Natural Language Generation for Assistive Vision Technologies

Authors: Rodriguez, J., Chen, S., Williams, D., Kim, D., & Patel, A.
Date: 2024
Institution: iVision AI Research Division
Contact: research@ivisionai.org

Abstract

This paper presents a novel approach to natural language generation (NLG) for assistive vision technologies designed specifically for blind users. We introduce ContextSpeak™, an adaptive NLG system that produces contextually relevant, spatially accurate descriptions of visual scenes. Our system implements a hierarchical priority framework that ranks information based on navigational importance, spatial proximity, and user preferences. Evaluations with 56 blind participants demonstrate significant improvements in environmental comprehension, navigation efficiency, and user satisfaction compared to existing systems. Results show that context-aware language generation with spatial precision enables more intuitive understanding of environments, addressing a critical gap in current assistive technologies.

1. Introduction

Translating visual information into useful verbal descriptions represents a significant challenge for assistive technologies serving blind users. While computer vision systems have made remarkable progress in object detection and scene recognition, the conversion of this visual data into natural language remains an underdeveloped area—particularly for navigation assistance applications.

Existing approaches to natural language generation in assistive vision typically focus on image captioning paradigms derived from general-purpose computer vision applications. However, these approaches fail to address the specific informational needs of blind users navigating physical environments. General-purpose image captions often emphasize visual aesthetics or general scene composition rather than spatial relationships and navigational hazards that are crucial for independent mobility.

This paper addresses this gap by presenting ContextSpeak™, a natural language generation system specifically designed for navigation assistance. Our approach differs from traditional image captioning in several key ways:

1. Prioritizing information based on navigational relevance
2. Incorporating precise spatial coordinates and distances
3. Adapting to user movement and environmental changes
4. Personalizing outputs based on individual preferences and needs

We demonstrate that these adaptations significantly improve the usefulness of verbal descriptions for real-world navigation, enabling more confident and independent mobility for blind users.

2. Related Work

2.1 Image Captioning for Accessibility

Traditional image captioning models have been applied to accessibility applications with mixed results. Anderson et al. (2018) presented a bottom-up and top-down attention mechanism for image captioning that improved description quality, but their approach focused on general visual content rather than navigational information.

Wang et al. (2021) adapted image captioning specifically for blind users by emphasizing object identification but did not address spatial relationships adequately. Similarly, Zhang and Li (2022) proposed a system for detecting accessibility-relevant objects but generated descriptions lacking the spatial precision necessary for navigation.

2.2 Spatial Language in Assistive Technologies

The importance of spatial language in assistive technologies has been recognized by several researchers. Mascetti et al. (2020) explored how blind users interpret verbal spatial directions, finding that precise distance information and consistent reference frames significantly improved navigation performance.

Kumar et al. (2022) demonstrated that blind users strongly prefer directional information that references their body orientation (e.g., "2 meters ahead" rather than "north") and benefit from standardized distance units. However, their implementation did not integrate with real-time object detection or adapt to changing environments.

2.3 Context-Aware Description Generation

Context awareness in language generation has been explored by Hersh and Johnson (2023), who developed a system that adapts descriptions based on user history and environmental familiarity. Their work showed improvements in user satisfaction but did not focus on real-time navigation assistance.

Singh et al. (2021) proposed a context-sensitive description framework that considered user tasks, but their approach required explicit task specification rather than adapting automatically to navigation contexts.

3. System Design

ContextSpeak™ implements a comprehensive approach to natural language generation for navigation assistance. Figure 1 illustrates the overall architecture of the system.

3.1 Information Prioritization Framework

The core innovation of ContextSpeak™ is its hierarchical prioritization framework that determines what information to include in verbal descriptions and in what order. This framework consists of four priority levels:

3.1.1 Level 1: Critical Safety Information

The highest priority is assigned to information that directly impacts user safety:

- Obstacles at head or foot level
- Approaching stairs, drops, or level changes
- Moving objects on collision course
- Sudden environmental changes

These elements are always communicated first, with explicit distance measurements and urgency markers when appropriate (e.g., "Caution: Step down 2 feet ahead").

3.1.2 Level 2: Primary Navigation Elements

The second priority level focuses on elements that define the navigable space:

- Pathways and corridors
- Doors and entrances/exits
- Major landmarks for orientation
- Significant boundaries (walls, fences)

These elements provide the structural framework for navigation and are described with clear spatial relationships to the user (e.g., "Doorway 3 meters ahead, slightly to your right").

3.1.3 Level 3: Functional Objects

The third priority level includes objects that may be functional targets:

- Furniture and fixtures
- Interactive elements (buttons, handles)
- Signage and information displays
- People and service points

These elements are described with their functional affordances when relevant (e.g., "Chair 2 meters to your left, facing away from you").

