



RESEARCH ARTICLE

OPEN ACCESS

SIGNSPEAK: AN AI-BASED UTILITY FOR VISUALLY IMPAIRED INDIVIDUALS

Sharanjit Kaur¹, Manju Bhardwaj², Kriti Misra³ and Preeti Marwaha^{*4}^{1, 3, 4}Department of Computer Science, Acharya Narendra Dev, College, (University of Delhi), New Delhi, India.²Department of Computer Science, Maitreyi College, (University of Delhi), New Delhi, India.¹<https://orcid.org/0000-0002-9437-5633>, ²<https://orcid.org/0000-0002-8696-3953>³<https://orcid.org/0009-0006-7957-3558>, ^{4*}<https://orcid.org/0009-0000-7425-4749>E-mail: sharanjitkaur@andc.du.ac.in, mbhardwaj@maitreyi.du.ac.in, kritimisra87@gmail.com, *preetimarwaha@andc.du.ac.in

ARTICLE INFO

Article History

Received on: March 05, 2026

Reviewed on: April 05, 2026

Accepted on: July 01, 2026

Published on: July 31, 2026

*Keywords:*Visually Impaired Assistance,
Text-To-Speech,
Signspeak,
Yolov11s,
Gemini.

ABSTRACT

Visually impaired people (VIP) need assistive technology for smooth and confident navigation on pavements along the roads. VIPs find it hard to follow the speech generated through a text-to-audio converter due to the presence of advertisement text. Automating the process of audio conversion from the signboards by integrating the capabilities of emerging AI models in a fraction of a second will be useful for their confident navigation. In this context, SignSpeak, an AI-based assistive utility, is designed to convert signboard text to speech using visual inputs, which may be captured through smart sensors or canes. The proposed system follows a three-step (detection-extraction-conversion) pipeline. Initially, a fine-tuned lightweight YOLOv11s model is used to generate a bounded box for the relevant text region, which is input to Gemini model, a multimodal large language model for text extraction. Lastly, the extracted text is converted into audio using the Google Text-to-Speech tool. We fine-tuned the YOLOv11s model on 200 signboard images and tested its performance on an additional curated dataset of 42 real-world signboard images. Experimental findings indicate that the proposed integrated pipeline achieved superior performance compared to the solitary usage of the Gemini model for text extraction followed by its audio conversion.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

One basic need of every human being is to visit different places with comfort and ease, but this need becomes a challenge for a Visually Impaired Person (VIP). To begin with, the VIPs usually moved around with the help of a human guide, this help although reliable could not be afforded by all. Moreover, the human guide has a disadvantage of making the VIPs feel constrained due to lack of sense of freedom and privacy, making them feel like an assisted person rather than being independent. Alternatively, dogs were also trained to assist VIPs but this service also has its own challenges. Dog training is a tedious task and the dog itself requires a lot of care and attention (i.e. feeding, grooming and walking), which is even more difficult for VIPs [1]. Similarly, other assisting tools like white cane are extensively used but they provide a limited set of features [2]. With the advent of internet-based technologies, a variety of assistive solutions are available for VIPs. All these systems, services, devices and applications help them with their daily chores and provide them a safer mobility option. These are included under one umbrella term: Assistive Technology. In the 1960s, the term “assistive technology” was used for technologies designed to handle issues related to information transmission, navigation and orientation for the especially abled [2] [3]. The invention of smartphones has not only enhanced the connectivity for common man but has provided an excellent set of assistive tools by providing audio assistance through Google maps, road navigation applications for sensing hurdles etc. Inclusion of artificial intelligence technologies with smartphones has transformed the smartphone into an essential tool for most VIPs [4-6]. Modern technologies are now accessible to visually impaired people, encouraging greater independence, which include screen readers on smartphones, portable camera-based systems, electronic pens, Android assistant EyeMate, the “Self-Sufficient Smart Vision Stick” [7] etc. Each of these provide different functionality. While navigating, mobile application Google Lens can be used for capturing images but it reads out everything captured, which may not be useful for their navigation.

Google's Lookout uses computer vision and GenAI to make everyday tasks, such as reading documents, object identification etc. However, reading text from the signboards using available applications, while walking on a pavement or sidewalk is still a challenge because of its surroundings, text quality and its orientation. Among various assistive technologies, systems that can detect, read, and convert text into speech hold great potential for enhancing information accessibility and independence for the visually impaired [4] [8] [9]. YOLOv11s, Gemini, and Google Text-to-Speech (gTTS) are cutting-edge technologies whose capabilities can be integrated to address this need. We propose SignSpeak as a cutting-edge technology that follows a three-step (detection-extraction-conversion) pipeline to identify and read aloud the text on the sign boards. In the first step, YOLOv11s is used as a fine-tuned detector to detect the signboard region from captured images in real time. Subsequently, Gemini, a large language multimodal model, is deployed to extract the text from the detected regions [10]. Finally, gTTS, a high-quality text-to-speech converter, is used to convert recognized text into speech [11]. It is to be noted here that our focus is to reliably convert the fetched text from the signboard images to audio, once images are captured effectively from the smartphone integrated with an object detection sensor. The following are the major contributions of the proposed work:

- Fine-tuning of YOLOv11s model to extract desired text regions from the signboards after distinguishing between signboards and banners of varying shapes and sizes.
- Integration of three cutting-edge technologies viz. YOLOv11s, Gemini and gTTS for converting text from signboards into speech.
- Self-annotation of 200 signboard images used for fine-tuning of YOLOv11s as well as curating a separate test set of 42 signboard images to evaluate the performance of the pipeline.
- Deploying a Streamlit-based application namely SignSpeak1 with an easy-to-use interface for VIPs.
- Extensive experimentation to show the effective performance of the proposed integrated pipeline.

Organization of the paper: Section II reviews related work on current deep-learning based assistive technologies for VIPs following a brief introduction to existing traditional methods. Section III briefly describes the tools explored followed by details of the proposed methodology. Section IV reports the experimental results and analysis along with discussion, while the conclusion and future scope are given in Section V.

II. THEORETICAL REFERENCE

This section gives an overview of the related work on current deep-learning based assistive technologies for VIPs after a brief introduction to traditional methods.

