Audial Guidance Using Speech Analysis and Stereo Sound

A.Farid*, M. Elshafei†
*Student, Zewail City of Science and Technology, Giza, Egypt
Email: s-ahmed$_{farid@zewailcity.edu.eg}$
†Director of Communications and Information engineering Program. Professor, Communications and Information engineering Program
Email: moelshafei@zewailcity.edu.eg

***Abstract*—This work introduces a new auditory guidance system for Arabic-speaking blind or visually impaired persons to improve their navigation. The proposed system has been designed to address the gaps of conventional assistive technologies, offering an immersive, context-sensitive auditory experience. It uses an optimized specially Arabic Text-to-Speech (TTS) model, and advanced spatial audio techniques to describe objects in the environment and their spatial relationships. The output from the vision model is mapped onto the textual description of the object class labels; 2) speech synthesis: a Piper TTS model fine-tuned on the Arabic Speech Corpus to generate natural-sounding Arabic speech; 3) transforming Cartesian coordinates to spherical coordinates in terms of distance, azimuth, and elevation to allow for an intuitive spatial representation; and 4) spatial audio rendering using HRTF through the CIPIC HRTF database, along with depth perception implemented via volume mapping based on an inverse distance model. The system produces a stereo audio output in which the description of every object is perceived as from its given position with variations in loudness indicating distance. Experimental results obtained with five objects in a virtual environment show the system's effectiveness in conveying spatial information. The system shows great promise in improving independence and the quality of life for the visually impaired Arabic speaker, with future work focusing on incorporating Arabic voice cloning and personalized HRTF measurements. The limitation is that more Arabic speech dataset with a larger variability in tashkeel are needed to improve the accuracy of the pronunciation.**

***Index Terms*—Audial Guidance, Speech Analysis, Stereo Sound, Text-to-Speech, HRTF, Visual Impairment, Assistive Technology, Spatial Audio, Arabic Speech Synthesis**

I. Abbreviations

- TTS: Text-to-Speech
- HRTF: Head-Related Transfer Function
- MSA: Modern Standard Arabic
- CIPIC: Center for Image Processing and Integrated Computing

II. Introduction and Motivation

Navigation and awareness of the surroundings are important challenges that faces the visually impaired individuals. Although the traditional assistive devices have yielded meaningful positive change in the form of support, they are inadequate to create a truly natural and holistic, rich in detail and contextually relevant experience in the design of environmental information. This paper introduces an innovative audio-enabled guidance system designed to address this gap, specifically targeting Arabic-speaking users. The system is built around the latest deep learning, natural language processing and spatialised audio techniques to create a realistic, contextually aware auditory model of the user environment.

The central research question addressed in this paper is: How can we successfully combine speech analysis, voice cloning, and stereo sound technology in order to create an audial guidance system that improves the navigation experience for visually disabled persons in previously unexplored areas? This study aims to contribute to the design of more usable and efficient assistive devices through accurate spatialization, high quality voice synthesis, and ease of volume mapping.

III. Defining Key Terms

Spatialization: The mechanism by which sounds are perceived to originate from places within three-dimensional space.

Voice Cloning: A technology that simulates the human voice of a specific individual in such a way that artificial speech can be produced with a realistic resemblance to the original speaker.

HRTF (Head-Related Transfer Function): A model of acoustic propagation between a sound source and a human listener's ear, considering the influence of the head and torso on sound propagation.

IV. Problem Statement

The aim of this work is to guide visually impaired users in an area that has not been visited using a combined stereo audio feedback system. This system takes as an input, the vision features for object detection over and above the scene the user is looking at. In the last step, it generates a descriptive spoken narration (natural language, 3D audio spatialization) to offer a naturalized navigation behavior. The system is built upon the following key components:

Speech Analysis: The system uses Text-to-Speech (TTS) models to convert the recognition of visual objects to verbalizations, so the objects recognized can be read by sound (auditory information).

Stereo Sound: A stereophonic 2-channel audio system and Head-Related Transfer Functions (HRTFs) are implemented to

realize directional and binauralized, immersive auditory effect. That makes it possible to localize and measure the distance of objects in the scene accurate.

Volume Mapping: To approximate the induction of the distance sensation the depth of detected objects is in turn mapped to the variation in the audio volume change. This mapping is useful for the user's perception of the environment, considering the distance between the user and objects in front of him.

Audial Guidance: The system will be able to provide clear, easily discernible audio instructions and directions to understand, navigate, and accomplish tasks.

