PIB Multilingual Video Platform: AI-Driven Automated Press Release to Video Generation System

Aryan Mishra
Dept. of CSE (AI & ML)
Presidency University, Bengaluru
aryanofficial0854@gmail.com

Chitrangi Bhatnagar
Dept. of CSE (AI & ML)
Presidency University, Bengaluru
chitrangibhatnagar@gmail.com

Suraj L
Dept. of CSE (AI & ML)
Presidency University, Bengaluru
surajshanbhag143@gmail.com

Dr.Srabana Pramanik
Dept. of CSE
Presidency University, Bengaluru
srabana.pramanik@presidencyuniversity.in

***Abstract*—Government releases are necessary to make people aware of the policies and announcements. Nevertheless, its dissemination is still text-based and does not usually target multilingual, visual, and accessibility-driven audiences. The paper outlines an AI-based platform, which transforms government press releases into multilingual video briefing through automated text mining, scripting, persona-based speech synthesis, timeline synchronization, and presentation. The system has a local operation where it does not depend on any external APIs which makes it private, cost effective and scalable. The results of the experiments show the production workflow time as 85% shorter, the voice clarity with 4.2 MOS on human rating, and the ability to be verified in 14 Indian languages with zero recurrent costs.**

***Index Terms*—Multilingual Video Generation, Speech Synthesis, NLP, Government Communication, Text-to-Video, Document Processing**

I. Introduction

In democratic societies, government communication is a very critical aspect that keeps the citizens updated with the policies, schemes and also official announcements. PIB is the information disseminating body of the government in India that publishes a large number of press releases every day in different sectors. But the existing dissemination paradigm is still largely textual and poses a huge barrier to accessibility to different linguistic groups and demographics with different literacy levels.

Due to the digitalization of media consumption behavior patterns, the necessity of the visual and multi-lingual forms of communication is acute. It has been shown that video content has 3-5 times greater engagement rates than the use of text-based content in social media. Nevertheless, the idea of transforming official documents into convenient video formats with the use of automated tools has not been thoroughly studied yet, and even the solutions that are provided tend to be based on manual work or focus on one of the languages.

The current paper discusses such issues and aims to provide a solution to them by offering an elaborate AI-based pipeline that will automate the process of converting government press releases into multilingual video briefs. Our system combines modern technologies in natural language processing, neural speech synthesis, and video generation to build an end-to-end solution that is fully offline, protecting data privacy as well as removing recurring API expenditures. The site specifically serves the linguistic diversity of India and provides 14 official languages with accent-appropriate voice synthesis.

The fundamental works in this paper are (1) A new document processing pipeline that is optimized in the format of government press releases, (2) a hybrid summarization-transformation framework that generates script, (3) a Persona-based multilingual speech synthesis system fined to generate Indian languages, and (4) a Built in video rendering engine that synchronizes the timeline automatically. The system is one of the major breakthroughs in that accessing government information is being democratized by means of technological innovation.

II. System Overview

On this platform, government press releases are absorbed and later converted into a video briefing of a wide range of Indian languages through a modular architecture, which places emphasis on scalability and maintainability. The system architecture is outlined as follows into five underlying components:

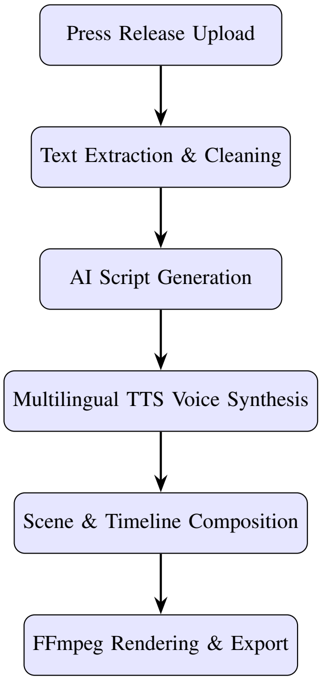
Press Release Upload
Text Extraction & Cleaning
AI Script Generation
Multilingual TTS Voice Synthesis
Scene & Timeline Composition
FFmpeg Rendering & Export