II.1 TRADITIONAL METHODS

An early work by Shen and Coughlan [3] demonstrated a real-time sign-reading prototype for the identification of environmental signs followed by its audio conversion. It established the importance of visual to audio technologies for assisting the VIPs by providing speech-based feedback, audio translation and many more. However, these machine learning based approaches were using hand-crafted features along with rule-based algorithms and were effective in a controlled environment [4] [12].

II.2 DEEP LEARNING METHODS

Recent advancements in deep learning-based approaches have enabled assistive technology for real time text detection, recognition and scene understanding. Utilizing these approaches has empowered the real-time object detection and navigation system, smartphone applications to improve the accessibility for VIPs [4]. Reading text from complex images has been made possible for VIPs by identification of textual regions and using deep learning techniques to identify text, and ultimately converting to speech. A hybrid approach combining deep learning with particle swarm optimization proposed by Pandey et al., recognizes text from natural scenes and an audio output is produced via text-to-speech converter [13]. Ganeshan et al. employ CNN-LSTM models to detect the caption associated with the given image, the identified caption is then converted into voice message using Text-to-Speech API. [14]. A ResNet-based system for object detection and text recognition by Begum et al. produces an audio output, and the visually impaired individuals can easily use this to read medication labels etc. [15] In [16], the authors design a system to assist visually impaired or text-reading challenge users by employing VGG-16 for text detection and recognition followed by Augmented reality visualization to display the content as a 3D image on a smartphone screen. One of the notable advancements is the OCR integration with the Text-to-Speech (TTS). A student-led initiative using the CNN-LSTM models on the embedded hardware achieved over 93% recognition accuracy from the live camera feeds and has affirmed the feasibility of deploying such pipelines on the resource-constrained devices [12]. Research has also highlighted the importance for the user experience during the audio rendering by preserving the structural elements to improve comprehension [5]. Similarly, work in Augmented reality has shown promise for vocalizing visual symbols [6]. Suresh et. al. proposed a wearable assistive system for the real-time auditory guidance in the form of a smart backpack by integrating sensing and on-device intelligence to be used in a controlled environment [17]. Recently, Moreira et. al. presented an integrated mobile application using Google Gemini artificial intelligence model for the VIPs to understand visual content [18]. Although tremendous efforts are being made for the effective use of deep-learning based assistive systems, challenges still exist in terms of their applicability and robustness. In this article, we are focusing on real-time capturing of text from the signboards and its reliable audio conversion for the smooth navigation of VIPs on the sidewalks and pavements.

III. MATERIALS AND METHODS.

This section briefly outlines the selection process for deciding the AI models and tools effective for getting the desired outcome in Sec. III.1, followed by description of methodology in Sec. III.2.

III.1 SELECTION OF AI MODELS AND TOOLS

III.1.1 Text Detection Models

The first step in the proposed pipeline involves accurate detection of text regions from the signboard images. Different tools were explored and their effectiveness was analysed. Traditional OCR engines like Tesseract [19] failed to work with low-resolution images and complex backgrounds. PaddleOCR [20] could detect and recognize text but showed inconsistencies in accessing small-sized text in the signboard images. Deep-learning based detectors like EAST [7] and CRAFT [16] provided better output, but produced imprecise bounding boxes and suffered from compatibility issues with modern frameworks. Finally, YOLOv11s, a lightweight model, was selected for object detection because of its localization accuracy, and seamless compatibility with current deep learning libraries. Also, Khanam and Hussain demonstrate that appropriate fine-tuning of YOLOv11s for a specific task produces precise bounded boxes, which ensures reliable input for the subsequent phase [21]. However, for proper fine-tuning, the input images in the YOLO model need to be annotated, hence, an open-source platform Label Studio is used to generate bounding boxes for signboard regions [22].

III.1.2 Text Recognition and Extraction Models

After the detection of text regions, multiple recognition frameworks were evaluated for their text recognition capability. During the early stages, PaddleOCR offered integrated detection and recognition but showed limited accuracy on small datasets and lacked robust cropping mechanisms [20]. EasyOCR [23], which is built on the CRAFT algorithm, supports multilingual scripts but produces inconsistent results, particularly with Tamil signboards which are present in the dataset. More advanced solutions such as LLAMA OCR 3.2-90B Vision, demonstrated strong recognition under complex conditions, but is constrained by high deployment costs and reliance on external APIs [24]. Finally, Gemini 3 was adopted as the recognition engine because of its high accuracy across various inputs, its support for multilingual scripts, and seamless integration with YOLOv11s [10]. This integration enables an efficient end-to-end text reading pipeline for fast inference required for real-world deployment.

III.1.3 Text-to-Speech Conversion Tools

Initially, the pyttsx3 engine for English synthesis was explored, as it provides offline functionality with adjustable voice and rate settings [25]. However, its limited applicability and quality response for non-English languages were the major concerns. We finally selected gTTS (Google Text-to-Speech) that not only covers these limitations but is also compatible with diverse linguistic contexts [26]. In addition to these tools, different Python-based libraries are used for the development, training and deployment of the proposed tool SignSpeak. Python Imaging Library (PIL) is used for the image conversion and the augmentation required for the dataset preparation [27]. OpenCV is used for the image processing task for delivering consistent input formats for the detection and recognition [28]. The Ultralytics PyTorch framework is used for the fine-tuning of YOLOv11s, offering a modular API for detection tasks [7]. Finally, the Google Generative AI SDK is incorporated for integrating the Gemini 3 via API calls [10].

III.2 PROPOSED METHODOLOGY

The methodology used for developing the SignSpeak encompasses data preparation, and integration of three major AI-based technologies for text detection, text extraction and text-to-speech conversion. The proposed approach (detection-extraction-conversion) takes images of signboards as input, detects texts on the signboards, followed by their conversion to speech, which is delivered as output to the visually impaired individuals.

III.2.1 Data Preparation and Labeling

A publicly available dataset, namely the Indian Signboard Image dataset consisting of 200 images from Kaggle, is used [29]. The signboard images contain multilingual text (English, Hindi, and Tamil) under varying lighting, orientations, and background conditions (see Figure 1). Label Studio is used to manually annotate all the images, drawing bounding boxes around text regions to generate high-quality ground-truth data in the YOLO format [7]. In addition, 200 background images without signboards are also incorporated into the dataset to enhance the detection robustness of the model.



