УДК 004.934 **doi: 10.26906/SUNZ.2026.2.165**

Daniil Raptanov, Olesia Barkovska, Mykhailo Shylenko, Oleksandr Holovchenko, Diana Ivakhnenko

Kharkiv National University of Radio Electronics, Kharkiv, Ukraine

A STUDY OF THE ACCURACY OF BIMFORMING METHODS IN THE CONTEXT OF AN INCLUSIVE INTERNAL NAVIGATION SYSTEM

Abstract. Relevance. Voice control of elements in inclusive navigation systems is critical for ensuring the independence and safe mobility of people with visual impairments in public spaces, particularly in large retail spaces. However, existing speech-to-text (STT) systems face a significant drop in recognition accuracy due to the highly dynamic and non-stationary acoustic noise in supermarkets. **The object** of this study is audio stream preprocessing and spatial filtering (beamforming) in a voice control system under conditions of dynamic, non-stationary noise. The problem lies in the insufficient selectivity of standard audio signal processing algorithms under conditions of background noise in a store, which leads to a critical increase in the word error rate (WER) and makes the smart cart control system vulnerable. The objective of the article is to evaluate the impact of external factors (number, spatial topology of placement, and power level of acoustic noise sources) on the accuracy of spatial filtering (beamforming) methods for subsequent voice command recognition through computer simulation. **As a result** of the study, the acoustic environment and microphone array were simulated using the Pyroomacoustics library. A comparison was conducted between three methods: Delay-and-Sum (DAS), Max-UDR, and Max-SINR. The study showed that the Max-SINR algorithm provides the highest signal-to-noise ratio gain (Delta SNR from 7.9 to 9.1 dB) and is mathematically robust to changes in the distance to interference sources and their power. The DAS method proved to be the least effective (5.35–5.95 dB) and demonstrated sensitivity to changes in distance. It was established that the key factor in signal degradation is the configuration of noise sources, among which the cross topology is the most difficult to filter.

Keywords: inclusive navigation system, visual impairment, speech recognition, spatial filtering, beamforming, Delay-and-Sum, Max-UDR, Max-SINR, dynamic noise.

Introduction

Problem Statement. Existing speech-to-text (STT) systems suffer from low recognition accuracy due to the presence of noise, reverberation, and multi-channel audio sources [1]. The use of microphone arrays in combination with direction-of-arrival (DoA) and beamforming algorithms significantly improves the signal-to-noise ratio in noisy environments [2]. Modern solutions often combine classical DSP algorithms, such as Delay-and-Sum (DAS), MVDR, or GSC, with deep learning methods. However, neural network models require significant computational resources and are sensitive to the acoustics of the rooms in which they were trained [3].

Despite advances in spatial filtering algorithms, their application in inclusive systems for retail spaces remains understudied [4, 5]. In the highly dynamic and non-stationary acoustic noise conditions of a retail floor, standard cloud-based and local STT methods exhibit significant degradation in accuracy. Accordingly, the adaptation of beamforming methods at the hardware level for smart shopping carts designed to assist visually impaired people in supermarkets is critically necessary to improve the signal-to-noise ratio (SNR) prior to the semantic processing stage.

The problem lies in the insufficient selectivity of standard audio signal processing algorithms under conditions of dynamic non-stationary noise, which leads to a critical increase in word error rates and makes the shopping cart control system vulnerable to background store noise. The subject of the study is the preprocessing of the audio stream and spatial filtering in a voice control system under conditions of dynamic non-stationary noise. The subject of the study is the effectiveness of spatial filtering algorithms (DAS, UDR, SINR) for improving the recognition of user voice commands in an inclusive smart cart system.

The scientific novelty lies in the introduction of an analytical stability metric (robustness index), which allows for a quantitative assessment of the degradation in the performance of spatial filtering algorithms as the topology of nonlinear disturbances becomes more complex (e.g., an increase in the number of shoppers, carts, or other noise sources). The practical significance is determined by the optimal configuration of the spatial filtering module, which maximizes the ASNR parameter and stabilizes the entire speech-to-text conversion pipeline. This ensures the reliable operation of an inclusive voice interface for the shopping cart without requiring significant computational resources, thereby promoting the independence of visually impaired users while shopping.

Analysis of Recent Research and Publications. The field of voice command recognition is explored in this paper in the context of using inclusive navigation systems with voice interfaces for people with visual impairments, underscoring the relevance of the chosen topic. Given the current situation in Ukraine, three key user groups can be identified for whom standard voice assistants are inaccessible:

- military personnel and civilians who have lost their sight due to trauma face an urgent need for spatial and psychological adaptation to new conditions and require tools to restore social autonomy in public spaces;

- elderly people with age-related vision impairments (cataracts, glaucoma, etc.) are unable to read small print on products, price tags, and navigation signs in large retail spaces;

- people with congenital or acquired blindness and low vision are critically dependent on reliable assistive technologies for spatial orientation, avoiding obstacles, and independent self-care.

Voice control aids in the socialization and safe mobility of people with disabilities in public spaces. Therefore, the processing and analysis of voice commands to

control elements of an inclusive system must be implemented for these groups of people with disabilities to enhance their independence, autonomy, and sense of security, which is a pressing task.

Examples of control scenarios for different population groups are provided in the table below (Table 1). The system must adapt to the user's specific needs, ensuring the performance of vital functions even with impaired vision.

Table 1 – **Management Scenarios for People with Limited Mobility**

Command category	Query example	System action
Navigation query	" Show me the way to …"	Calculating the optimal route to the desired department or shelf, with voice guidance and obstacle warnings.
Subject query	" I need …"	Searching for products in the product database, announcing prices and expiration dates, and providing assistance right at the shelf.
Emergency assistance	"Help me get outside", "We need help from healthcare workers "	Calculating the shortest route to the exit, automatically calling an assistant or the sales floor manager via the internal communication system.

Current research has already demonstrated that the use of microphone arrays, combined with sound source localization and mixed signal separation algorithms, can significantly improve the signal-to-noise ratio in multichannel environments. Existing studies show that the use of beamforming and DoA (direction of arrival) methods [6, 7, 13] significantly improves noise suppression while using omnidirectional and relatively inexpensive sensors. The geometry of the microphone array plays an important role and is directly related to the hardware component of the module. Some studies indicate a significant improvement in audio quality when using two- and three-dimensional arrays, with the number of sensors being the primary factor affecting the method's performance.

These methods are particularly critical for embedded real-time speech recognition systems, given limited computational resources and the need for rapid system response. Methods such as beamforming and DoA analysis are actively combined with classical DSP algorithms (e.g., Wiener filters, spectral subtraction) and deep learning methods (GRU, CNN). This allows for effective reduction of background noise, amplification of speech, and ensures stable operation of STT modules even in noisy environments [8, 9, 14]. This study focuses on the audio stream preprocessing module and noise suppression in the presence of user speech disturbances in the system described above. Furthermore, given the variety of speech impairments, microphone array configurations, background noise, and other environmental characteristics, the question arises: which beamforming method and/or signal arrival direction estimation method is optimal under specific conditions. Since the beamforming method provides a significant improvement in the output audio in the presence of background noise, and the hardware component requires a microphone array (two or more microphones), while the software component includes complex mathematical algorithms such as FaS (Filter And Sum), MVDR (Minimum Variance Distortionless Response), GSC (Generalized Sidelobe Canceller), etc.–the best and most accurate method for conducting the experiment will be the use of actual hardware and algorithms optimized for it [10–12]. The Pyroomacoustics utility is proposed as a modeling environment for the room and microphone array, as it provides classic implementations of the most popular algorithms.