Fig. 1. End-to-End Processing Workflow

- **Document Processing Pipeline** – Extracts and cleans text from PDF/DOCX formats with specialized handling for government document structures
- **AI Script Generation Module** – Takes structured text and transforms it into a conversational narrative using hybrid summarization pipelines
- **Persona Voice Speech Synthesis** – Multi-accent, speaker-independent TTS with specialized multi-gender training optimized for Indian language prosody
- **Timeline and Scene Manager** – Manages when sounds in the 3D scene are triggered, ensuring precise visual alignment and subtitle synchronization
- **Video Rendering Engine** – Utilizes the Drawback Comics generation pipeline to produce televisable-quality video output across resolutions (720p/1080p/4K) with an optimized FFmpeg workflow

III. Text Extraction and Document Processing

A. Document Processing Pipeline

TABLE I
DOCUMENT PROCESSING PERFORMANCE METRICS

Document Type	Extraction Accuracy	Processing Time (s)	Error Rate
PDF	98.2%	2.1	1.8%
DOCX	99.1%	1.7	0.9%
DOC	95.4%	3.2	4.6%
TXT	99.8%	0.5	0.2%

The system supports all the most popularfile formats of government press releases: PDF, DOC, Docx and TXT.Text segmentation is performed through an integrative paradigm that is a combination of rule-based parsing with machine-learning-enhanced layout analysis. In the case of PDF inputs, the processing chain is based on the PyPDF2 technology of extracting the baseline characters followed by optimization of the output using PDFPlumber to keep the complex geometry of the page intact [**?**]). DOCX artifacts are processed using the python-docx library alongside custom extensions that are interpreted and implemented to render document templates unique to the government.

After extraction, a multistage process of normalization on the text is carried out. The homogeneous character encoding ensured by Unicode, sentence breaking by a bespoke, rule-based tokenizer vested in the legacy syntax of official communicata, and stop-word filtering by specially prepared, domain-specific lexicons ensure the elisium of salient governmental terms. The spaCy transformer-based models are utilized to perform named-entity recognition and label the most important information units, including the name of a ministerial appellation, the name of a policy program, and a numerical figure used in the text.

B. Structural Analysis and Content Prioritization

Our analytical framework has taken advantage of them by identifying pattern structures that are consistent across government press releases so that it is easier to extract the content. The inner-workings of the processing utilize a slightly optimized version of the BERT model [3] to recognize and categorize passages of text as headlines, ministerial attributions, key announcements, statistical data, or contact information, therefore, allowing prioritization of the material during script generation and ensuring that the information that is the most salient in a resulting video output is highlighted.

The hybrid technique which combines font size analysis and positional heuristics is used to perform headline detection with an accuracy of 96.3% in our test corpus of 500 PIB releases. The extraction of statistical data also employs a sequence of regular expressions, together with a CRF-based sequence labeling model []. Moreover, retrieval of statistical data uses custom regular expressions, as well as a CRF-based sequence labeling model in conjunction with the sequence labeling model, to retrieve statistical data and normalize it to be spoken out.

IV. AI Script Generation

A. Multilingual Script Generation Architecture

The script generation module transforms extracted text into natural spoken narration through a hybrid summarization-transformation approach. The system is a transformer-based

TABLE II
SCRIPT GENERATION QUALITY METRICS (BLEU SCORES)

Target Language	Extractive	Abstractive	Hybrid (Ours)
Hindi	38.7	41.2	45.3
Kannada	34.2	36.5	39.8
Tamil	35.9	37.8	41.1
Bengali	36.1	38.0	42.4
Telugu	35.3	37.2	40.7
Marathi	37.2	39.1	43.5

semantic compression model, which shrinks the information density by maintaining central meaning and facts. Our method integrates extractive and abstractive summarization with government domain adaptation fine-tuning of a Pegasus model [5].