Figure 1: An Annotated Sample Sign Board.
Source, Author, (2026).

These images with the road scenes are included as background (negative samples) without bounding box annotations [30]. The inclusion of background-only images helps the model learn contextual features and reduces false-positive detections during inference. This resulted in a combined dataset of 400 images, on which the pipeline is built and validated. An independent test dataset of 42 signboard images is also curated to evaluate the performance of the proposed pipeline.

III.2.2 Fine-Tuning of YOLOv11s Detector for Text Detection

YOLOv11s is highly effective for object detection but needs to be fine-tuned to road signboards to incorporate domain-specific optimizations [31]. The model is fine-tuned for Indian roadside sideboard detection using a combined dataset of 400 images (200 signboards + 200 road images) as discussed in Sec. III.2.1. It is to be noted that a few signboard images have multiple signs. Hence, Label Studio used for annotation returned 215 annotated objects for 200 signboards. The image dataset is split into training and validation sets with an 80:20 ratio, resulting in 320 training images and 80 validation images. The training set with 320 images consists of 166 signboards consisting of 178 annotated objects, and 149 background images respectively. The validation set contained 34 images with 37 annotated instances of the signboards and 46 background images as negative samples, indicating background-only scenes without visible road signs. Including such background images helped improve the detector's robustness by reducing false-positive predictions. Fine-tuning is conducted for 60 epochs with a batch size of 8 images, where each image resolution is set to 640×640 using the Ultralytics framework in a GPU-accelerated Google Colab environment. The model is trained for a single class, *Road Signboard*, using YOLO *txt* format annotations in the presence of a few background images. Data augmentation techniques such as horizontal flipping, mosaic augmentation, and RandAugment are also enabled to improve the generalization of the model.

Table 1: Distribution of signboard images and corresponding annotated objects along with background images used for training and validation during the fine-tuning of YOLOv11s model.

Signboard Data	Total Signboards (a)	Total Annotated Objects in Signboards	Total Background Images (b)	Total Images (a+b)
Image Instances	200	215	200	400
Training Data	166	178	154	320
Validation Data	34	37	46	80

Source: [29], [30].

III.2.3 Pipeline of SignSpeak

SignSpeak employs a three-stage pipeline: image detection module, text extraction module, and text-to-speech conversion module. The input image is first analyzed to detect the signboards, followed by text recognition and extraction, and lastly, the text is converted to audio, as illustrated in Figure 2.

III.2.3.1 Image Detection Module

The proposed pipeline begins with an object detection module, which uses a lightweight YOLOv11s model. The used model is fine-tuned specially for signboard detection. Its capability to capture localized regions of interest followed by generating bounding boxes around the signboards, makes it suitable for poor-quality images too. Its lightweight architecture allows for real-time inference without impacting the accuracy of the retrieved localized region.

III.2.3.2 Text Extraction Module

The localized regions are further processed by a text recognition AI model leveraging natural language prompts to extract textual content. Using the Gemini 3 model not only allowed for handling of various font styles, but its human-readable prompt returned the text without requiring task-specific fine-tuning. The integration of a vision-language model in this context demonstrates the advantage of natural prompt engineering, allowing SignSpeak to seamlessly handle multilingual signboards and varied environmental conditions, including distorted images, low resolution, or background noise, by integrating YOLOv11s with Gemini 3 to ensure high reliability in practical deployment scenarios.

III.2.3.3 Text-to-Speech Conversion Module

The SignSpeak pipeline concludes by transforming the recognized text into natural-sounding speech using a text-to-speech engine to get an accessible auditory output. The selection of the Google text-to-speech (gTTS) tool for providing a natural, multilingual speech output makes the proposed pipeline suitable for delivering clear feedback for texts written on signboards.

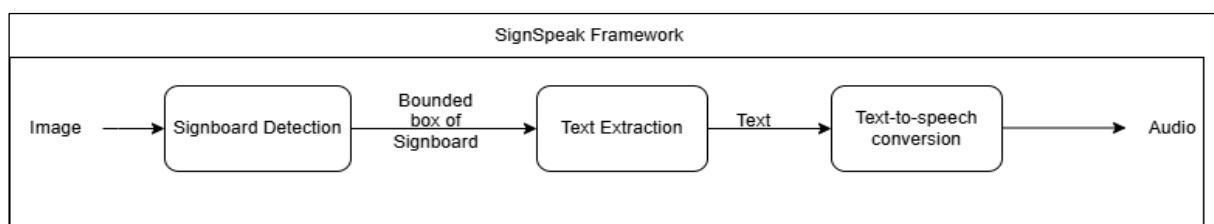
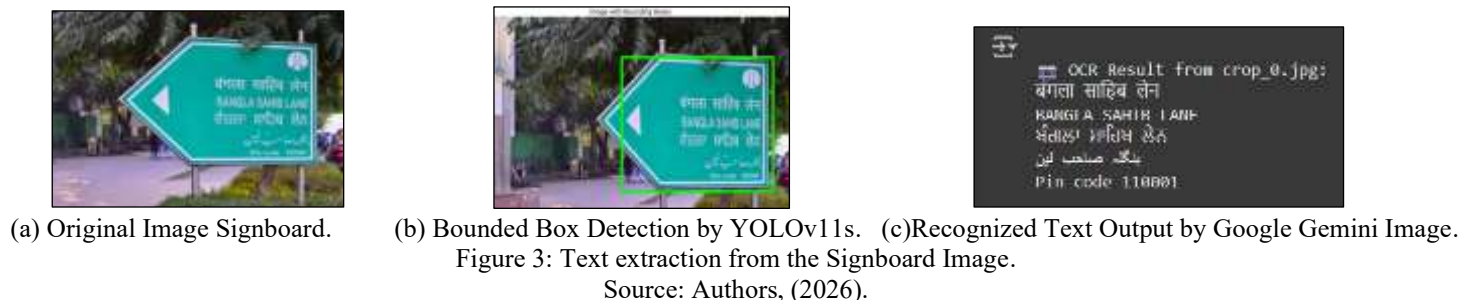


Figure 2: Pipeline of Three Phases in the Proposed Signspeak Framework for Extracting Text from Signboard for Text-to-Speech Conversion.

