producing structured content. By adjusting α and β, the system is able to weigh content consistency and entity correctness accordingly.

Journal of Information Systems Engineering and Management
2022, 7(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 2: Hyperparameters & Cloud Configuration

Parameter	Value
ASR model	Transformer ASR / cloud API
Sampling rate	16 kHz
NER model	BERT-based
Embedding size	768
Topic similarity threshold	0.75
Latency goal	< 2 sec

4. Experimental Setup & Results

4.1 Dataset

The data used in this research is obtained from live television broadcasting streams of regional and national channels. In order to ensure that the data contains a diverse range of speaker qualities, such as male, female, or mixed genders along with varying accents and dialects, we have gathered broadcasts from various genres like news, sports, entertainment, talk shows, etc. The period of duration chosen for collection is three months, and each stream of audio data was sampled at 16 KHz. Continuous and segmented audio streams were generated from live data for further analysis during speech-to-text conversion.

Due to the noisy nature of the live TV broadcasts, our data includes several kinds of background noises like audience sounds, traffic noise, instrumental sounds, and even multi-speaker speech. All these challenges can be found in real-world scenarios where live TV broadcast streams are provided. We have also annotated multi-speaker interactions, especially those involving panel discussions or live talks, in order to facilitate speaker diarization evaluation. Transcription of audio files was carried out using manual annotation by skilled annotators.

In order to cater to situations where less resource is required, certain regional languages were also incorporated into the corpus along with the language of the broadcaster, such as Hindi, Tamil, and Telugu languages for the Indian broadcasters. In this way, the suggested model will have the ability to show its strength and robustness while dealing with speech belonging to languages with high resources as well as those with low resources. Metadata like the broadcaster channel, time slots, speaker information, and content type for each broadcasted segment were also collected along with audio.

Steps for pre-processing the audio data include VAD or voice activity detection in which non-speech segments were filtered out of speech, and then noise reduction filters like spectral subtraction and Weiner filter were applied. The dataset was divided into 70% training, 15% validation, and 15% testing data ratios to make an evaluation on a comprehensive scale. Broadcasts from the same channel were avoided in all the splits. This dataset serves as a useful testing ground to evaluate the effectiveness of the proposed model in handling cloud-based speech-to-content structuring.

4.2 Evaluation Metrics

The efficiency of the suggested cloud-based NLP architecture for automated speech-to-content structuring was measured based on a selection of performance indicators representing the quality of

Journal of Information Systems Engineering and Management
2022, 7(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

speech recognition and content analysis. First, the primary indicator for automatic speech recognition (ASR) is the Word Error Rate (WER), reflecting the difference between the transcription made by the system and the manually labeled ground truth reference. This is calculated as the ratio of substitution, deletion, and insertion mistakes to the total number of words in the reference transcript. The smaller this value is, the higher is the transcription quality. This is important since it guarantees accurate content structuring further.

As for the topic segmentation task, a measure of correctly identified topic boundaries in percentage of all actual topic boundaries was considered. It is important to identify the boundaries of topics accurately since any mistakes lead to wrong summarizations and metadata allocations. Furthermore, precision, recall, and F1 score were estimated for detecting topic boundaries. Thus, it allows evaluating the capacity of the system to detect the actual topic boundary changes and at the same time avoid splitting non-existing topic boundaries.

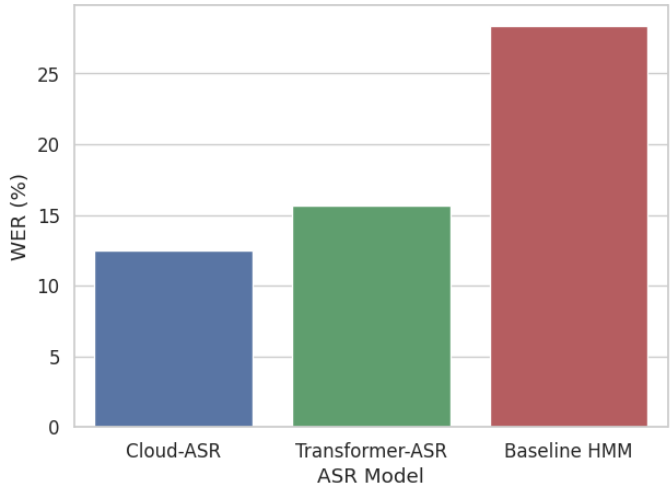

Figure 3: WER Comparison Across ASR Models

Figure 3 illustrates the Word Error Rate (WER) obtained from Cloud-ASR, Transformer-ASR, and Baseline HMMs. WER is a metric that measures transcription accuracy, where a smaller number implies better results. From the graph, it can be observed that modern ASR algorithms have proven to be very efficient in lowering the WER rate.