Table 2 – **An Overview of Classical Beamforming and DoA Methods**

Method name	Features	Areas of application
Delay–and–Sum (DAS) [15]	- low computational load; - low resolution.	simple hearing aids; educational projects; arrays with a large number of microphones.
Functional Beamforming [16]	higher resolution compared to DAS; better suppression of side lobes and specular sources; requires calibration/correction for accurate dB measurements; sensitive to the choice of power exponent.	anechoic chambers; wind tunnel tests; industrial noise diagnostics.
GCC–PHAT (Generalized Cross–Correlation + Phase Transform) [17]	high resistance to reverberation; low computational complexity; works primarily with a single pair of microphones (TDOA); it is difficult to localize multiple speakers simultaneously.	smart speakers; IoT devices; basic video conferencing systems.
SRP–PHAT (Steered Response Power) [18]	reliable in noisy/reverberant environments; does not require knowledge of the number of sources; high computational cost.	robotics; professional conference systems; acoustic monitoring of premises.
MVDR / Capon (Minimum Variance Distortionless Response) [19]	adaptive method; better noise and interference suppression than DAS; sensitive to microphone calibration errors.	noise cancellation systems; telecommunications; speech processing.
MUSIC (Multiple Signal Classification) [20]	ultra-high resolution; can distinguish sources that are close to one another; high computational complexity; performs poorly in reverberant environments.	laboratory acoustic studies; high-precision measurements; radars.
ESPRIT (Estimation of Signal Parameters via Rotational Invariance Techniques) [21]	less computationally intensive than MUSIC; requires a specific array geometry.	specialized instruments with fixed geometry; radars.

In addition to the classical methods listed in Table 2, it is worth highlighting deep machine learning, which is based on standard DoA techniques. Modern voice assistants, smartphones, and AR/VR headsets use such adaptive methods to improve recognition quality. A neural network can learn to account for complex room acoustics, nonlinear distortions, and noise; however, at its core, it is a "black box" that requires large datasets for training and is dependent on the specific acoustics under which the network was trained. The Pyroomacoustics library contains implementations of popular beamforming methods, some of which feature advanced noise suppression algorithms (Table 3).

Table 3 – **An overview of the methods implemented in the Pyroomacoustics library**

Method name (API)	Description	Applications	Requirements
Delay–and–Sum (DAS) rake_delay_and_sum_weights	A basic algorithm that compensates for signal propagation delays to the microphones and sums them up.	Basic SNR enhancement and spatial filtering.	Coordinates of the target source and the array microphones.
MVDR rake_mvdr_filters	Minimizes noise variance while maintaining a unit gain in the direction of the target.	Suppression of directional interference and uncorrelated noise without distorting the useful signal spectrum.	Coordinates of the target source; estimation of the spatial covariance matrix of noise.
Max–SINR rake_max_sinr_filters rake_max_sinr_weights	Maximizes the ratio of the useful signal including selected early reflections to the sum of noise and interference.	Optimizing reception in reverberant rooms in the presence of strong competing sound sources.	The coordinates of the useful and interfering sources, or the interference covariance matrix.
Max–UDR rake_max_udr_filters rake_max_udr_weights	It separates energy into useful (direct sound + early reflections) and harmful (late reverberation + noise).	Speech reverberation and noise suppression (counteracting the harmful blurring of the spectrum).	Source coordinates and room parameters for separating early and late reflections.
Perceptual rake_perceptual_filters	Calculates filters in the time domain, taking psychoacoustics into account. Relaxes the suppression constraints within a short window (30 ms).	Improved speech perception (the Haas effect, integration of early reflections by the auditory system).	Source coordinates and specified time window.
One–Forcing rake_one_forcing_filters rake_one_forcing_weights	Calculates the time-domain filters of a beamformer with a single response to multiple sources.	Signal extraction with strict response constraints for multiple specified directions/reflections.	The exact coordinates of the target sources and their images.
Distortionless rake_distortionless_filters	Ensures undistorted transmission of the target signal in a multipath environment by imposing strict constraints on the phase and amplitude of the target paths.	Signal extraction in reverberant environments where even the slightest phase or amplitude distortion is unacceptable.	Knowledge of the room's impulse response (RIR) or the exact coordinates of the image sources.

Three methods were selected for the experiment: Delay and Sum, Max–SINR, and Max–UDR. The latter two are suitable for reverberant environments and have straightforward requirements. The Delay and Sum method was chosen as a simple and computationally efficient beamforming method.

The research is conducted in a computer simulation environment to model the room and the microphone array. This brings the experiment as close as possible to real-world conditions: reverberation, diffraction, stationary noise, and room modes.

It is worth noting that computer simulation is not the only valid way to test the research hypothesis, namely, to find a correlation between the beamforming method and the clarity of the output audio (removed noise) in the presence of significant speech disturbances from the user. The next step in working on the chosen topic will be an empirical study using real hardware with real microphone arrays for subsequent statistical processing; however, this requires preliminary empirical validation under controlled conditions.

Since the accuracy of the second stage depends on the clarity of the text obtained in the first stage, it becomes necessary to implement spatial filtering (beamforming) methods even before the STT stage. In environments with high acoustic noise, such as a retail store, standard STT methods suffer from a decline in accuracy. Therefore, a spatial filtering (beamforming) stage is required at the hardware level to improve the signal-to-noise ratio (SNR).

To this end, this paper proposes to investigate the change in the difference between the input and output signal-to-noise ratios (SNR) for various beamforming methods depending on:

- the topology of noise source placement;
- the distance of noise sources from the microphone array;
- noise power.

Thus, **the aim of the study is** to evaluate the influence of external factors (number, placement topology, and power level of acoustic noise sources) on the accuracy of beamforming methods for subsequent voice command recognition, using computer modeling.

To achieve the stated goal, the following tasks must be addressed:

- analysis of beamforming methods for use under conditions of variable external factors;
- development of a model for a voice command processing and analysis subsystem in an inclusive indoor navigation system;
- evaluation of the accuracy of the Delay–and–Sum (DAS), Max–SINR, and Max–UDR methods for

determining the speaker's location based on the number, spatial distribution, and power levels of acoustic noise sources;

- analysis of the obtained results.

Future research will focus on conducting field experiments using real hardware under **variable** external conditions.

Main Content

The proposed voice command processing subsystem, shown in Figure 1, is based on an interconnected sequence of steps: first, it isolates and cleans the speaker's voice of noise, converts it into text, and then, using a pre-trained language model, analyzes the semantics, verifies the request based on context, and generates a final command for execution or a voice response, thereby renewing the interaction cycle.

In the first stage of the system's operation (speech-to-text conversion), an unstructured analog or digital audio stream is converted into a formalized text sequence. This process is based on Voice Activity Detection mechanisms, which allow the useful signal to be separated from the background noise of the sales floor. The use of a specialized hardware interface ensures stable data transmission to the Speech-to-Text module, where phoneme recognition and the formation of a textual representation of the query occur using acoustic and linguistic models. The quality of the output data at this level is critical for the entire system, since recognition errors directly determine the accuracy of the subsequent interpretation of commands.