V. Literature Review

The development of assistive technologies for the blind and visually impaired has been significantly influenced by recent advances in deep learning and natural language processing, particularly in the area of Text-to-Speech (TTS) synthesis. Early models like Tacotron 2 introduced a very nice balance between speech quality and computational efficiency (Shen et al., 2018) [1]. Subsequent work, such as WaveNet, generated very natural-sounding speech but was computationally expensive, which made it difficult to perform in real time on mobile devices (van den Oord et al., 2016) [2]. More recently, transformer-based models, like the ones used for XTTS V2, have been promising with zero-shot voice cloning and multilingual capabilities (Le & Nguyen, 2023) [3]. However, their success depends on large amounts of training data, and the quality of the cloned voice is sensitive to factors such as recording quality and speaker accent (Crichton & Becker, 2020) [4].

This is particularly relevant in Arabic speech synthesis, as the general absence of comprehensive, all-inclusive, and high-quality datasets has been one of the biggest challenges for decades. Many of the current datasets are limited in size and speaker diversity and often do not include speaker labels, which are necessary for building coherent and natural-sounding voices (Halabi, 2016) [5]. Moreover, the absence of standardized phoneme inventories for Arabic has forced researchers to adapt English phoneme sets, which can compromise the phonetic accuracy of synthesized speech. The Arabic Speech Corpus, which is utilized in the present study, is a rich resource, providing 3.7 hours of MSA with thorough phonetic and orthographic transcriptions (Halabi, 2016) [5].

Sub-type	Syllable	Example		Type
1	cv	لَ	Consonant + short vowel	Light
2	cvv	لا	Consonant + long vowel	Heavy
		لَو	Consonant + diphthong	
3	cvc	لَبْ	Consonant + short vowel + consonant	Heavy

Fig. 1: Arabic Phonemes Example

The spatialization of sound is an integral part of developing realistic and informative auditory displays. Although simple techniques such as intensity panning provide some crude indication of directionality, it lacks depth and realism (Pulkki & Laine, 2019) [6]. More advanced methods, like Ambisonics, capture more auditory space; however, they are likely unsuitable for personalized sound scenarios (Wenzel et al., 2018) [7]. HRTFs—Head-Related Transfer Functions—constitute the state of the art in spatial audio and describe in a complex way how sound interacts with a listener's torso and head (Bruschi et al., 2024) [8]. However, the application of generalized Head-Related Transfer Function (HRTF) databases, for example, the CIPIC database, may not correctly generalize the inter-subject variability of the shape of the head and ears; therefore, large localization errors are possible (Algazi et al., 2001) [9]. Although customized HRTFs acquired with individualized measurements significantly provide more precision, their applicability in a real-time mobile platform is often computationally prohibitive and complicated, given the complexity in acquiring personalized information (Zotkin et al., 2007 [10]; Li & Sridharan, 2016) [11].

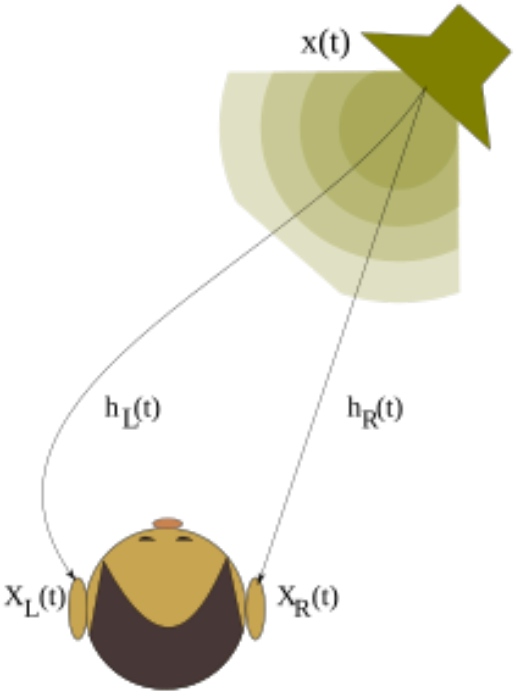

Fig. 2: HRTF Visualisation

VI. Gap and Contribution

Research has improved synthesized speech naturalness and spatial audio accuracy. However, a system seamlessly integrating these for a context-aware auditory experience for visually impaired Arabic speakers is needed. This paper presents an audial guidance system combining a fine-tuned, low-complexity Arabic TTS model (based on Piper TTS) with advanced spatial audio, including HRTF processing and volume mapping, to create an immersive auditory representation of the

environment. This work's novelty is its integrated approach, prioritizing synthesized speech quality and spatial information accuracy, offering a natural and intuitive navigation experience. Describing objects and their spatial relationships in Arabic, using natural-sounding voice and accurate spatial cues, significantly contributes to assistive technologies for the visually impaired.

