Міністерство освіти і науки України
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Міністерство освіти і науки України
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

Кваліфікаційна наукова
праця на правах рукопису

Кухарічева Катерина Андріївна

УДК 621.391.83

ДИСЕРТАЦІЯ
ПІДВИЩЕННЯ РОБАСТНОСТІ СИСТЕМ АВТОМАТИЧНОГО РОЗПІЗНАВАННЯ МОВИ ДО ДІЇ ЗАВАД

17 – Електроніка та телекомунікації
171 – Електроніка

Подається на здобуття наукового ступеня доктора філософії.

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело.
_____/ Кухарічева К.А.

Науковий керівник Продеус Аркадій Миколайович, доктор технічних наук, професор

Київ - 2023

АНОТАЦІЯ

Кухарічева К.А. Підвищення робастності систем автоматичного розпізнавання мови до дії завад. – Кваліфікаційна робота на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 171 «Електроніка». – Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», МОН України, Київ, 2023.

Дисертаційна робота присвячена дослідженню методів навчання систем автоматичного розпізнавання мовлення (АРМ) та методів оцінювання якості мовних сигналів, що забезпечують підвищення точності систем автоматичного розпізнавання мовлення без суттєвого ускладнення налаштування таких систем.

Зміст дисертаційного дослідження викладено в чотирьох розділах, де представлено та обґрунтовано основні результати роботи.

Актуальність дисертаційної роботи обґрунтовано у вступі, де сформульовано мету та задачі дослідження, описано методи дослідження, надано інформацію про наукову новизну та практичне значення одержаних результатів.

У першому розділі виконано огляд існуючих підходів до підвищення робастності систем АРМ до дії шумової та реверберaційної завад. Описано два напрями підвищення точності розпізнавання: напрям, що базується на попередній корекції сигналу та напрям, згідно якому виконується адаптація системи до дії завад. В існуючих наукових працях, присвячених вивченню систем АРМ, недостатньо вивчено напрям, згідно якому виконується адаптація системи до дії завад шляхом навчання на сигналах, спотворених завадами.

У другому розділі наведено результати дослідження таких факторів як реверберація та кліпування сигналу, що можуть істотно вплинути на ефективність роботи системи АРМ. Розглянуто міри визначення величини кліпування, особливості використання об’єктивних показників розбірливості

мовлення та запропоновано способи моделювання реверберації в приміщенні, що є корисним при створенні систем АРМ, стійких до дії завад.

У третьому розділі представлено короткий огляд мір якості систем автоматичного розпізнавання мови й, зокрема, мір якості, що використовуються при оцінюванні точності розпізнавання в програмному комплексі The Hidden Markov Model Toolkit (HTK). Також представлено результати експериментальних досліджень, спрямованих на підвищення робастності систем АРМ до дії шумової завади. При цьому отримано оцінки потенційних можливостей різних сполучень режимів навчання та роботи систем АРМ.

У четвертому розділі представлено результати експериментальних досліджень, спрямованих на підвищення стійкості систем АРМ до дії реверберaційної завади. При цьому визначено ефективність роботи системи для різних варіантів навчання та роботи за умов спотворення сигналів реверберaційною завадою.

Представлені в дисертації нові теоретичні та практичні результати можна рекомендувати до використання при розробці та експлуатації систем автоматичного розпізнавання мовлення, а також в навчальному процесі вищих навчальних закладів України для підготовки інженерів-акустиків. Отримані результати вже застосовано в освітньому процесі кафедри акустичних та мультимедійних електронних систем за спеціальністю 171 Електроніка, зокрема в освітній науковій програмі «Електроніка», а також в освітній професійній програмі “Акустичні електронні системи та технології обробки акустичної інформації” Національного технічного університету України “Київський політехнічний інститут імені Ігоря Сікорського”.

В дисертаційній роботі отримано наступні наукові результати:

1. Вперше для реальних мовленнєвих сигналів отримано кількісні оцінки ступеня підвищення точності розпізнавання мовлення, спотвореного шумом різної природи та інтенсивності, шляхом

навчання системи автоматичного розпізнавання на спотворених шумом сигналах.

2. Вперше для реальних мовленнєвих сигналів отримано кількісні оцінки ступеня підвищення точності розпізнавання мовлення, спотвореного реверберацією, шляхом навчання системи автоматичного розпізнавання на спотворених реверберацією сигналах.
3. Вдосконалено метод оцінювання розбірливості мовлення непрямим методом, із використанням міри якості сигналів у вигляді барківського спектрального спотворення.
4. Уточнено висновки щодо залежності розбірливості мовлення від щільності відбить звуку та часу реверберації, із використанням імовірнісних моделей імпульсних характеристик приміщень.
5. Вдосконалено спосіб виявлення ефекту кліпування мовленнєвих сигналів та об’єктивного оцінювання якості мовленнєвих сигналів, спотворених кліпуванням, що базується на використанні коефіцієнта ексцесу як міри спотворення сигналів.

Практичне значення отриманих результатів полягає у наступному:

1. Встановлено умови досягнення високої точності розпізнавання в системах автоматичного розпізнавання мовлення для сигналів, що спотворені шумами різної природи та інтенсивності, за наявності різної апріорної інформації щодо відношення сигнал-шум, що дозволяє забезпечити робастність системи автоматичного розпізнавання шляхом відносно простого її налаштування;
2. Встановлено умови досягнення високої точності розпізнавання в системах автоматичного розпізнавання мовлення для сигналів, спотворених реверберацією в приміщеннях із різним часом реверберації, за наявності різної апріорної інформації щодо часу реверберації, що дозволяє забезпечити робастність системи

автоматичного розпізнавання шляхом використання певних правил її налаштування;

3. Встановлено працездатність та ефективність оцінювання розбірливості мовлення непрямим методом, із використанням міри якості сигналів у вигляді барківського спектрального спотворення, що дозволяє оцінювати розбірливість мовлення, спотвореного реверберацією, за наявності еталонного неспотвореного сигналу;
4. Отримано залежності розбірливості мовлення від щільності ранніх відбить звуку та часу реверберації, із використанням імовірнісних моделей імпульсних характеристик приміщень, що дозволяє обґрунтувати результати прогнозування та оцінювання розбірливості мовлення в різних точках приміщення;
5. Встановлено можливість автоматизації виявлення кліпування, оцінювання його ступеня, а також об’єктивного оцінювання якості мовленнєвих сигналів, спотворених кліпуванням.