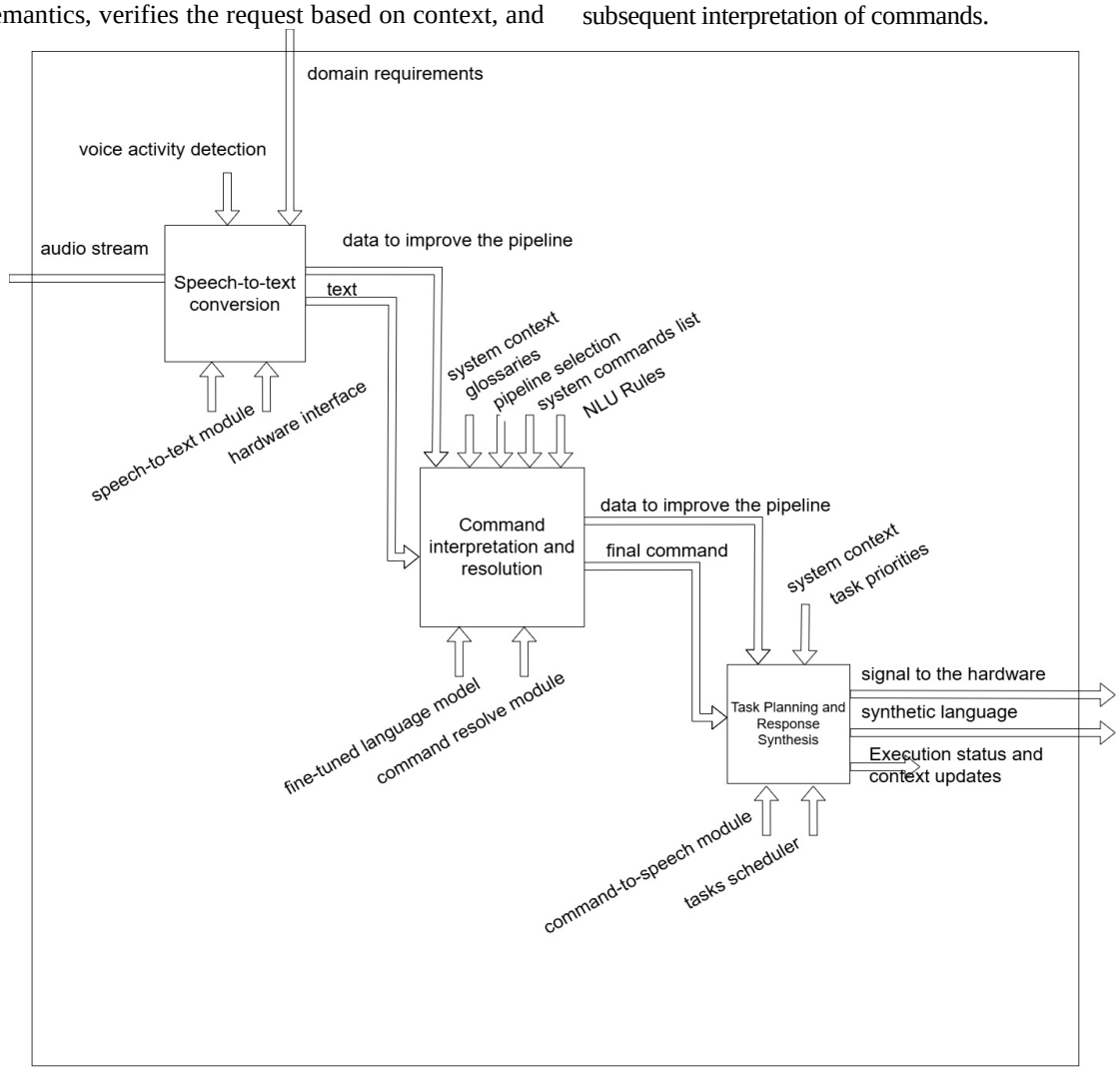


Fig. 1. General model of a voice command processing system

The central component of the system (command interpretation and resolution) is responsible for transforming the received text into structured, machine-readable commands through semantic resolution mechanisms. The interpretation process is based on the use of fine-tuned language models (Fine–tuned LLMs) and natural language understanding rules (NLU Rules), which allow for the identification of user intent within a specific domain. To minimize ambiguity, the system leverages external knowledge in the form of system context and specialized glossaries of terms related to the product assortment and navigation within the marketplace. The result of this stage is the generation of a final command that accurately reflects the user's true intent, taking into account the current situation and the selected processing pipeline.

The final stage of the architecture implements the logic for controlling the intelligent cart's actions and generating feedback. The task scheduling module ranks the received commands according to set priorities and coordinates their execution through context updates. In parallel with sending signals to the hardware, the system generates a voice response using the Command-to-

Speech module, ensuring natural interaction with the user. This approach allows for a closed-loop control cycle, where each action not only fulfills a request but also updates the system state for the correct processing of subsequent dialogue iterations.

The experimental studies were conducted taking into account the following potential influences, limitations, and requirements:

- determining the effect of the distance to noise sources (1 m, 3 m, 5 m) and their intensity relative to the user's voice (quieter than the voice: +10 dB; equal to the voice: 0 dB; louder than the voice: –10 dB) on the input signal;

- calculation of the SNR for beamforming methods under various topologies (a “cross” topology for 4 noise sources, a “ring” topology to simulate diffuse noise from 8 noise sources, and a point-source noise configuration, Fig. 2);

- calculating a coefficient to compare the ability of methods to suppress multiple noise sources without significantly degrading the output signal.

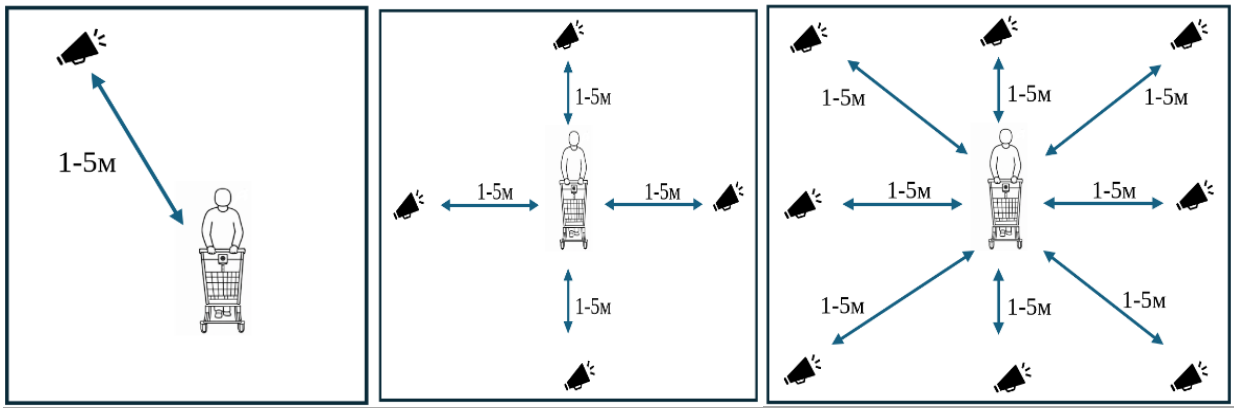


Fig. 2. Modeling of acoustic noise source configurations relative to the useful signal source: a – point source, b – “cross” configuration, c – “ring” configuration

The ΔSNR metric is calculated based on the measurements at the microphone input and output, SNR_{in} та SNR_{out} respectively. The input signal-to-noise ratio is calculated using the formula:

$$SNR_{in} = 10\,log_{10}(\frac{P_{signal}}{P_{noise}}),$$

where P_{signal} - clear voice power, P_{noise} - noise power.