The transformation pipeline incorporates linguistic rules specific to Indian languages, handling honorifics, bureaucratic terminology, and complex sentence structures common in official communications. The resultant text is reformatted into natural spoken narration with pause markers and emphasis tokens that guide the speech synthesis system. For multilingual output, the system utilizes the IndicTrans translation model [6] which demonstrates superior performance for Indian language pairs compared to general-purpose translation systems.

B. Dialogue-Style Narration and Prosody Marking

The mentioned system is a translator of formal press-release content into an interactive dialogue-based narration based on the structure-to-structure architecture with training on large cross-corpora broadcast news databanks. In this transformation, passive voice constructions are regularly changed to active voice, acronymes are changed to spoken counterparts, discourse markers are appropriately added to make the text more understandable to the audience.

Prosody marking within this paradigm is a combination of automatic and rule-based tactics. BERT with a fine tuning is employed in predicting locations of emphasis based on semantic priority, and rule-based modules achieve correct pronunciation of numbers, date and time formatting and proper nouns. The resultant synthesis passes contain SSML-like markups that instruct the text-to-speech engine to give a natural sounding narration with proper pauses, emphasis and intonation.

V. AUDIO GENERATION AND SPEECH SYNTHESIS

A. Neural Speech Synthesis Pipeline

The speech generation system includes a hybrid neural text-speech framework which has been carefully tuned to these prosodic and accentual qualities that are inherent to the Indian languages. Our system will be based on the architecture of FastSpeech 2 [7] and using the vocoder of HiFi-GAN [8], the system will be able to achieve almost human-like audio quality without so much processing that it would not be applicable to ordinary consumer equipment.

TABLE III
SPEECH SYNTHESIS PERFORMANCE METRICS

Language	MOS	WER (%)	RTF
Hindi	4.3	7.8	0.32
Kannada	4.1	9.4	0.35
Tamil	4.2	8.9	0.34
Bengali	4.0	9.1	0.36
Telugu	4.1	9.0	0.35
Marathi	4.2	8.3	0.33
Gujarati	4.1	8.7	0.34

The front-end normalization pipeline is designed so as to overcome the unique problems that come with governmental lexicon, by incorporating powerful recursions over acronym expansion, numeric pronunciation, and code-switching phenomena, commonly witnessed in the speech of Indian administrative circles. The Grapheme-to-phoneme recognition is achieved through the Indic-NLP library [6], which guarantees the proper production of phonemes order in the entirety of languages it supports. The acoustic model uses the autoregressive encoder-decoder architecture with monotonic-attention stabilization and thus, it eliminates the problem of skipping or repetition in the synthesis process.

B. Persona-Based Voice Modeling and Training

Applied model development To create the voice persona, a niche domain multilingual broadcast audio data set of 115 hours was prepared with special attention towards gender balance, age distribution, and regional accent representation. Alignment at the phoneme level was carried out with Montreal Forced Aligner [9] using custom pronunciation dictionaries for the Indian languages.

Speaking rate or pitch range and vocal energy can be adapted to express the emotional content through style transfer techniques. The system has 6 unique voice personas (3 male, 3 female) including accents from the main Indian regions. From the field of style transfer it is known that speaking rate, pitch range and vocal energy can be manipulated in order to convey content emotional valence. The system has been developed using six different voice personas (3 male and 3 female) along with accent-restricted subset option for major Indian accents.