Source: Authors, (2026).

III.2.3.4 Example: Inference Process

Figure 3 demonstrates the workflow for processing a single image for text extraction. The input image (Fig. 3a) is passed as input to the fine-tuned YOLOv11s model, which returns the coordinates of a bounding box for the detected signboard (Fig. 3b). The bounded box is used for cropping the image so that only a valid portion is sent to the Gemini model with a prompt asking it to extract the text (Fig. 3c). The resulting text is stored for subsequent audio conversion by the text-to-speech conversion module.



IV. RESULTS AND DISCUSSIONS

This section describes the experiments conducted to evaluate the performance of the proposed AI-based utility tool using the established performance metrics for evaluating text extraction from the signboard and subsequent text-to-audio conversion efficiency.

IV.1 FINE-TUNING AND PERFORMANCE EVALUATION OF YOLOV11S MODEL

During YOLOv11s fine tuning for signboard detection, visual outputs from selected batches were saved to monitor the evolution of the model's detection performance. A batch refers to a group of images processed together before updating the model parameters, with the batch number indicating the sequence of processing during training. Recall that a batch of 8 images are used for the experimentation.

IV.1.1 Fine Tuning of the YOLO Model for Domain Specific Object Detection

The dataset consisting of 400 images is split into an 80:20 ratio, where 320 images are allocated for training and the remaining 80 images are reserved for validation (see Sec. III.2.2: Table 1 for details). The model is trained for 60 epochs with a mini-batch size of 8 images, allowing the training process to iteratively update model parameters across the entire training set multiple times to enable an effective domain-specific learning and convergence. Selection of parameters is done such that all the training samples are processed in the manageable batches to get stable gradient updates.

IV.1.2 Label Distribution and Bounding Box Analysis

Figure 4 illustrates the label distribution and bounding box distribution. The model identifies 178 annotated objects from a training set of 320 images (166 signboards + 154 background images) as shown in Figure 4a. Recall that the model is used for identification of only one class, which is Road Signboard. Figure 4b shows all bounding boxes overlaid, where each rectangle is one labeled object. All boxes are drawn on the top of each other, indicating their concentration on the middle part of the signboards. Figures 4c and 4d show the object center positions (x v/s y) and distribution of width v/s height of all boxes respectively. It is clear from these two figures that most signboards are mostly near the center and are having size variation from small to medium respectively.

IV.2.2 Training and Validation Performance of The YOLO model

Figure 5 shows the training and validation performance of the YOLO model where the x-axis represents the number of epochs and the y-axis represents the score of the selected metric. The first three plots in top and bottom rows show the performance of the models on the training and validation set in terms of loss respectively. The first and second plots in the top row show the loss that incurs during learning of bounding boxes and signboard class during different epochs. It is clear from the first two plots that learning is poor initially which improves with increasing epochs. Low loss with increasing epochs vindicates the good learning of the model. A similar pattern is observed in the third plot of the top row and revealed that advanced bounding box fine tuning loss also decreases with epochs. The first three plots in the bottom row of Figure 5 shows the performance of the model on the validation set. Spikes in early epochs may have contributed to minor label normalization inconsistencies in the validation data resulting in high classification error in these epochs. The fourth and fifth plots in both top and bottom rows show the Precision, Recall and mAP (Mean Average Precision) of the trained model on the validation set. It is clear that the Precision and Recall stabilize at approx. 0.92 and 0.94 respectively, capturing the effective learning of the model on the provided data. During early epochs, the model is unstable, which causes high classification error. Once training progresses and the model adjusts to the data, the loss quickly stabilizes and decreases as vindicated by the mAP plots shown as last two plots in the bottom row. The metric mAP50 considers a prediction accurate if it overlaps with the ground truth bounding box by at least 50%, which is around 0.92. The last box shows the metric mAP90-95, which checks for 90-95% overlap in the detected and reference images and is around 0.6 which is good for the object detection model.

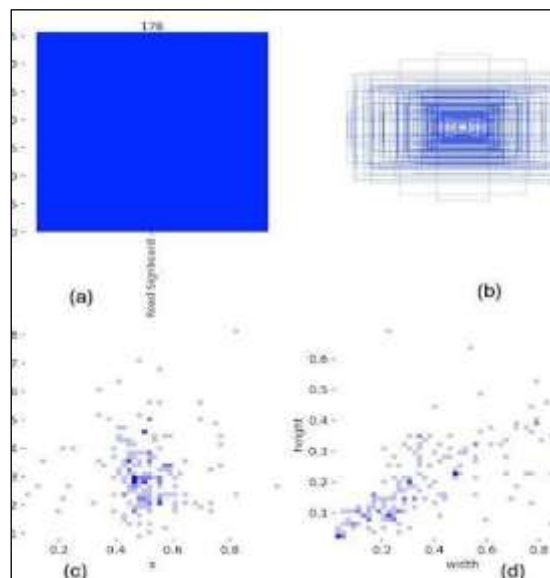


Figure 4: Label Distribution and Bounding box analysis.
Source: Authors, (2026).

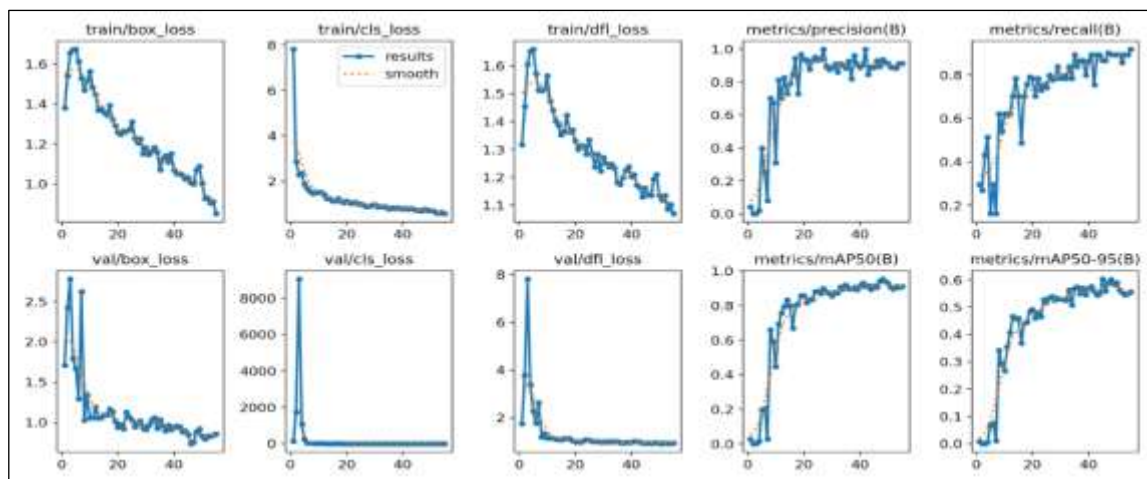


Figure 5: Training and Validation Performance of YOLOv11s model.
Source: Authors, (2026).

