UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

FIB

BLINDWIKI 2.0: AN INTELLIGENT VERSION

ADRIA LISA BOU

Thesis supervisor
KARINA GIBERT OLIVERAS (Department of Statistics and Operations Research)

Thesis co-supervisor
XAVIER ANGERRI TORREDEFLOT (Department of Statistics and Operations Research)

Degree
Master's Degree in Innovation and Research in Informatics (Advanced Computing)

Master's thesis

Facultat d'Informàtica de Barcelona (FIB)

Universitat Politècnica de Catalunya (UPC) - BarcelonaTech

TO DO.

Abstract

[1] TO DO.

Keywords

TO DO.

Contents

1. Introducción

BlindWiki es una red de audio geolocalizada que permite a personas con discapacidad visual total o parcial compartir sus experiencias mediante grabaciones sonoras utilizando teléfonos inteligentes. Creada en 2014 por Antoni Abad, la plataforma no se limita a documentar dificultades y barreras urbanas, sino que constituye un repositorio de experiencias, opiniones e historias que genera una cartografía colaborativa y creativa de lo invisible.

El proyecto tuvo su origen como una iniciativa para dar voz a las personas con discapacidad visual, permitiéndoles documentar y compartir su percepción única del entorno urbano. Desde su creación, BlindWiki ha experimentado una notable expansión internacional, desarrollándose en ciudades como Roma (2014-2015), Sydney (2015), Berlín y Wrocław (2016), Venecia (2017), Valencia (2020) y São Paulo (2022), adaptándose a cada contexto cultural y lingüístico.

La aplicación móvil de BlindWiki, disponible tanto para Android como iOS, permite a los participantes grabar audio específico del lugar y publicarlo inmediatamente en la plataforma. Los usuarios pueden desplazarse por sus ciudades mientras publican y reciben descripciones de audio geolocalizadas, historias, obstáculos o crónicas previamente contribuidas a través de la app. Esta funcionalidad fa-

cilita la creación de un mapa sensorial colectivo que enriquece la experiencia de navegación urbana para personas con discapacidad visual.

1.1 Motivación y Objetivos

La necesidad de rediseñar la aplicación surge de los avances tecnológicos en los dispositivos móviles y sus sistemas operativos, así como de la comunidad internacional de usuarios que demanda mejoras en la accesibilidad y funcionalidad.

El stack tecnológico de la aplicación original se compone de PhoneGap/Cordova como framework base para el empaquetado, e Ionic 1 con Angular.js para la interfaz de usuario y lógica de la aplicación. Esta arquitectura, aunque permitió un desarrollo multiplataforma eficiente en su momento, ha quedado obsoleta frente a los estándares actuales de desarrollo móvil, lo que dificulta la implementación de nuevas funcionalidades y afecta al rendimiento en dispositivos modernos.

Para abordar estos desafíos, se ha desarrollado BlindWiki 2.0, una aplicación completamente rediseñada utilizando tecnologías modernas y eficientes. El nuevo stack tecnológico se basa en **React Native** y **Expo**.

Además, la expansión internacional del proyecto ha llevado a la necesidad de superar las barreras lingüísticas que limitan la comunicación entre participantes de diferentes países. Para lograr este objetivo, es preciso implementar una capa de inteligencia artificial basada en modelos avanzados de procesamiento de lenguaje natural y audio.

Específicamente, se utilizan tres componentes principales:

- **Identificación automática de idioma:** Mediante el modelo TO DO. , que permite detectar el idioma original de cada grabación de audio.
- **Transcripción de audio a texto:** Utilizando el modelo TO DO. para convertir las grabaciones de audio en texto en su idioma original.
- **Traducción de audio a audio:** TO DO.

1.2 Objectivos

2. State of the Art

Speech-to-speech (S2S) translation and automatic speech recognition (ASR) have seen transformative advances with multimodal AI models. Below is an analysis of state-of-the-art systems, focusing on noise robustness, language coverage, open-source availability, and benchmark performance.