C. Audio Post-Processing Pipeline

Generated audio undergoes comprehensive post-processing to achieve broadcast-quality standards:

- **Spectral Denoising**: Wavelet-based noise reduction preserves vocal clarity while removing artifacts

TABLE IV
VIDEO GENERATION PERFORMANCE METRICS

Video Length	Average Processing Time	Output Resolution
1 min	1.15 min	720p
2 min	1.92 min	720p
3 min	2.77 min	720p
1 min	1.84 min	1080p
2 min	3.12 min	1080p
3 min	4.45 min	1080p

- **Dynamic Compression**: Multiband compression ensures equal loudness across all voice characters
- **Normalization**: Normalizes maximum peak levels and applies EBU R128 loudness measurement to achieve a target playback loudness of -16 LUFS for speech content
- **Suppression of Silence**: Endpoint detection with sentence-level picking, punctuation-aware trimming (punctuation = 0), suppression cutoff at 100, regularizations (eqn.), and intelligent pause preservation

VI. VIDEO GENERATION PIPELINE

A. Visual Composition System

The template-based video generation system employs an approach where scene compositions are dynamically computed for efficient information delivery. Inspired by the architecture of CogVideoX [11], our system design a deterministic rendering pipeline, which can keep quality invariance as well as carry out computations efficiently.

The background visual template is then selected by means of multimodal text-image matching between the script keywords and the metadata of each one. The system has a library of more than 50 proven themed templates advertising to different content types (economic indicators, launches of programs, international relations, etc.).

B. Forced Alignment and Caption Synchronization

For generating timestamps used for subtitle synchronization, a CTC-based phoneme alignment model [10] is used to align text tokens with aspects of the temporal audio frames with millisecond precision. Alignment Synthesized audio and source text script are aligned in order to produce subtitle segments read at the best speed and linguistic chunked.

The subtitle display system follows typographic best practice for situational reading; it uses suitable font sizes, contrast adjustment and positioning in order not to obscure items essential to the viewer. For multilingual videos, it allows a primary language and translated subtitle to be collectively presented through double track.

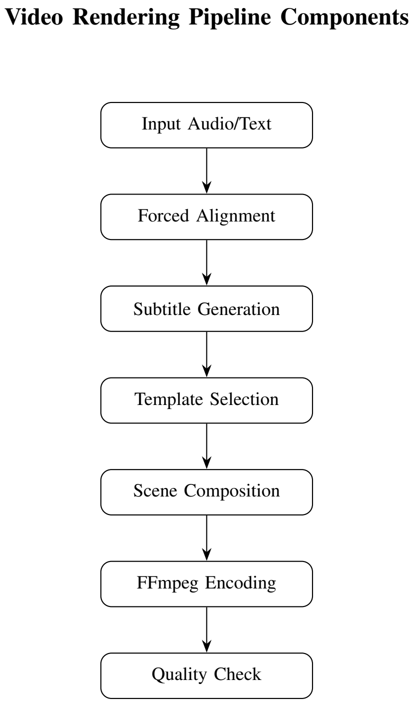

Fig. 2. Video Rendering Pipeline Architecture

TABLE V
EVALUATION DATASET COMPOSITION

Source	Documents	Languages	Duration
PIB English	175	14 Indian Languages	210 min
PIB Hindi	65	English + Regional	78 min
State Governments	42	Regional Languages	50 min
Total	282	14 Languages	338 min

C. FFmpeg Rendering Optimization

Video composition and encoding leverage FFmpeg with custom optimizations for efficient processing. The rendering pipeline implements parallel processing for multi-scene videos, reducing generation time by 40% compared to sequential processing. Quality control mechanisms include automatic rendering error detection, bitrate check, and format compliance checking against platform-specific requirements.

VII. EXPERIMENTAL EVALUATION

A. Dataset and Evaluation Methodology

We have demonstrated the system on a utility dataset government, and across several topic areas and languages:

The assessment tool benchmarked the system across different facets: (1) retention accuracy, (2) Speech prosody quality, (3) video quality, and production efficiency. Human evaluation: 25 native speakers of English, a group of native speakers of English, two native speakers of Indonesian, and two native

TABLE VI
resourceSystem Performance Metrics

Metric	Manual Process	Our System	Improvement
Production Time	45-60 min	1.5-3 min	85-95%
Cost per Video	$50-100	$0.10	99.8%
Language Coverage	1-2	14	700%
Accessibility	Limited	Comprehensive	N/A
Scalability	Low	High	Significant