Table 2: Confusion Matrix showing Performance of YOLO model on the Validation Set.

Predicted	Actual	
	Road Signboard	Background
Road Signboard	34 (TP)	7 (FP)
Background	3 (FN)	NA

Source: Authors, (2026).

For the single-class *Road Signboard* detection task, the confusion matrix in Table 2 shows the accuracy of the model in terms of TP (True Positive), FN (False Negative) and FP (False Positive) which are 34, 3, and 7 respectively. Out of actual 37 signboard objects, 34 are reported correctly and the rest 3 are reported as background. Also, 7 background images with no signboard are predicted as signboard images (FP). Here, the true negatives are not explicitly defined due to the enormous number of the possible background regions. The true positive detections correspond to correctly localized and classified road signs, while false negatives represent missed detections of annotated signs. False positives indicate cases where the model predicted a road sign in background regions. These results confirm that the model successfully detects the majority of road instances while maintaining reasonable discrimination against the background.

IV.2.3 Model Evaluation through Box and PR Curves

Figure 6 represents the Box and PR curves for the validation set. In all subfigures, the light blue curve captures the model's performance in detecting signboard class only whereas the dark blue curve shows the overall model's performance in the presence of background images without signboards in them. Figure 6a represents the F1 confidence curves where the x-axis represents the confidence and y-axis represents the F1 score. The curves depict the variation in model performance with change in the confidence threshold. It is observed that the model performance is best with F1-score 0.88 at confidence threshold ≈ 0.53 . Figure 6b describes the Precision-Confidence curve where the x-axis and y-axis show the confidence threshold and Precision respectively. It is clear from the figure that Precision is more than 0.8 for confidence > 0.3 and is maximum 1 for confidence value ≈ 0.73 .

It vindicates the model's capability to detect signboard images with high precision starting from low confidence value. Figure 6c validates the model's capability to detect images precisely by missing out few of the actual signboard images (see right side of the Precision-Recall curve where the X-axis shows recall and the Y-axis represents precision), which suggests model is able to detect most objects and make few mistakes at mAP@0.5. Detection of all classes with 0.916 vindicate the model's good performance in detecting sign board part from the images. Figure 6d represents the Recall-Confidence curve showing variation in Recall (y-axis) with change in confidence threshold (x-axis). It is observed that the recall decreases after the 70% confidence i.e. the model starts missing many objects. The YOLO model performs best in the low confidence range with recall above 70%. But its performance deteriorates with high confidence.

IV.2 PERFORMANCE EVALUATION OF THE SIGNSPEAK PIPELINE ON TEST SET

We use independent test data to evaluate the performance of the integrated approach in the following sections. For which, we separately evaluate the performance of each of the three modules using suitable metrics to prove its effectiveness on unseen signboard images.

IV.2.1 Evaluation of Signboard Detection from the Images

We evaluate signboard detection on the dataset with 62 images containing 42 real world images and 20 banners (with no signboards) images curated from the internet. All 62 images are used sequentially to evaluate its working. We set a threshold of 0.6 (60%) to reduce the false detection of banners as signboards. At this threshold, the model achieves a higher precision along with fewer false positives while maintaining a good recall as vindicated during fine tuning of the YOLO model. The model detected 38 signboard images out of 42 images as TP and the remaining 4 images were detected as background leading to False negative. It reported only one banner image falsely classified as Road Signboard (Table 3), NA values in TN cell held same justification for validation set. For the single-class *Road Signboard* detection task, the model performance with precision of 0.97 and a recall of 0.90 affirms its capability to capture the most of the signboards correctly with a minimal false detection.

Table. 3: Confusion Matrix showing Performance of YOLO model on the Test Set with 42 Signboard Images and 20 banner images.

	Predicted Signboard	Predicted Background
Actual Signboard	38 (TP)	4 (FN)
Actual Background	1 (FP)	NA

Source: Authors, (2026).

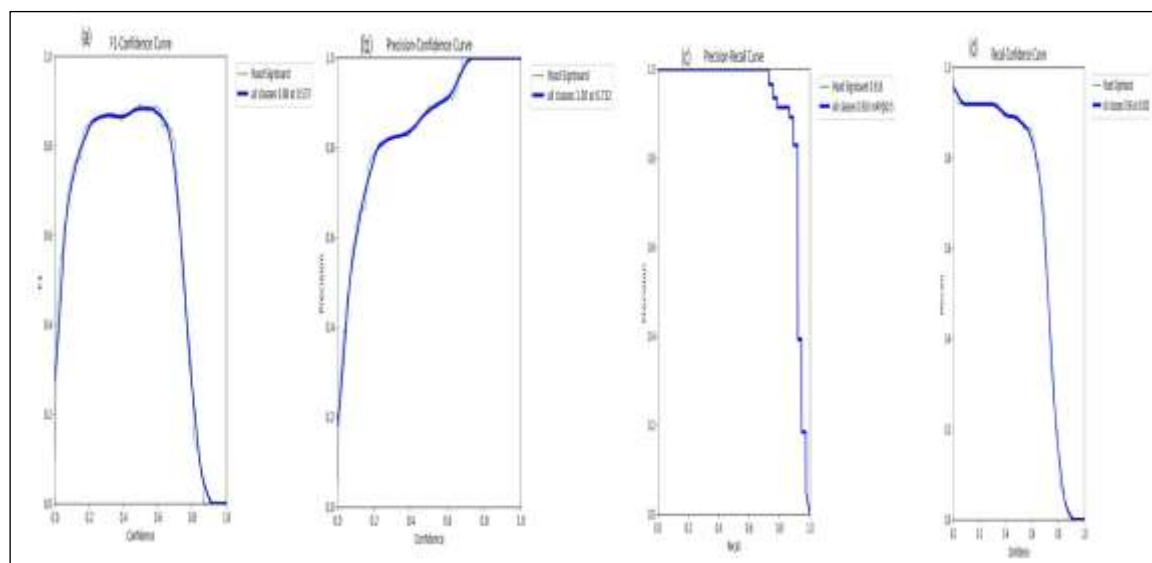


Figure 6: Performance Evaluation of YOLOv11s Model on the Validation Set using three Standard Metrics with varying Confidence thresholds.