Measurements at the SNR_{out} are determined in the same way, but after applying the beamforming method.

The performance metrics for evaluating the beamforming method under conditions of dynamic non-stationary interference are determined by the formula:

$$SNR_{improvement} = \Delta SNR = SNR_{out} - SNR_{in}.$$

The resulting ΔSNR value (expressed in decibels) serves as a quantitative measure of the effectiveness of spatial signal selection in the presence of additive noise. An increase in the ΔSNR value directly correlates with an increase in the probability of correct operation of the STT module. In particular, high values of this indicator indicate successful compensation for acoustic interference, which allows minimizing the level of phonetic distortions at the hardware interaction level. Conversely, low values of ΔSNR indicate insufficient selectivity of the algorithm, which leads in the future to degradation of the input stream and a critical increase in word error rate, making the system vulnerable to non-stationary noise in the trading floor. In other words, maximizing ΔSNR will lead to the stabilization of the entire pipeline.

The system’s robustness to changes in the noise topology is determined by an analytical metric introduced in this paper to compare beamforming methods. The metric shows how much the algorithm’s performance “drops off” when conditions in the store become more complex (the number of shoppers, carts, or noise increases). It is calculated as the ratio of the average signal gain under complex conditions to the gain under baseline (ideal) conditions.

Results of the research conducted. Discussion

To systematize the research results, taking into account all the variable parameters described in the previous section, this study includes the following experiment: evaluating performance as a function of distance and noise level, as well as assessing robustness to the topology and number of noise sources.

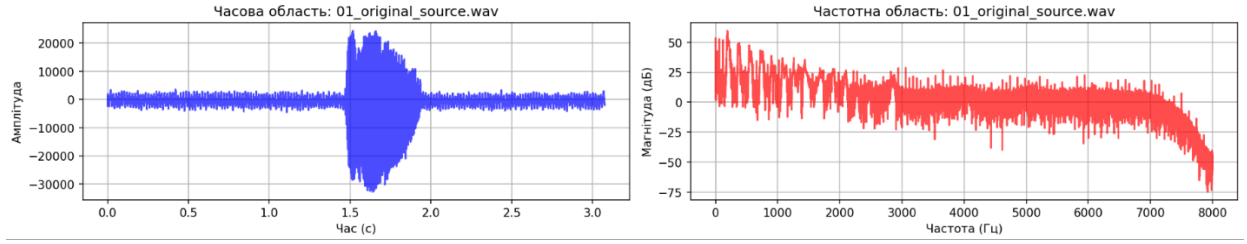
For the experiment, healthy speech recordings were selected from the specialized TORGO dataset, created for the study of dysarthria [21]. Typically, such recordings are clearer and cleaner than those in standard general-purpose datasets. This approach allows us to focus on working with noise and testing hypotheses, and simplifies the overall data preparation process.

For the experiment, 1,000 healthy speech sequences were selected, along with 3 types of noise geometry (point, cross, ring) at 3 different distances (1, 3, and 5 meters) and with varying signal-to-noise ratios (–10, 0, and 10 dB). Each recording was simulated using all three beamforming methods (DaS, UDR, and SINR). The output consisted of 81,000 records in a .csv file with the following columns: method, geometry, distance, SNR, input SNR, output SNR, and delta SNR.

Next, graphs were generated to demonstrate the dependencies of the variables on external factors. Figures 3 - 5 shows the dependence of delta SNR on distance for the DaS, UDR, and SINR methods, respectively.

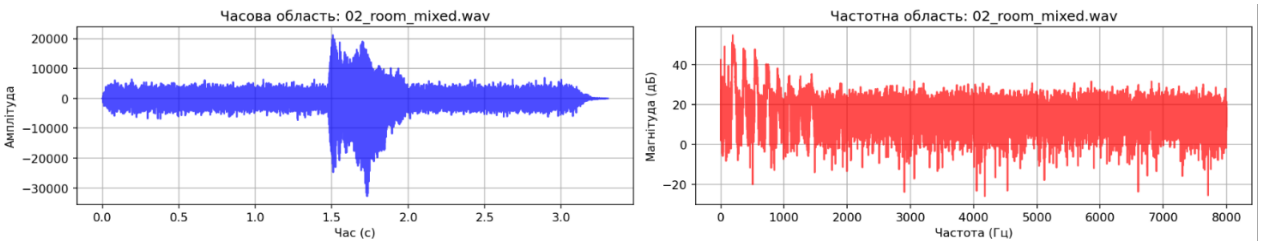
Fig. 3 shows an example of an audio signal from the dataset in the time and frequency domains.

Next, using the Pyroomacoustics library, a simulation of a room with dimensions of 15x15x5 was created, corresponding to the average size of open spaces in supermarkets or shopping malls. After adding noise and the desired signal, the overall audio signal changes (Fig. 4). The speech signal and noise signal are shown separately in Fig. 5 and 6, respectively. After applying the beamforming algorithm, noise is removed from the mixed signal and audio quality is improved (Fig. 7). The microphone array and algorithms ensure a stable spatial gain (array gain). This gain is a characteristic of the system's geometry and the mathematics of the algorithm, and it does not degrade when the total noise power changes. This manifests as the Delta SNR being independent of the initial signal-to-noise ratio. The graphs for conditions of –10 dB, 0 dB, and 10 dB are identical for each individual method.



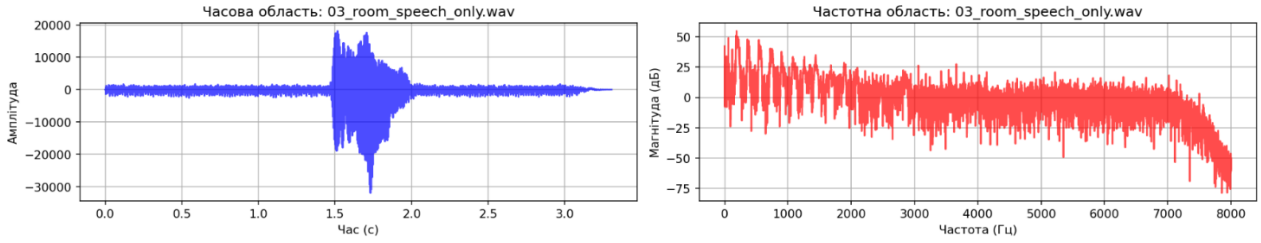
a b

Fig. 3. Visualization of the input audio signal: a – time domain, b – frequency domain