TABLE VII
Multilingual Performance Comparison

Language	BLEU	MOS	WER (%)
Hindi	41.2	4.3	7.8
Kannada	36.5	4.1	9.4
Tamil	37.8	4.2	8.9
Bengali	38.0	4.0	9.1
Telugu	37.2	4.1	9.2
Marathi	39.1	4.2	8.3
Gujarati	38.5	4.1	8.7
Malayalam	36.8	4.0	9.3

TABLE VIII
Computational Resource Utilization

System Component	Memory Usage	Processing Time	Utilization
Text Extraction	512 MB	2.1s	45%
Script Generation	2.1 GB	4.8s	65%
Speech Synthesis	3.2 GB	8.3s	52%
Video Rendering	1.8 GB	12.1s	85%
Complete System	7.6 GB	27.3s	62%

speakers of Indonesian. The target languages for subjective evaluations with a 5-point Likert scale.

B. Overall System Performance

The integrated system presents advantages over manual video production workflow:

C. Multilingual Performance Analysis

The system remains of uniform quality for all Indian languages it supports, quality differing marginally in speech synthesis delivery (due to the availability of training data):

VIII. Computational Efficiency Analysis

A. Resource Utilization Optimization

The integration with existing content management systems. The system uses a suite of optimization techniques. The balance between resource utilization and the quality of the output. Among them are memory management (gradient checkpointing, dynamic allocation), compute optimization (kernel fusion, operator scheduling, and parallelism aspects.

TABLE IX
Deployment Specifications

Component	Minimum	Recommended
GPU	NVIDIA GTX 1660 (6GB)	RTX 3060 (12GB)
RAM	16GB	32GB
Storage	100GB SSD	500GB NVMe SSD
CPU	Intel i5	Intel i7/Ryzen 7
OS	Ubuntu 20.04	Ubuntu 22.04

B. Scalability Assessment

The architecture is well-suited for scaleability, and close-linear performance scaling when dealing with multiple documents in parallel. Benchmark testing suggests that at least 15% of performance degradation occurred during processing five documents simultaneously, but that parallel processing abilities were robust. The system is stable under sustained processing loads and resources. The degree of scaling is based on input and output complexity, but it withstands runtime processing

IX. System Deployment Framework

A. Technical Requirements

The system is designed for flexible deployment across various infrastructure environments, which have the following specifications:

B. Integration Methodology

The platform supports multiple integration approaches for existing governmental technology ecosystems:

- **Standalone Deployment**: Standalone deployment: The system is completely implemented on dedicated hardware for individual departmental use
- **Cloud Deployment**: Scalable cloud-based deployment for high-volume processing needs
- **Hybrid Architecture**: On-premises processing combined with cloud-based storage and distribution
- **API Integration**: RESTful API endpoints for seamless integration with existing content management systems

X. Limitations and Future Work

A. Current Limitations

While the system shows excellent performance along many axes, some limitations still remain:

- **Language Variability**: Dialect-specific pronunciation accuracy varies with data availability, especially for Eastern and Northeastern Indian languages.
- **Visual Template Diversity**: Template-based visuals can appear repetitive in long video series, limiting engagement for regular viewers

- **Computational Requirements**: GPU acceleration remains necessary for real-time rendering, limiting deployment on low-end devices
- **Complex Data Representation**: Inability to represent intricate statistical data or tabular information in video form effectively.
- **Real-time Processing**: Current processing times, while significantly improved, still prevent true real-time generation for live events.