3.1.4 Level 4: Contextual Information

The lowest priority level includes additional contextual information:

- Aesthetic features
- Ambient conditions
- Historical or cultural context
- Non-essential objects

This information is included only when higher-priority elements have been addressed and verbosity settings permit (e.g., "Room has bright natural lighting from windows on the right wall").

3.2 Spatial Reference System

ContextSpeak™ implements a user-centered spatial reference system that:

1. Maintains the user as the primary reference point
2. Provides distances in standard units (feet/meters)
3. Uses clock-face directions for angular precision when appropriate
4. Adapts reference frames based on user movement

The system converts absolute spatial coordinates from computer vision into relative positions that remain intuitive during navigation. For example, as a user approaches an object, descriptions automatically update from "Table 10 feet ahead" to "Table 5 feet ahead" to "Table directly in front of you."

3.3 Language Generation Architecture

The language generation component employs a template-based approach augmented with dynamic elements to balance consistency with natural-sounding output. Figure 2 illustrates this architecture.

3.3.1 Template Library

The system includes a comprehensive library of description templates organized by:

- Object category (350+ categories)
- Spatial relationship (42 relationship types)
- Information priority level
- Description length (brief, standard, detailed)

Templates are structured with slots for dynamic content while maintaining natural syntax and flow. For example:

[OBJECT_TYPE] [DISTANCE] [DIRECTION] from you, [ADDITIONAL_ATTRIBUTE]

This template might generate: "Chair 3 feet ahead of you, facing toward you"

3.3.2 Context Management

The Context Management module maintains:

- Recently described objects to prevent redundancy
- User movement history to detect perspective changes
- Environmental change tracking to highlight new elements
- User interaction patterns to inform personalization

This context awareness enables more natural conversational flow and prevents information overload through excessive repetition.

3.3.3 Natural Language Refinement

The final stage applies linguistic refinements to improve naturalness:

- Appropriate discourse markers and transitions
- Natural variation in phrase structure
- Contractions and informal constructions when appropriate
- Prosodic markers for text-to-speech emphasis

These refinements help avoid the robotic quality often associated with template-based generation, resulting in more natural and engaging descriptions.

3.4 Personalization Framework

ContextSpeak™ includes a comprehensive personalization framework that adapts to individual user preferences and needs. The system allows customization along several dimensions:

3.4.1 Information Density

Users can adjust the amount of information provided through five verbosity levels:

- Minimal: Safety-critical information only
- Concise: Essential navigation elements with minimal detail
- Standard: Balanced information covering navigation and functional objects
- Detailed: Comprehensive descriptions including contextual information
- Complete: Maximum available information about the environment

3.4.2 Description Style

The system offers multiple description styles to match user preferences:

- Functional: Emphasizing practical aspects and uses
- Spatial: Focusing on precise locations and measurements
- Directional: Prioritizing guidance and wayfinding
- Descriptive: Including more sensory and contextual details

3.4.3 Adaptive Learning

The system implements a feedback loop that learns from user interactions:

- Tracking which descriptions prompt user inquiries
- Monitoring navigation success following different description types
- Observing patterns in user movement after receiving information
- Recording explicit preferences through the user interface

This adaptive component allows the system to evolve with user needs without requiring explicit reconfiguration.

4. Implementation

4.1 Technical Architecture

ContextSpeak™ is implemented as a modular system that integrates with object detection and distance estimation components. The core modules include:

1. **Priority Manager:** Implements the hierarchical prioritization framework
2. **Spatial Processor:** Converts absolute coordinates to user-relative positions
3. **Template Engine:** Selects and fills appropriate description templates
4. **Context Tracker:** Maintains historical and environmental context
5. **Personalization Manager:** Handles user preferences and adaptive learning

The system is implemented in Python with optimized components in C++ for performance-critical operations. Natural language processing utilizes the spaCy library with custom extensions for spatial language.

4.2 Integration with Vision Systems

ContextSpeak™ is designed to work closely with computer vision systems, particularly our DenseVision™ object detection framework and DepthSense™ distance estimation technology. Integration points include:

- Shared scene graph representation for detected objects
- Unified confidence scoring for detection and description
- Coordinated update cycles based on environmental changes
- Consistent object identification across systems

This tight integration ensures that verbal descriptions accurately reflect the visual analysis, maintaining user trust in the system.

4.3 Performance Optimization

Several optimizations enable real-time performance on mobile devices:

1. **Incremental updates:** Generating descriptions only for changed elements
2. **Priority-based processing:** Allocating computational resources based on information importance
3. **Template caching:** Reusing common description patterns
4. **Batched language refinement:** Processing multiple descriptions simultaneously

These optimizations result in an average description latency of 110ms following object detection, which ensures that verbal feedback remains synchronized with the user's movement through space.

5. Evaluation

We conducted comprehensive evaluations of ContextSpeak™ through both technical assessments and user studies.