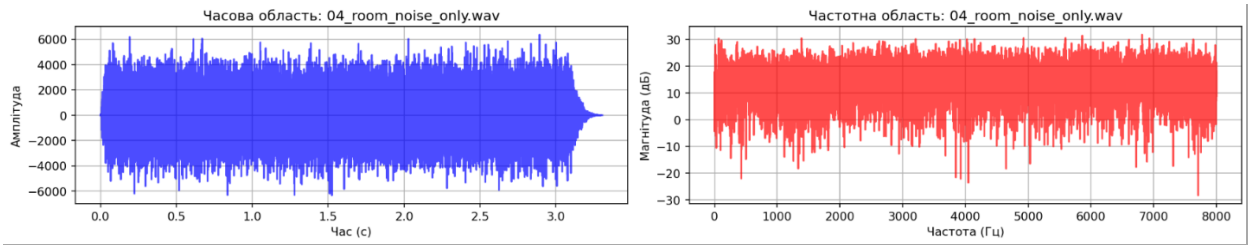
a b

Fig. 4. Visualization of a mixed audio signal in a simulation: a – time domain, b – frequency domain



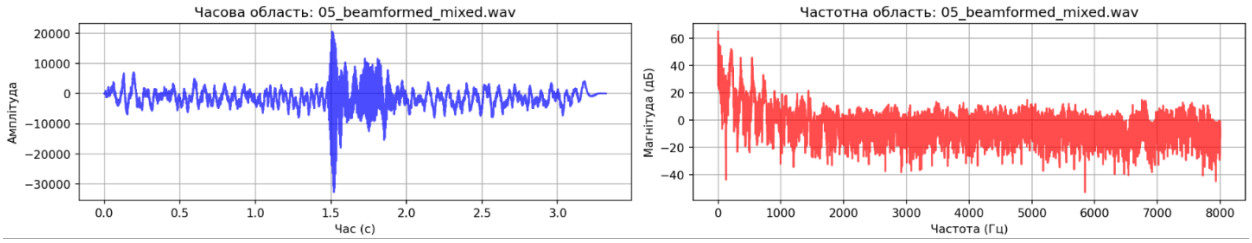
a b

Fig. 5. Visualization of a broadcast audio signal in a simulation: a) time domain, b) frequency domain



a b

Fig. 6. Visualization of a noise audio signal in a simulation: a – time domain, b – frequency domain



a b

Fig. 7. Visualization of the audio signal after applying the beamforming algorithm: a – time domain, b – frequency domain

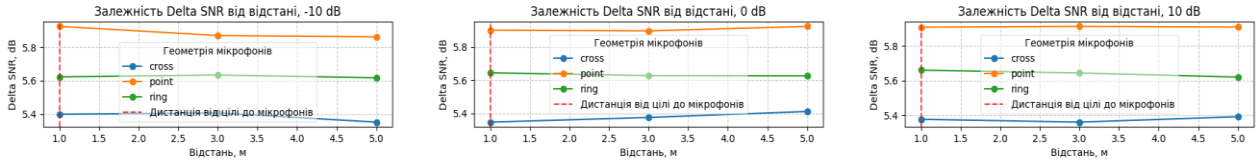
The results of experiments conducted for various beamforming methods under varying conditions in simulation mode are presented below.

All three algorithms demonstrate the same hierarchy of performance depending on how the noise sources are distributed. Point achieves the best result (highest delta SNR) among all methods. It is easiest for beamformers to focus the spatial "zero" in the beam pattern to suppress a single localized source. The ring configuration has average performance: noise surrounds the array, complicating the task compared to a single point, but the algorithms are still capable of filtering it reasonably well. The most challenging scenario for all methods is the cross configuration, which yields the smallest gain in the useful signal. This distributed configuration creates the most complex interference field for processing.

The algorithms differ significantly in terms of overall performance and response to the distance of noise sources. DaS shows the smallest improvement (delta SNR ranging from ~5.35 to ~5.9 dB, Fig. 8). It is the only method that demonstrates a dependence on the distance to the noise source. As point noise is moved from 1 to 5 meters away, performance slowly decreases, whereas for cross noise, it tends to increase slightly.

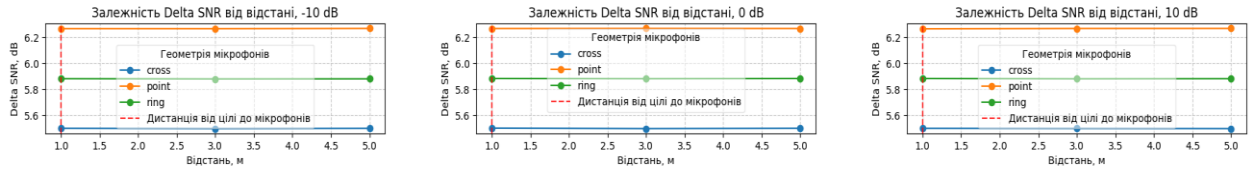
UDR shows average results (ranging from ~5.5 to ~6.25 dB), which are noticeably better than those of the classic DaS (Fig. 9). The graphs consist of horizontal lines. This means that the suppression performance is completely independent of the distance to the noise sources and depends solely on their spatial type (point, ring, cross).

SINR demonstrates the highest performance (delta SNR ranging from ~7.9 to ~9.1 dB, Fig. 10). The algorithm directly maximizes the signal-to-interference-plus-noise ratio and, as expected, performs best at optimizing the weights for interference suppression. As in the previous case, the result is maximally stable and does not depend on the distance to the noise source in the studied range.



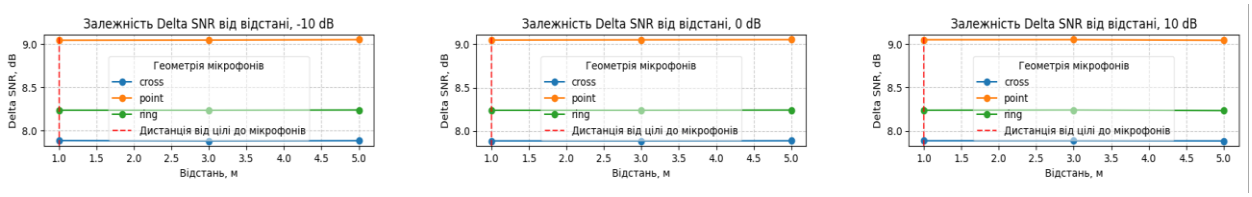
a b c

Fig. 8. Delta SNR as a function of distance, DaS method: a – louder, –10 dB; b – equal, 0 dB; c – quieter, 10 dB



a b c

Fig. 9. Delta SNR as a function of distance, UDR method: a – louder, –10 dB; b – equal, 0 dB; c – quieter, 10 dB



a b c

Fig. 10. Dependence of delta SNR on distance, SINR method: a – louder, –10 dB; b – equal, 0 dB; c – quieter, 10 dB

Next, the dependence of delta SNR on the noise level (SNR dB) was analyzed at distances of 1, 3, and 5 meters (Fig. 11–13). The distance from the target to the noise is 0, 2, and 4 meters, respectively.

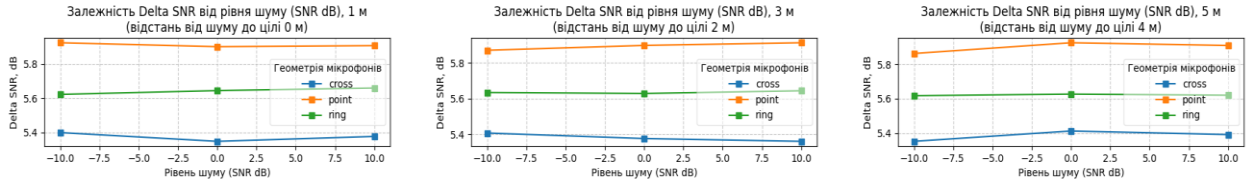
For all methods, the delta SNR value remains stable as the noise level changes. This confirms the linear nature of signal processing in these algorithms: the output signal-to-noise ratio changes in direct proportion to the input, so their difference (delta SNR) remains constant.