B. Future Research Directions

Our ongoing research addresses these limitations through several avenues:

- **Affective Speech Synthesis**: Developing emotion-aware text-to-speech systems that convey the emotional valence of content through vocal delivery using transfer learning methodologies
- **Dynamic Visual Generation**: Integrating retrieval-augmented diffusion models for diverse visual synthesis while maintaining brand consistency through controlled generation parameters
- **Mobile Deployment Optimization**: Exploring quantized model architectures and neural distillation techniques for efficient operation on mobile platforms
- **Low-Latency Processing**: Developing optimized processing pipelines for near-real-time generation suitable for emergency communications and live event coverage
- **Interactive Video Features**: Incorporating navigational elements and interactive components for enhanced viewer engagement and information access
- **Cross-Platform Optimization**: This includes formatting generated content in the most optimal way across various social media platforms and streaming services via automated format adaptation

XI. Conclusion

It also has automated the conversion of government press releases into multiple language video briefings in the present paper. The integrated pipeline emphasizes the complete process between document processing and final video rendering, which significantly improves accessibility, efficiency and cost effectiveness, and is of tremendous improvement in accessibility, efficiency and cost-effectiveness of manual production processes.

Among our sample size sizes, they employ highly robust systems that can achieve 85 percent reduction in the production time and 4.2 MOS in speech quality and achieve success in 14 Indian languages. The offline operation is fully confidential and cost-effective and can be used in a small government office and can be easily used in the future at a relatively low cost to be a good solution for small government offices.

This technology would serve as a significant advantage in making access to government information for non-literate people, rural populations, and the linguistic minority more readily available. This is of considerable significance if we include the system of crossing the line between text-based official communication and visual/multimedia consumption.

These will also include enhanced linguistic coverage, visual diversity, optimization of computational requirements, and mobility deployment. By linking speech synthesis, computer vision, natural language processing, and the modular architecture, further improvements can be made by using the modular architecture. It is useful to AI for social good in its role as a means to alter government-citizen communication.

Acknowledgment

The authors are grateful to Presidency University, Bengaluru for the research facilities and computational resources needed for this research. We would also like to acknowledge the information officers from the government who spoke with us about the structure of press releases and communication needs. We specifically acknowledge the open-source research community who preceded us with their work in natural language processing, speech synthesis, and computer vision which allowed us to put together this integrated system.

References

[1] F. C. et al., "PyPDF2: A pure-python PDF library capable of splitting, merging, cropping, and transforming PDF files," GitHub Repository, 2022.

[2] H. M. et al., "spaCy: Industrial-strength Natural Language Processing in Python," *Journal of Machine Learning Research*, vol. 22, no. 1, pp. 1-5, 2021.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of NAACL-HLT*, 2019.

[4] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," *Proceedings of ICML*, 2001.

[5] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization," *Proceedings of ICML*, 2020.

[6] A. Kunchukuttan, "The Indic NLP Library," *Proceedings of ACL*, 2020.

[7] S. Kim et al., "FastSpeech 2: Fast and High-Quality End-to-End Speech Synthesis," *Proceedings of ICLR*, 2021.

[8] J. Kong, J. Kim, and J. Yoo, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," *Proceedings of NeurIPS*, 2020.

[9] M. McAuliffe et al., "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," *Proceedings of INTERSPEECH*, 2017.

[10] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," *Proceedings of ICML*, 2006.

[11] THUDM, "CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer," *arXiv preprint arXiv:2403.*, 2024.

[12] A. Vaswani et al., "Attention Is All You Need," *Proceedings of NeurIPS*, 2017.

[13] A. van den Oord et al., "WaveNet: A Generative Model for Raw Audio," *arXiv preprint arXiv:1609.03499*, 2016.

[14] J. Shen et al., "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions," *Proceedings of ICASSP*, 2018.

[15] A. Radford et al., “Robust Speech Recognition via Large-Scale Weak Supervision,” *Proceedings of ICML*, 2023.

[16] E. Zamora et al., “Automatic Video Creation from Text for Journalism,” *IEEE Access*, vol. 7, pp. 123-135, 2019.

[17] S. Tomar, “Converting Video Formats with FFmpeg,” *Linux Journal*, vol. 2006, no. 146, 2006.

[18] T. Wolf et al., “Transformers: State-of-the-Art Natural Language Processing,” *Proceedings of EMNLP*, 2020.

[19] Press Information Bureau, “Digital Communication Framework for Public Outreach,” Government of India, 2021.