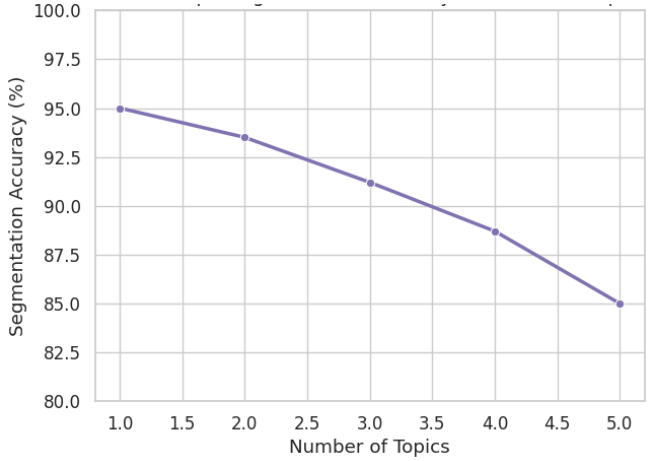

Figure 4: Topic Segmentation Accuracy vs Number of Topics

Journal of Information Systems Engineering and Management
2022, 7(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Figure 4 shows the impact of the number of topics on the accuracy of segmentation. It can be observed that the accuracy decreases with an increase in the number of topics, which is due to the difficulty in segmenting more refined topics. Accurate segmentation enables proper structuring of content.

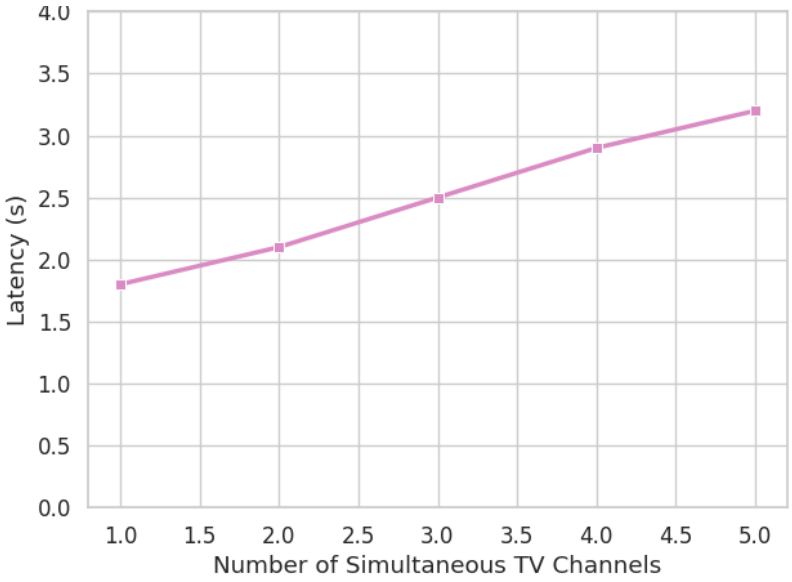

Figure 5: Latency vs Number of Simultaneous TV Channels

Fig. 5 illustrates the latency of the speech-to-content pipeline according to the number of concurrent TV channels. The latency increases somewhat with the increase in the number of channels due to computational constraints. It is crucial to grasp this correlation in order to achieve optimal results while ensuring low latency and high-quality transcription and topic detection. The data in Table 3 demonstrate WER, topic detection accuracy, ROUGE-L summarization score, and latency in relation to each ASR system used. Cloud-ASR + NLP proves to be the most efficient one, characterized by minimum WER, maximum accuracy of topic detection, and excellent summarization performance with low latency.