2.1 Key Models in Speech Translation

2.1.1 SeamlessM4T (Meta AI)

A unified multimodal model supporting S2S translation, S2T, T2T, T2S, and ASR across 100 languages. Key features:

- **Noise Robustness:** Outperforms Whisper-Large-v2 by 38% on background noise (Fleurs benchmark) and 49% on speaker variation.
- **Language Coverage:** Handles 100 → English and English → 35 for S2S, 95 → English and English → 95 for T2T, and ASR for 96 languages.
- **Open Source:** Full model weights, training code, and aligned speech dataset (SeamlessAlign) released.
- **Safety:** Reduces added toxicity by up to 63% compared to SOTA models.

2.1.2 Whisper (OpenAI)

Primarily an ASR model (speech-to-text) supporting 97 languages:

- **Noise Robustness:** Less robust than SeamlessM4T; SeamlessM4T reduces WER by 45% on overlapping languages in noisy conditions.
- **Limitations:** No native S2S support; cascaded systems required for full translation.

2.1.3 Coworker-Proposed Models

Details unavailable due to inaccessible document (`estat_art_t2_t3_blind_wiki.docx`). Provide the file to incorporate these into the analysis.

2.2 Benchmarks Explained

Fleurs:
- Evaluates multilingual ASR and S2TT across 102 languages.
- Tests robustness to noise/speaker variations.
- Metrics: BLEU (translation), WER (ASR).

CVSS:
- Focuses on S2ST quality using ASR-BLEU (speech output transcribed and compared to reference text).

CoVoST 2:
- Measures S2TT from English to 15 languages.
- Metric: BLEU.

Flores:
- Text-to-text translation benchmark for 204 languages.
- Metric: chrF++ (character-level F-score).

Blaser 2.0:
- Modality-agnostic metric for translation quality, correlating with human judgments.

2.3 Performance Comparison

2.4 Critical Analysis

SeamlessM4T dominates in multitasking capability, supporting direct S2S without cascaded systems. Its 100-language coverage and noise resilience make it ideal for real-world applications.

Table 1: Comparison of Speech Translation Models

Model	**Languages Supported (ASR/S2S)**	**Noise Robustness (Fleurs WER ↓)**	**Open Source**	**Fleurs S2TT (BLEU ↑)**	**CVSS S2ST (ASR-BLEU ↑)**	**Flores T2T (chrF++ ↑)**
SeamlessM4T-Large	96 / 100 → 35	38% better than Whisper	Yes	20.4 (X→Eng)	58.7	54.3 (Eng→X)
Whisper-Large-v2	97 / None	Baseline	Yes	16.2 (X→Eng)	N/A	N/A
Coworker Models	Not specified	Not specified	Unknown	Not specified	Not specified	Not specified

Whisper remains strong for ASR but lacks native S2S and lags in translation benchmarks.

Open-Source Gap: SeamlessM4T's release of training data and tools sets a new standard for reproducibility, unlike proprietary models like AudioPaLM-2.

For applications requiring robustness to ambient noise and broad language support, SeamlessM4T is the current SOTA. Integrate coworker-proposed models into this framework once their documentation is available.

3. Specification and design of the solution

3.1 Overview

Disseny de blind wiki 1.0 vs disseny de blind wiki 2.0.

Comentar breument el perquè escullo React Native per fer la app.

Detall sobre el model de traducció el·legit.

4. Development of the work

Esquema Gnatt de la meva planificació

Aquesta secció representa que ha de ser un diari del meu progrés???

Detalls de implementació.

5. Experimentation and evaluation of the work

Mesures del rendiment algorísmiques del sistema de traducció podrien sert interessants. Mesures pràctiques derl sistema en el entorn de producció podrien ser interessant. Potser hi ha certs paràmetres que es poden ajustar en funció del servidor del que disposem.

6. Sustainability analysis and ethical implications

Anàlisi del cost de manteniment.

Parlar de privacitat? Acció benèfica?

Potser és més interessant enfocar aquesta secció al tema de la accessibilitat digital.

7. Conclusion

Conclusió.

References

[1] Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen. CVSS corpus and massively multilingual speech-to-speech translation. In *Proceedings of Language Resources and Evaluation Conference (LREC)*, pages 6691–6703, 2022.