Source: Authors, (2026).

IV.2.2 Evaluation of Text Extraction from the Detected Signboards

Model Gemini 3, which does not require fine-tuning, is used for text recognition. A prompt-based approach is used to retrieve the text from the images using Gemini. For example, a prompt: "What text do you see in this image?" returns the visible text in the provided image without any explanation. We manually extracted the text from all 42 signboard images in the test set and saved it as the Ground-Truth text, which is used subsequently for comparing the performance of the proposed pipeline (referred as Pipeline A) with the other standard method Pipeline B (Gemini 3 recognition only).

(a) Original SignBoard	(b) Pipeline A Output	(c) Pipeline B Output
	दीप चन्द बन्धु अस्पताल Deep Chand Bandhu Hospital बुनकर कालोनी Bunker Colony इन्दरलोक Inderlok दिल्ली सरकार	दीप चन्द बन्धु अस्पताल Deep Chand Bandhu Hospital बुनकर कालोनी Bunker Colony इन्दरलोक Inderlok दिल्ली सरकार DELHI GOVT. D.A. FLATS RESIDENTIAL COMPLEX
	जन पथ लेन JAN PATH LANE सं न पथ लेन PIN.COD.110001	**Sign on the far left:** * GENERAL WILLIAMS MASONIC POLYCLINIC * NOW AVAILABLE * WHOLE BODY CT SCAN FACILITY * TEL. 23746313, 23363073, 23342343 **Blue signs behind the white sign (from left to right):** * PHYSIOTHERAPY * X-RAY * ULTRASOUND * C.T. SCAN **Directional arrow sign in the foreground:** * **Hindi:** जन पथ लेन * **English:** JAN PATH LANE * **Punjabi:** ਜਨ ਪਥ ਲੇਨ * **Urdu:** جن پتھ لین * **Pin Code:** PIN. COD. 110001

Figure 7: Comparison of Text Extracted by Pipeline A (YOLO and Gemini 3) and Pipeline B (by Gemini only) with the Original Signboard Text.

Source: Authors, (2026).

It is to be noted that Pipeline A (YOLOv11s detection and Gemini 3 recognition) was designed to first localize signboard regions before performing text recognition, whereas Pipeline B (Gemini 3 recognition only) operated directly on the full images without prior detection. Figure 7 shows the comparison of both the Pipelines A (Fig 7 b) and B (Figure 7c) respectively w.r.t to two original signboard images (Figure 7a). It is observed that Pipeline A performed better by extracting the relevant text from the localized region of the signboard marked by YOLO, whereas line B extracted the entire text from the image provided making navigation of VIPs difficult. We also use BLEU [32] and ROUGE [33] metrics to compare the text extracted from these two pipelines (Table 4). For the Pipeline A, high value of performance metrics (BLEU score of 81.06, an average Rouge-1 F1 score of 0.93 and an average Rouge-L F1 score of 0.92) compared to those of Pipeline B (BLEU score of 60.33, an average ROUGE-1 F1 score of 0.85 and an average ROUGE-L F1 of 0.85) demonstrate the effectiveness of integrating a signboard detection stage prior to the recognition. The improvement can be attributed to YOLOv11s ability to isolate relevant text regions, thereby minimizing the background noise and distortions that often hinder the accuracy. The higher BLEU score confirms the improved n-gram precision, while the ROUGE score reflects the greater overlap with the Ground-Truth text, signifying enhanced completeness of recognition. Overall, the results provide quantitative evidence that a modular approach (Pipeline A), which uses a detection-recognition pipeline surpasses the recognition-only approach (Pipeline B) for the real-world scenarios. It is concluded from the results that the proposed integrated approach for text extraction is reliable and robust for practical deployment.

Table 4: Performance comparison of Pipeline A and Pipeline B in terms of Text Extraction using Test Set.

Pipeline	Bleu score	Avg Rouge-1 F1	Avg Rouge-L F1
A (Yolo+ Gemini)	81.06	0.93	0.92
B (Gemini only)	60.33	0.85	0.85

Source: Authors, (2026).

IV.2.3 Evaluation of Text-to-Speech Conversion

We use TTS-ASR (Text-to-speech Automated Speech Recognition), which uses a loop where text is converted into speech, followed by its conversion back to text again using a round-trip approach [34]. The round-trip time allowed for computing the Word Error Rate (WER) [35] and Character Error Rate (CER) [36], which were computed after text normalization (lowercasing, punctuation removal, and extra space removal) to ensure reliable comparison.

Then we calculated the WER and CER scores. For this phase, text captured from Pipeline A and Pipeline B from test images were converted to audio using Google Text-to-speech converter (gTTS), followed by its processing by the ASR pipeline using a round-trip approach. The benefit of using the round-trip approach is to check whether the system can reconstruct the original text accurately after passing through both speech synthesis and recognition stages. Here, Pipeline A clearly outperformed Pipeline B (Figure 8) by achieving a lower WER and CER, indicating its text-preserving capability in the TTS-ASR evaluation loop.