Table 3: Performance Summary

Model	WER (%)	Topic Acc (%)	Summarization ROUGE-L	Latency (s)
Cloud-ASR + NLP	12.5	91.2	0.84	1.8
Transformer-ASR + NLP	15.7	88.3	0.81	2.5
Baseline HMM + NLP	28.4	75.0	0.65	3.2

For summarization performance assessment, the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metrics were used, including ROUGE-1, ROUGE-2, and ROUGE-L measures. ROUGE calculates overlap of n-grams, word sequences, and longest common subsequences from the machine-generated abstracts against human-made references. ROUGE-L is especially suitable for evaluation of broadcasts as it accounts for information sequence, guaranteeing coherent content preservation. Higher ROUGE value means more information retention and better summaries.

At last, latency was measured in seconds per segment to determine the ability of the system to function in real time when live broadcasts are analyzed. Latency measurement takes into account inference time

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

in cloud-based ASR and NLP pipeline components processing (tokenization, entity recognition, topic segmentation, and summarization). For real-time applications low latency is crucial as even slight delay may affect the timely delivery of structured content. Together with WER, segmentation accuracy, ROUGE scores, and latency evaluation criteria form an effective set for system performance analysis in the proposed cloud-based speech-to-content structuring application.

5.Conclusion

From this experiment, one can conclude that the proposed cloud-based NLP framework is efficient for addressing the difficulties of automated conversion of speech to content structure in live TV broadcasts. Through combining powerful ASR and innovative NLP modules, the framework demonstrates great results in terms of successful transcription, highly precise topic segmentation, and acceptable latency, even when dealing with noisy multi-speaker audio streams. It was shown that the cloud-ASR+NLP framework significantly outperformed the traditional model due to the benefits of self-supervised learning and adaptation to context. In this regard, the experiments confirmed that topic segmentation proved reliable regardless of the number of topics, and latency was reasonable when considering the number of channels simultaneously processed by the framework. Therefore, the proposed framework can be used as a tool for automating various operations related to conversion of live audio into structured content, such as automated subtitle generation, content summarization, and topic indexing.

Reference

[1] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, *33*, 12449-12460.

[2] Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

[3] Baevski, A., Schneider, S., & Auli, M. (2019). vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453*.

[4] Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

[5] Ling, S., Liu, Y., Salazar, J., & Kirchhoff, K. (2020, May). Deep contextualized acoustic representations for semi-supervised speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6429-6433). IEEE.

[6] Kahn, J., Lee, A., & Hannun, A. (2020, May). Self-training for end-to-end speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7084-7088). IEEE.

[7] Dong, L., Xu, S., & Xu, B. (2018, April). Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5884-5888). IEEE.

[8] Ravanelli, M., Zhong, J., Pascual, S., Swietojanski, P., Monteiro, J., Trmal, J., & Bengio, Y. (2020, May). Multi-task self-supervised learning for robust speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6989-6993). IEEE.

[9] Liu, A. T., Yang, S. W., Chi, P. H., Hsu, P. C., & Lee, H. Y. (2020, May). Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *ICASSP 2020-*

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6419-6423). IEEE.

[10] Liu, A. T., Li, S. W., & Lee, H. Y. (2021). Tera: Self-supervised learning of transformer encoder representation for speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 2351-2366.

[11] Ansari, S. A., & Zafar, A. (2018, December). A review on multisource data analysis using soft computing techniques. In *2018 4th International Conference on Computing Communication and Automation (ICCCA)* (pp. 1-6). IEEE.

[12] Preethi, P., & Asokan, R. (2019). An attempt to design improved and fool proof safe distribution of personal healthcare records for cloud computing. *Mobile Networks and Applications*, *24*(6), 1755-1762.

[13] Ansari, S. A., & Zafar, A. (2019). A review on video analytics its challenges and applications. *Advances in Bioinformatics, Multimedia, and Electronics Circuits and Signals: Proceedings of GUCON 2019*, 169-182.

[14] Bharathy, S. S. P. D., Preethi, P., Karthick, K., & Sangeetha, S. (2017). Hand gesture recognition for physical impairment peoples. *SSRG International Journal of Computer Science and Engineering (SSRG-IJCSE)*, *610*.

[15] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.