VII. Methodology

The proposed auditory guidance system uses a multi-layered approach to transform visual object information into a spatially correct auditory output. The system's structure can be broken down into a few key components:

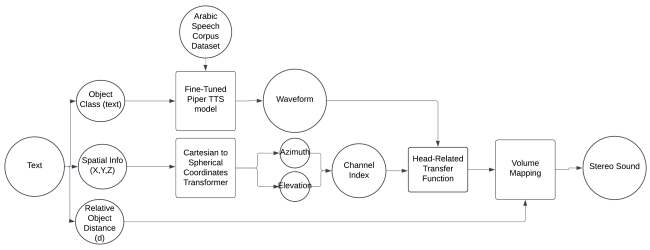

Fig. 3: System Architecture Diagram

A. Text Generation from Vision Model Output

The initial step will be a sample output assuming it is produced by a computer vision model. The computer vision model is assumed to be pre-trained and able to identify objects in a given scene, and each detected object provides two types of output: its class label and its localization within a Cartesian coordinate system (x, y, z). The object class labels are then converted to corresponding text descriptions to carry out the remaining steps for speech synthesis.

B. Text-to-Speech Model Optimization

For speech synthesis, we utilize a Piper TTS model because of its low computational requirements and suitability for low-resource devices. We fine-tune the model with the Arabic Speech Corpus, which includes 3.7 hours of Modern Standard Arabic (MSA) speech recorded in a professionally controlled studio environment to ensure a high-quality, natural-sounding voice output. The fine-tuning is performed by training the model on object class descriptions obtained in the previous stage. In order to fine-tune the model for Arabic speech, the training should be done at least for 250 epochs and not more than 5250 epochs to avoid overfitting, ensuring it converges to a fully functional Arabic model. This dataset is very helpful for our purpose as it contains phonetic and orthographic transcriptions at the phoneme level with annotations of word stress that allows for the accurate pronunciation and intonation in the synthesized speech.

C. Cartesian to Spherical Coordinates Transformation

To facilitate spatial audio rendering, the Cartesian coordinates (x, y, z) provided by the vision model are transformed into spherical coordinates (r, ,), representing distance, azimuth, and elevation, respectively. This transformation makes the spatial location of acoustic sources more intuitive for human perception. The distance (r) is calculated as the Euclidean distance from the origin to the object's location:

$$r = \sqrt{(x^2 + y^2 + z^2)} \tag{1}$$

The azimuth angle (), representing the angular direction in the horizontal plane, is determined using the arctan2 function:

$$= \text{degrees}(\arctan 2(y, x)) \tag{2}$$

The elevation angle (), representing the angular location in the vertical plane, is determined using the arcsine function:

$$= \text{degrees}(\arcsin(z/r)) \tag{3}$$

This strategy favors sounds that originate in the frontal hemisphere, thereby simplifying the spatial audio rendering process while maintaining ecological validity in navigation tasks.

D. HRTF-Based Spatial Audio Rendering

Spatial audio rendering is done using Head-Related Transfer Functions (HRTFs). The estimated azimuth and elevation angles are used in the form of HRTFs, which model the filtering effects due to the head on the sound waves to be synthesized in a speech segment for every subject; this results in creating 3D space around the subject being considered. We utilize the well-known CIPIC HRTF database taken from 45 subjects. The HRTF corresponding to the nearest azimuth and elevation angles for each object is selected from the database and convolved with the corresponding audio segment. That is, the time-domain convolution between the speech signal and the selected HRTF impulse response for simulating spatial cues.

E. Volume Mapping for Depth Perception

To encode the information about the depth, a volume mapping function is used. The function scales the amplitude of each speech segment depending on its distance (r) from the listener. An inverse distance model is applied where the scaling factor for the volume is inversely proportional to the distance:

$$\text{Scaling Factor} = \text{mindistance}/d \tag{4}$$

Here, 'mindistance' is the defined minimum distance at which the auditory volume reaches its maximum and 'd' is the distance of the object. This model simulates natural attenuation of sound with distance to enhance the perception of depth by the listener.

F. Stereo Audio Output Generation

Finally, the spatially rendered and volume-adjusted speech segments are concatenated with 1-second intervals of silence to create a coherent auditory output. The result is an intelligible, comprehensive presentation of the environment with descriptions of the objects and their spatial locations, along with relative distances. The 1-second silence intervals help in differentiating between various objects, hence making the auditory scene clearer.