5.1 Technical Evaluation

5.1.1 Methodology

The technical evaluation focused on:

- Description accuracy: Correspondence between descriptions and actual scene elements
- Information relevance: Appropriate prioritization of navigational information
- Linguistic quality: Naturalness and clarity of generated language
- Computational efficiency: Processing time and resource usage

We compared ContextSpeak™ against three alternative approaches:

- Generic image captioning (Microsoft Azure Cognitive Services)
- Basic object detection with fixed templates
- An earlier research prototype focusing on accessibility

Evaluation used a corpus of 1,200 scenes representing diverse environments and navigation scenarios.

5.1.2 Results

Accuracy and Relevance:

Table 1 shows the accuracy and relevance metrics across systems.

Table 1: Description Accuracy and Relevance

System	Object Identification Accuracy	Spatial Accuracy	Navigation Relevance Score
ContextSpeak™	96.3%	93.8%	8.7/10
Image Captioning	89.2%	61.4%	4.2/10
Basic Object Detection	94.7%	79.2%	6.3/10
Research Prototype	92.5%	85.6%	7.4/10

Navigation Relevance Score was determined by expert evaluators rating the usefulness of descriptions for navigation tasks.

Linguistic Quality:

Figure 3 shows linguistic quality assessments by professional evaluators.

Computational Performance:

Table 2 presents computational metrics on a mid-range smartphone (Pixel 6).

Table 2: Computational Performance

System	Average Latency	CPU Usage	Memory Usage
ContextSpeak™	110ms	4.2%	78MB
Image Captioning	350ms	7.8%	256MB
Basic Object Detection	65ms	2.3%	42MB
Research Prototype	185ms	5.6%	124MB

5.2 User Study

5.2.1 Participants and Methodology

We conducted a user study with 56 blind participants (31 female, 25 male) aged 19-72 (mean = 43.5, SD = 14.2). Participants represented diverse backgrounds in terms of:

- Onset of blindness: 33 congenital, 23 acquired
- Mobility aid usage: 41 white cane users, 15 guide dog users
- Technology proficiency: 22 high, 26 medium, 8 low (self-reported)

The study employed a within-subjects design where participants completed navigation tasks using ContextSpeak™ and two alternative systems (counterbalanced order). Tasks included:

1. Navigating unfamiliar indoor environments
2. Locating specific objects in a room
3. Following a predetermined path with obstacles
4. Building and describing a mental map of a space

For each task, we measured task completion time, accuracy, and confidence ratings. We also conducted semi-structured interviews and collected think-aloud protocols during tasks.

5.2.2 Results

Task Performance:

Table 3 summarizes task performance metrics.

Table 3: Navigation Task Performance

Metric	ContextSpeak™	System A	System B	Improvement (vs. best alternative)
Task completion time	64.3s	103.7s	97.5s	34.1%
Object location accuracy	92.7%	76.4%	79.2%	17.0%
Path deviation	0.74m	1.35m	1.28m	42.2%
Mental mapping score	8.4/10	5.9/10	6.3/10	33.3%

User Experience:

Figure 4 illustrates subjective ratings across different aspects of the user experience.

Key findings include:

- Participants rated ContextSpeak™ significantly higher for "spatial clarity" (M=8.7/10) compared to alternatives (M=6.1/10, M=6.4/10)

- 51/56 participants (91%) preferred ContextSpeak™ overall
- 48/56 participants (86%) reported greater environmental confidence with ContextSpeak™

Personalization Impact:

We also evaluated the impact of personalization features:

- Participants using personalized settings completed tasks 22% faster than those using default settings
- Adaptive learning improved task performance by an additional 14% after three usage sessions
- 87% of participants rated personalization features as "very important" or "essential"

Qualitative Insights:

Thematic analysis of interviews revealed several key benefits of ContextSpeak™:

1. **Intuitive Spatial Understanding:** Participants highlighted how the system's consistent reference frame and precise distances created a more intuitive understanding of spaces.

"For the first time, I could build a mental picture of the room without touching everything. The way it described distances and directions just made sense." - P17, 37, congenitally blind

2. **Confidence in Movement:** Many participants noted increased confidence in their movements through space.

"Knowing exactly how far away things are changed everything. I could walk with purpose instead of shuffling cautiously." - P29, 54, acquired blindness

3. **Reduced Cognitive Load:** Participants appreciated the priority-based information presentation.

"Other systems overwhelm you with everything at once. This one told me what I needed to know, when I needed to know it." - P42, 28, congenitally blind

4. **Natural Interaction:** The conversational quality of descriptions was frequently mentioned as a benefit.

"It didn't feel like a machine reading off coordinates. It felt like someone describing the space to me in a natural way." - P08, 61, acquired blindness

6. Discussion

6.1 Key Contributions

ContextSpeak