At any distance and at any noise level, the pattern of suppression efficiency remains consistent depending on the interference geometry: a point source is filtered best, a ring source is filtered worse, and a cross-shaped source is filtered worst. DaS demonstrates the smallest spatial gain (5.35–5.95 dB, Fig. 11). In the first graph (noise distance of 1 m, coinciding with the coordinates of the useful signal), slight fluctuations in delta SNR are noticeable. This is because DaS is a data-independent method with fixed weighting coefficients; it does not adapt to the environment, so nonlinear interference effects may occur when signal and noise sources are spatially coincident.

UDR demonstrates higher efficiency (5.5–6.25 dB, Fig. 12). The graphs appear as horizontal lines at all distances. Adaptive weight estimation relies on a normalized spatial covariance matrix, which depends on the location of the sources rather than their absolute power. Therefore, the algorithm is robust to changes in noise intensity.

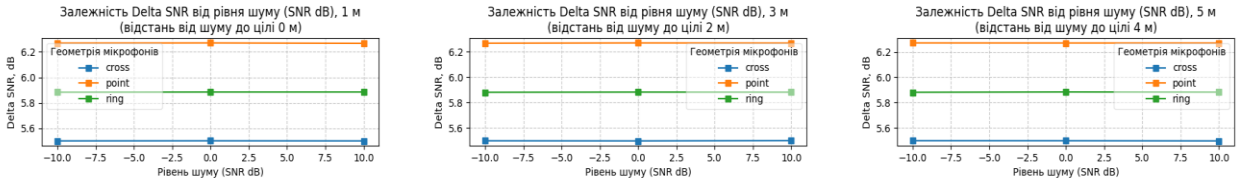
The SINR method demonstrates the highest efficiency (7.9–9.05 dB, Fig. 13). The algorithm also exhibits perfect linearity and stability regardless of the input

noise level and the distance to the noise source. Analytical maximization of the SINR criterion allows achieving the theoretical efficiency limit for a given microphone configuration.



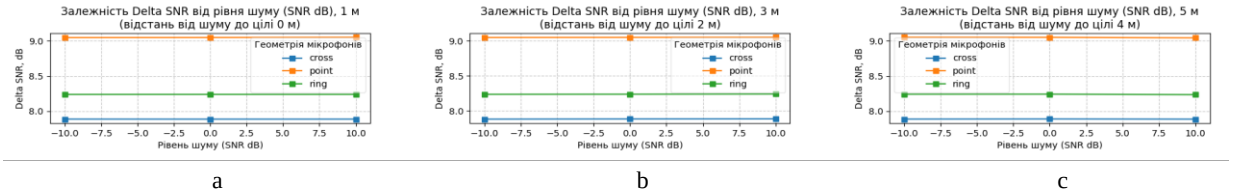
a b c

Fig. 11. Dependence of delta SNR on noise level, DaS method: a – 1 m, b – 3 m, c – 5 m



a b c

Fig. 12. Dependence of delta SNR on noise level, UDR method: a – 1 m, b – 3 m, c – 5 m

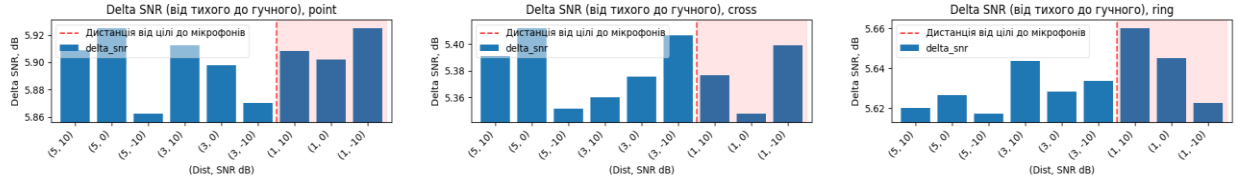


a b c

Fig. 13. Dependence of delta SNR on noise level, SINR method: a – 1 m, b – 3 m, c – 5 m

To confirm the previously established stability of adaptive algorithms for the given speech recognition task, the following histograms were created (Fig. 14–16).

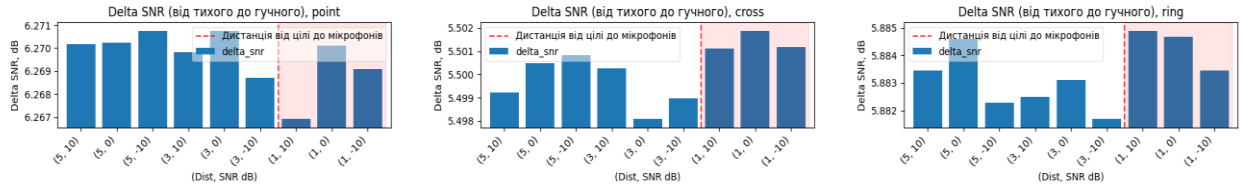
For UDR, Delta SNR fluctuations occur in the range of thousandths of a decibel (for example, from 6.267 to 6.271 dB for the point geometry, Fig. 15). For SINR, fluctuations are also within the range of thousandths or hundredths of a decibel (for example, from 9.0425 to 9.0525 dB for the point geometry, Fig. 16). Such changes are negligibly small. Visual fluctuations in the histograms are a consequence of the scaling of the Y-axis and reflect not a fundamental instability of the methods, but the limits of floating-point calculation accuracy and minor artifacts of the digital simulation of covariance matrices.



a b c

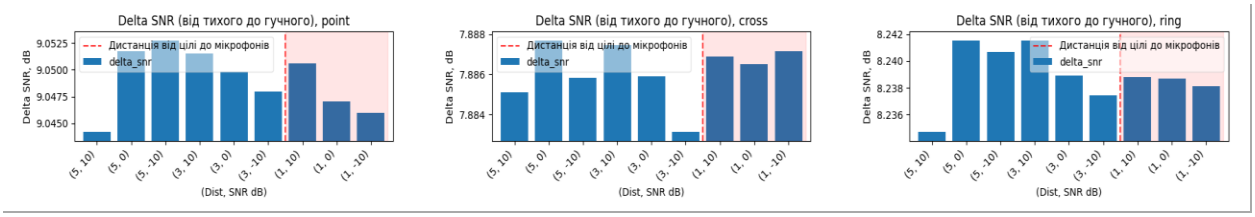
Fig. 14. The dependence of delta SNR on noise level and distance, DaS method: a – point interference topology, b – cross, c – ring

DaS is the only one to exhibit more noticeable fluctuations, reaching hundredths and tenths of a decibel (for example, from 5.86 to 5.92 dB, Fig. 14). The absence of adaptive weight calculation makes the algorithm sensitive to phase shifts that occur when the distance to noise sources and their intensity change. The area highlighted in pink (a distance of 1 m, coinciding with the location of the useful signal) exhibits specific behavior. When the interference is at the same radius as the target, spatial interference complicates the operation of the fixed directional pattern.



a b c

Fig. 15. The dependence of delta SNR on noise level and distance, UDR method: a – point interference topology, b –cross, c – ring



a b c

Fig. 16. The dependence of delta SNR on noise level and distance, SINR method: a – point interference topology, b – cross, c – ring

For the UDR and SINR methods, the baseline efficiency remains constant regardless of the combination of distance (1, 3, 5 meters) and input SNR (–10, 0, 10 dB). Even when noise is located in close proximity to the useful signal (at a distance of 1 m), adaptive algorithms retain their effectiveness. Minor deviations in this region (within thousandths of a dB) confirm the ability of spatial "zeros" to isolate interference regardless of their proximity to the useful signal's focal radius.