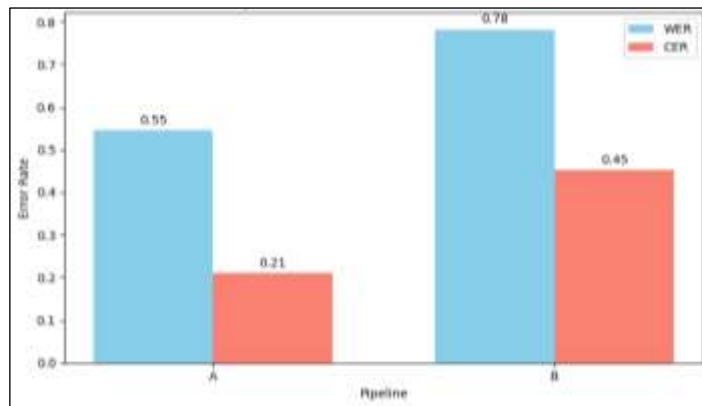


Figure 8: Comparison of WER and CER of Pipeline A and Pipeline B for Text-to-speech Conversion using Test Set.
Source: Authors, (2026).

IV.3 GUI FOR THE INTEGRATED SIGNSPEAK PIPELINE

A Streamlit-based application, namely SignSpeak is developed for the ease-of-use of users, whose interface is shown in Fig. 9. Once an image is detected by the sensor, the different phases of the proposed integrated framework are carried out in the background. For the validated signboard images, audio is generated to assist the commuters for safe mobility.



Figure 9: Lightweight GUI for Signspeak Utility.
Source: Authors, (2026).

IV.4 DISCUSSION

The experiments conducted in the former sections validate the efficacy of the integrated approach over the standalone usage of one model. The YOLOv11s performance demonstrates high accuracy and reliability even when it is trained on a small signboard dataset with manually annotated images. Its lightweight architecture enabled systematic inference while maintaining the detection quality. This balance between speed and accuracy makes the YOLOv11s suitable for the deployment for domain-specific tasks in low- resource environments, such as mobile or integrated systems [21]. Also, it improves the recognition performance of Gemini 3 in the text extraction phase, which is converted to reliable audio by gTTs. The integration of the detection-extraction-conversion phases is not only an efficient pipeline, but also delivers reliable output for navigation. The system maintains robustness under different conditions, including occlusion and resolution, that were incorporated while making the test set. The result highlights its potential for deployment as an assistive technology and its extension to multilingual accessibility because of the used Gemini 3 model.

V. CONCLUSION AND FUTURE SCOPE

SignSpeak introduces an assistive framework designed to enhance the awareness and navigation for visually challenged navigators on the pavements by detecting, recognizing and vocalizing the street/road signboards. This pipeline integrates the YOLOv11s model along with Gemini 3 for the high precision text recognition, ensuring the reliable performance in the real-world environments. To ease the usage, the recognized text is converted into speech using gTTS (google text to speech), providing natural auditory feedback to the users. Experimental evaluation affirmed the capability of the proposed integrated approach for text extraction over the stand-alone usage of Gemini model.

The proposed SignSpeak system for VIPs addresses the critical accessibility gap by enabling the visually challenged to access the street signs more efficiently and progress with greater reliance. As the current implementation of SignSpeak provides a strong foundation, several enhancements can be explored to broaden its application and usability. Inclusion of Multi-Object Detection along with important road elements like traffic lights, zebra crossings would enhance its applicability. Multi-language support may be included by extending the training data to support native languages (like Tamil, Hindi, Bengali, etc.) by leveraging the multilingual capabilities and global workflow of Gemini 3.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Sharanjit Kaur, Manju Bhardwaj, Kriti Misra and Preeti Marwaha.

Methodology: Sharanjit Kaur, Manju Bhardwaj, Kriti Misra and Preeti Marwaha.

Investigation: Sharanjit Kaur, Manju Bhardwaj, Kriti Misra and Preeti Marwaha.

Discussion of results: Sharanjit Kaur, Manju Bhardwaj, Kriti Misra and Preeti Marwaha.

Writing – Original Draft: Sharanjit Kaur, Manju Bhardwaj, Kriti Misra and Preeti Marwaha.

Writing – Review and Editing: Sharanjit Kaur, Manju Bhardwaj, Kriti Misra and Preeti Marwaha.

Resources: Sharanjit Kaur, Manju Bhardwaj, Kriti Misra and Preeti Marwaha.

Supervision: Sharanjit Kaur, Manju Bhardwaj, Kriti Misra and Preeti Marwaha.

Approval of the final text: Sharanjit Kaur, Manju Bhardwaj, Kriti Misra and Preeti Marwaha.

VII. ACKNOWLEDGMENTS

We acknowledge the research support provided by the Principal Acharya Narendra Dev College, New Delhi and financial assistance received from Anusandhan Kosh and DBT Star Scheme to carry out the proposed research work.