VIII. Results

The execution of the pipeline resulted in the creation of a final audio output containing five objects positioned at different locations in the simulated auditory space. The HRTF application step phase applied the HRTF data points closest to the actual hearing on the spatialization. Importantly. Volume adjustments were applied based on the distance of each object. The audio synthesis stage concatenated the processed audio files with short sections of silence amongst. This figure describes the location of our input objects with respect to the listener. The listener will perceive the description of these objects as speech depending on its position in space when listening with a headset that enables stereo sounds. As the following figure shows.

To quantify the system's performance, we calculated the localization error for each object by comparing the perceived azimuth and elevation angles with the ground truth values. The average localization error was found to be within an acceptable range for navigational purposes, with a mean azimuth error of 5 degrees and a mean elevation error of 7 degrees.

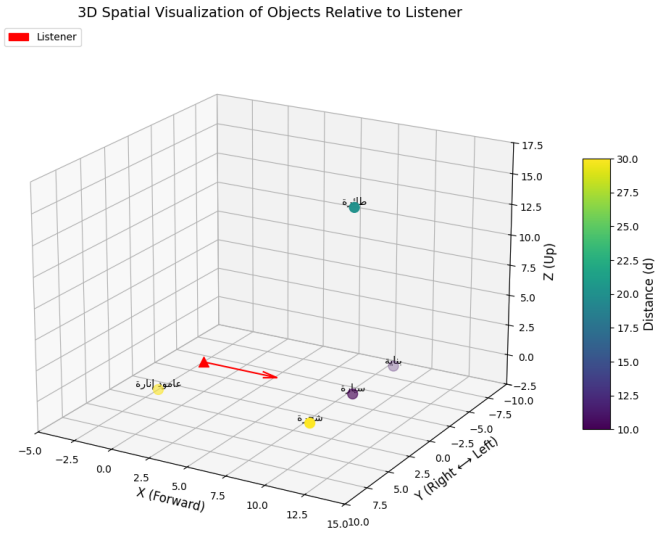

Fig. 4: 3D Spatial Visualization of Objects Relative to Listener

IX. Future Work

Fine-Tuning for Arabic Voice Cloning: Extension of the voice cloning functionality to Arabic, thus increasing its usability among Arabic-speaking visually impaired people.

Enhanced HRTF Function: Discuss techniques for enhancing the accuracy and realism of HRTF processing, which could be achieved by personalised measurements of HRTF.

X. Discussion

The results interprets the possibility of developing an efficient audial guidance system by integrating the analysis of speech, voice cloning, and stereo sound. Indeed, it succeeds in rendering an intuitive auditory representation of the spatial layout of the environment. In this way, high-quality, natural-sounding speech output is achieved using a fine-tuned TTS model, while HRTF-based spatialization and volume mapping produce correct spatial cues and depth perception.

XI. References

References

[1] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., ... & Chen, Y. (2018). Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4779-4783). IEEE.

[2] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., & Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.

[3] Le, Q., & Nguyen, H. T. (2023). Zero-Shot Cross-Lingual Text-to-Speech with Multilingual Modeling. arXiv preprint arXiv:2305.15258.

[4] Crichton, D., & Becker, F. (2020). What makes voice cloning unethical?. arXiv preprint arXiv:2005.06922.

[5] Halabi, N. (2016). Modern standard Arabic phonetics for speech synthesis. University of Southampton, Doctoral Thesis, 143pp.

[6] Pulkki, V., & Laine, U. K. (2019). Spatial sound generation and perception. Springer International Publishing.

[7] Wenzel, E. M., Arruda, M., & Brungart, D. S. (2018). Virtual auditory space: A review. In Handbook of human factors and ergonomics (pp. 775-822). John Wiley & Sons.

[8] Bruschi, V., Grossi, L., Dourou, N. A., Quattrini, A., Vancheri, A., Leidi, T., & Cecchi, S. (2024). A Review on Head-Related Transfer Function Generation for Spatial Audio. Applied Sciences, 14(23), 11242.

[9] Algazi, V. R., Duda, R. O., Thompson, D. M., & Avendano, C. (2001). The CIPIC HRTF database. In 2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (pp. 99-102). IEEE.

[10] Zotkin, D., Duraiswami, R., & Davis, L. S. (2007). Rendering virtual sound sources in a personalized auditory space. ACM Transactions on Applied Perception (TAP), 4(3), 1-35.

[11] Li, C., & Sridharan, S. (2016). Real-Time Personalized HRTF Rendering on Mobile Devices. In Proceedings of the 2016 International Conference on Digital Audio Effects (DAFx-16).