The results of the experiments for a single noise source are presented in Table 4. The results of the experiments for the four noise sources are presented in Table 5. The results of the experiments for eight noise sources are presented in Table 6.

Table 4 – **Evaluation of the effectiveness of beamforming methods as a function of distance and noise power in the presence of a point noise source**

Distance to the obstacle, m	Noise level, dB	SNR_{in} (Baseline)	Δ*SNR* (DAS)	Δ*SNR* (UDR)	Δ*SNR* (SINR)
1 m	+10 (Quiet)	44.67	5.91	6.27	9.05
1 m	0 (Equal)	34.67	5.90	6.27	9.05
1 m	–10 (Loudly)	24.67	5.93	6.27	9.05
3 m	+10 (Quiet)	44.67	5.91	6.27	9.05
3 m	0 (Equal)	34.67	5.90	6.27	9.05
3 m	–10 (Loudly)	24.67	5.87	6.27	9.05
5 m	+10 (Quiet)	44.67	5.91	6.27	9.04
5 m	0 (Equal)	34.67	5.92	6.27	9.05
5 m	–10 (Loudly)	24.67	5.86	6.27	9.05
The average change in SNR in a chaotic environment		34.67	5.90	6.27	9.05

Table 5 – **Evaluation of the effectiveness of beamforming methods as a function of distance and noise power (cross topology)**

Distance to the obstacle, m	Noise level, dB	SNR_{in} (Baseline)	Δ*SNR* (DAS)	Δ*SNR* (UDR)	Δ*SNR* (SINR)
1 m	+10 (Quiet)	45.14	5.38	5.50	7.89
1 m	0 (Equal)	35.14	5.35	5.50	7.89
1 m	–10 (Loudly)	25.14	5.40	5.50	7.89
3 m	+10 (Quiet)	45.14	5.36	5.50	7.89
3 m	0 (Equal)	35.14	5.38	5.50	7.89
3 m	–10 (Loudly)	25.14	5.41	5.50	7.88
5 m	+10 (Quiet)	45.14	5.39	5.50	7.89
5 m	0 (Equal)	35.14	5.41	5.50	7.89
5 m	–10 (Loudly)	25.14	5.35	5.50	7.89
The average change in SNR in a chaotic environment		35.14	5.38	5.50	7.89

Table 6 - **Evaluation of the effectiveness of beamforming methods as a function of distance and noise power (ring topology)**

Distance to the obstacle, m	Noise level, dB	SNR_{in} (Baseline)	Δ*SNR* (DAS)	Δ*SNR* (UDR)	Δ*SNR* (SINR)
1 m	+10 (Quiet)	45.44	5.66	5.88	8.24
1 m	0 (Equal)	35.44	5.64	5.88	8.24
1 m	–10 (Loudly)	25.44	5.62	5.88	8.24
3 m	+10 (Quiet)	45.44	5.64	5.88	8.24
3 m	0 (Equal)	35.44	5.63	5.88	8.24
3 m	–10 (Loudly)	25.44	5.63	5.88	8.24
5 m	+10 (Quiet)	45.44	5.62	5.88	8.23
5 m	0 (Equal)	35.44	5.63	5.88	8.24
5 m	–10 (Loudly)	25.44	5.62	5.88	8.24
The average change in SNR in a chaotic environment		35.44	5.63	5.88	8.24

Conclusions

This paper presents a system for processing and analyzing voice commands for inclusive smart carts, designed to operate in the dynamic, non-stationary noise environment of retail spaces.

An approach is proposed that uses microphone arrays and spatial filtering (beamforming) methods for continuous preprocessing of the audio stream to improve the signal-to-noise ratio (SNR) prior to the speech-to-text conversion stage.

Using spatial filtering algorithms, specifically Delay-and-Sum (DAS), Max-UDR, and Max-SINR, the system achieves high accuracy in extracting the useful signal in the presence of acoustic disturbances with various spatial topologies.

A comparative analysis of algorithm performance was conducted, the results of which showed that the Max-SINR method demonstrated the highest noise suppression efficiency (an SNR gain of 7.9 to 9.1 dB) compared to Max-UDR (5.5–6.25 dB) and DAS (5.35–5.95 dB).

The results confirm the effectiveness of the proposed approach for spatial signal selection, which is critically important in ensuring the reliability of voice control. The developed subsystem automatically compensates for the effects of background noise at the hardware level, significantly minimizing phonetic distortions and the vulnerability of the recognition system.

The use of adaptive beamforming algorithms, in particular Max-SINR, ensures mathematical stability of operation regardless of the strength of interference and the distance to it, making the solution promising for implementation in assistance and spatial orientation systems for people with visual impairments. Further research will focus on conducting field experiments using real hardware in variable outdoor conditions.

Conflicts of interest

The authors declare that they have no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study, as well as the results reported in this paper.

Use of artificial intelligence

The authors confirm that they did not use artificial intelligence technologies when creating the current work.