VIII. REFERENCES

- [1] M. Wahab, A. Helmy, A. A. Talib, H. A. Kadir, A. Johari, A. Noraziah, R. M. Sidek, and A. A. Mutalib, "Smart cane: Assistive cane for visually-impaired people," *International Journal of Computer Science Issues*, vol. 8, no. 4, 2011.
- [2] W. Elmannai and K. Elleithy, "Sensor-based assistive devices for visually-impaired people: Current status, challenges, and future directions," *Sensors*, vol. 17, no. 3, p. 565, 2017, doi: 10.3390/s17030565.
- [3] H. Shen and J. M. Coughlan, "Towards a real-time system for finding and reading signs for visually impaired users," in *Computers Helping People with Special Needs (ICCHP)*, vol. 7383, Springer, 2012, pp. 41–47.
- [4] A. Lavric *et al.*, "A comprehensive survey on emerging assistive technologies for visually impaired persons," *Sensors*, vol. 24, no. 15, p. 4834, 2024.
- [5] P. Mowar, T. Ganu, and S. Guha, "Towards optimizing OCR for accessibility," *arXiv preprint arXiv:2206.10254*, 2022.
- [6] J. Huang, M. Kinatader, M. J. Dunn, W. Jarosz, X.-D. Yang, and E. A. Cooper, "An augmented reality sign-reading assistant for users with reduced vision," *PLoS One*, vol. 14, no. 5, 2019.
- [7] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "EAST: An efficient and accurate scene text detector," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [8] M. Kannadhasan and M. M. S. Sree, "Machine vision as environmental narrator: A study of Seeing AI and Google Lookout," *International Journal of Innovative Research in Technology*, vol. 12, no. 8, 2026.
- [9] K. M. B. Laili, M. Z. Abdullah, S. Amiruddin, E. Sunardi, and J. Permatasari, "Integrating YOLOv8, EasyOCR, and gTTS for text detection in assistive technology for the visually impaired," *BIS Information Technology and Computer Science*, vol. 2, Apr. 2025, Art. no. V225008, doi: 10.31603/bistycs.185.
- [10] Gemini Team, "Gemini: A family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.
- [11] J. K. Lau, K. K. W. Kong, J. H. Yong, P. H. Tan, Z. Yang, Z. Q. Yong, J. C. W. Low, C. Y. Chong, M. K. Lim, and D. Lo, "Synthesizing speech test cases with text-to-speech? An empirical study on the false alarms in automated speech recognition testing," in *Proc. 32nd ACM SIGSOFT Int. Symp. Software Testing and Analysis (ISSTA)*, New York, USA, 2023, pp. 1169–1181, doi: 10.1145/3597926.3598126.
- [12] A. Prasanth, B. S. Kumar, and B. S. L. G. Varaprasad, "Portable camera based Text Reading of Objects for blind Persons", *International Journal of Applied Engineering Research*, vol 13, no. 17, pp. 12995-12999, 2018.
- [13] B. K. Pandey, D. Pandey, S. Wariya, G. Aggarwal, and R. Rastogi, "Deep learning and particle swarm optimisation-based techniques for visually impaired humans' text recognition and identification," *Augmented Human Research*, vol. 6, no. 1, p. 14, 2021.
- [14] J. Ganesan, A. T. Azar, S. Alsenan, N. A. Kamal, B. Qureshi, and A. E. H. Hassanien, "Deep learning reader for visually impaired," *Electronics*, vol. 11, no. 20, p. 3335, 2022.
- [15] A. Y. Begum, R. Krishnamoorthy, A. T. Gaikwad, N. Malarvizhi, D. S. Rao, and N. Adinarayana, "Visionary text recognition system for assisting the visually impaired in identifying medicines and products with hybrid deep learning," in *Proc. 3rd Int. Conf. Automation, Computing and Renewable Systems (ICACRS)*, 2024, pp. 1369–1374.
- [16] W. Yang, R. H. R. Zhong, and J. Lu, "CRAFT: Character region awareness for text detection", in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [17] N. M. Suresh, A. S. M. M., and R. P. Nair, "An assistive device for the visually impaired," in *Proc. Int. Conf. Power, Instrumentation, Control, and Computing (PICC)*, Thrissur, India, 2025, pp. 1–7, doi: 10.1109/PICC67314.2025.11291569.
- [18] F. W. R. Moreira and G. Hermes, "Development of a cross-platform mobile application using Gemini to assist visually impaired individuals," 2024.
- [19] R. Smith, "An overview of the Tesseract OCR engine," in *Proc. 9th Int. Conf. Document Analysis and Recognition (ICDAR)*, vol. 2, 2007, pp. 629–633.

- [20] PaddleOCR Team, “PaddleOCR 3.0 technical report,” *arXiv preprint arXiv:2507.05595*, 2025.
- [21] R. Khanam and M. Hussain, “YOLOv11: An overview of the key architectural enhancements,” *arXiv preprint arXiv:2410.17725*, 2024.
- [22] M. Tkachenko, M. Malyuk, A. Holmanyuk, and N. Liubimov, “Label Studio: Data labeling software,” Heartex, Open-source tool, 2020–2025.
- [23] M. A. M. Salehudin, S. N. Basah, H. Yazid, *et al.*, “Analysis of optical character recognition using EasyOCR under image degradation,” *Journal of Physics: Conference Series*, 2023.
- [24] J Chi, U Kam, H Zhan, E Smith, J Rando, Y Zhang, K Plawiak, ZD Coudert et. “Llama Guard 3 Vision: Safeguarding human-AI image understanding conversations,” *arXiv preprint arXiv:2411.10414*, 2024.
- [25] N. M. Bhat, “pyttsx3: Text-to-speech cross-platform library,” 2017. [Online]. Available: <https://pyttsx3.readthedocs.io>.
- [26] P. N. Durette *et al.*, “gTTS: Google text-to-speech Python library,” 2017. [Online]. Available: <https://github.com/pndurette/gTTS>.
- [27] S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu, “scikit-image: Image processing in Python,” *PeerJ*, vol. 2, p. e453, 2014, doi: 10.7717/peerj.453.
- [28] G. Bradski and A. Kaehler, *Learning OpenCV: Computer Vision with the OpenCV Library*. Sebastopol, CA, USA: O’Reilly Media, 2008.
- [29] DataClusterLabs, “Indian signboard image dataset,” Kaggle, 2024. [Online]. Available: <https://www.kaggle.com/datasets/dataclusterlabs/indian-signboard-image-dataset>
- [30] “Indian Driving Dataset (IDD),” Kaggle, 2024. [Online]. Available: <https://www.kaggle.com/datasets/mitanshuchakrawarty/new-idd-dataset>.
- [31] C. Li, S. Liu, and Z. Peng, “ESA-YOLO: An efficient scale-aware traffic sign detection algorithm based on YOLOv11 under adverse weather conditions,” *PLoS One*, vol. 20, no. 11, p. e0336863, 2025.
- [32] E. Reiter, “A structured review of the validity of BLEU,” *Computational Linguistics*, vol. 44, no. 3, pp. 393–401, 2018.
- [33] C.Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” in *Proc. Workshop on Text Summarization Branches Out*, Barcelona, Spain, 2004, pp. 74–81.
- [34] Y. Ren, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.Y. Liu, “Almost unsupervised text-to-speech and automatic speech recognition,” *arXiv preprint arXiv:1905.06791*, 2019.
- [35] A. Ali and S. Renals, “Word error rate estimation for speech recognition: e-WER,” in *Proc. 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia, 2018, pp. 20–24.
- [36] C. Park, H. Kang, and T. Hain, “Character error rate estimation for automatic speech recognition of short utterances,” in *Proc. 32nd European Signal Processing Conf. (EUSIPCO)*, Lyon, France, 2024, pp. 131–135.