REFERENCES

1. O. Barkovska, A. Havrashenko and P. Botnar, "The influence of reverberation, equalization and compression methods on speaker recognition," *2025 IEEE 6th KhPI Week on Advanced Technology (KhPIWeek)*, Kharkiv, Ukraine, 2025, pp. 1-5, doi: https://doi.org/10.1109/KhPIWeek61436.2025.11288718
2. Kulkarni, S., Thakur, A., Soni, S., Hiwale, A., Belsare, M. H., & Raj, A. B. (2025). A comprehensive review of direction of arrival (DoA) estimation techniques and algorithms. *Journal of Electronics and Electrical Engineering*, 138-186. https://doi.org/10.37256/jeee.4120255708
3. H. A. Kassir, Z. D. Zaharis, P. I. Lazaridis, N. V. Kantartzis, T. V. Yioultsis and T. D. Xenos, "A Review of the State of the Art and Future Challenges of Deep Learning-Based Beamforming," in *IEEE Access*, vol. 10, pp. 80869-80882, 2022, doi: https://doi.org/10.1109/ACCESS.2022.3195299
4. Barkovska Olesia, Vitalii Serdechnyi. Intelligent Assistance System for People with Visual Impairments. *Innovative technologies and scientific solutions for industries*, no. 2(28), June 2024, pp. 6–16. https://doi.org/10.30837/2522-9818.2024.28.006
5. Barkovska, O., Holovchenko, O., Storchai, D., Kostin, A., & Lehezin, N. (2025). Investigation of computer vision techniques for indoor navigation systems. Innovative technologies and scientific solutions for industries, (2(32), 5–15. https://doi.org/10.30837/2522-9818.2025.2.005
6. Xi, J., Xu, Z., Zhang, W., Xie, Y., & Zhao, L. (2025). Speech Enhancement Algorithm Based on Microphone Array and Multi-Channel Parallel GRU-CNN Network. Electronics, 14(4), 681. https://doi.org/10.3390/electronics14040681
7. Wang, J.-H., Le, P. T., Bee, W.-S., Putri, W. R., Su, M.-H., Li, K.-C., Chen, S.-L., He, J.-L., Pham, T., Li, Y.-H., & Wang, J.-C. (2024). Implementation of Sound Direction Detection and Mixed Source Separation in Embedded Systems. Sensors, 24(13), 4351. https://doi.org/10.3390/s24134351
8. Wang, J.-H., Le, P. T., Kuo, S.-J., Tai, T.-C., Li, K.-C., Chen, S.-L., Wang, Z.-Y., Pham, T., Li, Y.-H., & Wang, J.-C. (2024). Audio Pre-Processing and Beamforming Implementation on Embedded Systems. Electronics, 13(14), 2784. https://doi.org/10.3390/electronics13142784
9. Huang, P., Ullah, I., Wei, X., Ahamed, A. T., Hassan, N., & Shah, Z. H. (2025). Towards Energy-Efficient and Low-Latency Voice-Controlled Smart Homes: A Proposal for Offline Speech Recognition and IoT Integration. ArXiv.org. https://arxiv.org/abs/2506.07494
10. Ciccarelli, G., Barber, J., Nair, A., Cohen, I., & Zhang, T. (2022). Challenges and Opportunities in Multi-device Speech Processing. ArXiv.org. https://arxiv.org/abs/2206.15432
11. Haeb-Umbach, R., Heymann, J., Drude, L., Watanabe, S., Delcroix, M., & Nakatani, T. (2020). Far-Field Automatic Speech Recognition. ArXiv.org. https://arxiv.org/abs/2009.09395
12. Rascon, C. (2021). A Corpus-Based Evaluation of Beamforming Techniques and Phase-Based Frequency Masking. Sensors, 21(15), 5005. https://doi.org/10.3390/s21155005
13. Rowe, H. P., Gutz, S. E., Maffei, M. F., Tomanek, K., & Green, J. R. (2022). Characterizing Dysarthria Diversity for Automatic Speech Recognition: A Tutorial From the Clinical Perspective. Frontiers in Computer Science, 4. https://doi.org/10.3389/fcomp.2022.770210
14. Luria, M., Hoffman, G., & Zuckerman, O. (2017). Comparing Social Robot, Screen and Voice Interfaces for Smart-Home Control. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3025453.3025786
15. B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," in IEEE ASSP Magazine, vol. 5, no. 2, pp. 4-24, April 1988, doi: https://doi.org/10.1109/53.665

16. C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 24, no. 4, pp. 320-327, August 1976, doi: https://doi.org/10.1109/TASSP.1976.1162830
17. Rakerd, B., Hartmann, W.M. (2005). Localization of noise in a reverberant environment. In: Pressnitzer, D., de Cheveigné, A., McAdams, S., Collet, L. (eds) Auditory Signal Processing. Springer, NY. https://doi.org/10.1007/0-387-27045-0_51
18. J. Capon, "High-resolution frequency-wavenumber spectrum analysis," in Proceedings of the IEEE, vol. 57, no. 8, pp. 1408-1418, Aug. 1969, doi: https://doi.org/10.1109/PROC.1969.7278
19. R. Schmidt, "Multiple emitter location and signal parameter estimation," in IEEE Transactions on Antennas and Propagation, vol. 34, no. 3, pp. 276-280, March 1986, doi: https://doi.org/10.1109/TAP.1986.1143830
20. R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 37, no. 7, pp. 984-995, July 1989, doi: https://doi.org/10.1109/29.32276
21. Rudzicz, F., Namasivayam, A.K. & Wolff, T. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. Lang Resources & Evaluation 46, 523–541 (2012). https://doi.org/10.1007/s10579-011-9145-0

Received (Надійшла) 15.01.2026
Accepted for publication (Прийнята до друку) 22.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Раптанов Даніїл Андрійович - магістрант кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Daniil Raptanov - master's student of the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.
e-mail: daniil.raptanov@nure.ua, ORCID Author ID: http://orcid.org/0009-0001-9564-0080 .

Барковська Олеся Юріївна – кандидат технічних наук, доцент кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Olesia Barkovska – Candidate of Technical Sciences, Associate Professor at the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine;
e-mail: olesia.barkovska@nure.ua; ORCID Author ID: http://orcid.org/0000-0001-7496-4353;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=24482907700

Шиленко Михайло Павлович – магістрант кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Mykhailo Shylenko - master's student of the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.
e-mail: mykhailo.shylenko@nure.ua; ORCID Author ID: http://orcid.org/0009-0009-4084-4711.

Головченко Олександр Сергійович - аспірант кафедри Електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Oleksandr Holovchenko – Phd student of Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.
e-mail: oleksandr.holovchenko@nure.ua; ORCID Author ID: https://orcid.org/0009-0002-7582-1746.

Івахненко Діана Сергіївна - бакалавр кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Diana Ivakhnenko - bachelor's student of the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.
e-mail: diana.ivakhnenko@nure.ua; ORCID Author ID: https://orcid.org/0009-0005-0989-0016.

Дослідження точності роботи методів бімформінгу в контексті інклюзивної системи внутрішньої навігації

Д. А. Раптанов, О. Ю. Барковська, М.П. Шиленко, О. С. Головченко, Д. С. Івахненко

Анотація. Актуальність. Голосове керування елементами інклюзивних навігаційних систем є критично важливим для забезпечення автономності та безпечної мобільності людей з порушеннями зору в громадських місцях, зокрема у великих торговельних залах. Однак існуючі системи перетворення мовлення на текст (STT) стикаються із суттєвим зниженням точності розпізнавання через високодинамічний та нестаціонарний акустичний шум супермаркетів. **Об'єктом дослідження** є препроцесинг аудіопотоку та просторова фільтрація (бімформінг) у системі голосового керування за умов динамічного нестаціонарного шуму. Проблема полягає у недостатній вибірковості стандартних алгоритмів обробки аудіосигналу в умовах фонових завад магазину, що призводить до критичного зростання частки помилок у розпізнаних словах (WER) та робить систему управління розумним візком вразливою.. **Метою статті** є оцінка впливу зовнішніх факторів (кількості, просторової топології розміщення та рівня потужності джерел акустичного шуму) на точність методів просторової фільтрації (бімформінгу) для подальшого розпізнавання голосових команд шляхом комп'ютерного моделювання. **В результаті** роботи за допомогою бібліотеки Pyroomacoustics було змодельовано акустичне середовище та мікрофонну решітку. Проведено порівняння трьох методів: Delay-and-Sum (DAS), Max-UDR та Max-SINR. Дослідження показало, що алгоритм Max-SINR забезпечує найвищий приріст співвідношення сигнал/шум (delta SNR від 7,9 до 9,1 дБ) і є математично стійким до змін відстані до завад та їхньої потужності. Метод DAS виявився найменш ефективним (5,35–5,95 дБ) і продемонстрував чутливість до зміни дистанції. Встановлено, що ключовим фактором деградації сигналу є конфігурація джерел шуму, серед яких перехресна топологія (cross) є найскладнішою для фільтрації.

Ключові слова: інклюзивна система навігації, порушення зору, розпізнавання мовлення, просторова фільтрація, бімформінг, Delay-and-Sum, Max-UDR, Max-SINR, динамічний шум.