Ключові слова: автоматичне розпізнавання мовлення, точність розпізнавання, прихована марковська модель, шумова завада, реверберaційна завада, коефіцієнт ексцесу.

Список публікацій здобувача

1. Prodeus A., Didkovska M., Kukharicheva K. Comparison of Speech Quality and Intelligibility Assessments in University Classrooms // International Journal of Architectural Engineering Technology. 2021. Vol. 8. P. 52–60. https://doi.org/10.15377/2409-9821.2021.08.5
2. Prodeus A., Didkovska M., Kukharicheva K, Motorniuk D. Two Simplified Models Of Early Sound Reflections In a Room // Electronics and Control Systems. 2020. Vol. 3, (65). https://doi.org/10.18372/1990-5548.65.14991
3. Prodeus A., Kotvytskyi I., Didkovska M., Kukharicheva K. Kurtosis and Normalized Variance as Measures of Speech Signals Clipping Value //

Electronics and Control Systems, 2019. Vol. 4 (62). P.24-32. https://doi.org/10.18372/1990-5548.62.14378

4. Prodeus A., Kukharicheva K. Accuracy of Automatic Speech Recognition System Trained on Noised Speech // Electronics and Control Systems. 2016. Vol. 3 (49). https://doi.org/10.18372/1990-5548.49.11230
5. Prodeus A., Didkovska M., Kukharicheva K. Impact of University Classroom Size on the Relationship Between Speech Quality and Intelligibility // International Journal of Computing. 2022. Vol. 21 (3). https://doi.org/10.47839/ijc.21.3.2690
6. А. М. Продеус, І. В. Котвицький, М. В. Дідковська, В. С. Дідковський, К. А. Кухарічева, Д. Є. Моторнюк, О. О. Дворник Спосіб виявлення кліпування мовного та музичного сигналів // Патент UA 144291 U, МПК G01R 23/20, опубл. 25.09.2020.

SUMMARY

Kukharicheva K.A. Increasing the Robustness of Automatic Speech Recognition Systems to Interference – Qualification scientific work on the rights of the manuscript.

Thesis for the degree of Philosophy Doctor, in specialty 171 “Electronics”. – National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, 2023.

The dissertation is devoted to the study of methods of training automatic speech recognition systems and methods of evaluating the quality of speech signals, which ensure an increase in the accuracy of automatic speech recognition systems without significantly complicating the setup of such systems.

The content of the dissertation research is presented in four sections, in which the main results of the work are presented and substantiated.

The introduction substantiates the relevance of the dissertation, formulates the purpose and lists the objectives of the study, describes the research methods, and provides information about the scientific novelty and practical significance of the results.

The first section is devoted to a review of publications on increasing the robustness of ASR systems to noise and reverberation interference. Two methods of the recognition increase are described: signal correction methods and model adaptation approaches. The considered works on increasing the robustness of ASR systems lack knowledge of model adaptation approach in terms of the usage of different training techniques.

The second section presents the research results of the reverberation and signal clipping, which can significantly affect the efficiency of the ASR system. Measures for determining the amount of clipping, and features of the usage of speech intelligibility objective measures are considered, and the methods of reverberation modeling in the room are proposed, which are useful while creating ASR systems that are robust to the effects of interference.

The third section presents a brief overview of the quality measures of ASR systems and, in particular, the quality measures used in the assessment of recognition accuracy in The Hidden Markov Model Toolkit (HTK). The results of experimental studies aimed at increasing the robustness of APM systems to the effect of noise interference are also presented. At the same time, the estimates of the potential capabilities of training modes' various combinations of the ASR systems were obtained.

The fourth section presents the results of experimental studies aimed at increasing the ASR system's robustness to the impact of reverberation interference. At the same time, the efficiency of the system was determined for various modes of training and operation under reverberation interference.

The new theoretical and practical results presented in the dissertation can be recommended for the development and operation of ASR systems, as well as in the educational process of higher educational institutions of Ukraine for the acoustic engineering field of study. Results obtained are already implemented in the educational process at the Department of Acoustic and Multimedia Electronic Systems (specialty 171 "Electronics", educational programme "Acoustic Electronic Systems and Acoustic Information Processing Technologies) of National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

The following scientific results were obtained in the dissertation:

1. For the first time, the quantitative estimates of the degree of improvement in the speech recognition accuracy for the real speech signals distorted by the noise of different nature and intensity were obtained by training an automatic recognition system on noise-distorted signals.
2. For the first time, the quantitative estimates of the degree of improvement in the speech recognition accuracy for the real speech signals distorted by the noise of different nature and intensity were obtained by training an automatic recognition system on reverberant signals

3. The indirect method of assessing speech intelligibility using a signal quality measure in the form of Barkov spectral distortion has been improved.
4. The conclusions regarding the dependence of speech intelligibility on the sound reflections density and reverberation time have been refined, using probabilistic room impulse response (RIR) models.
5. The method of detecting the speech signals clipping effect and objective assessment of the quality of speech signals distorted by clipping, based on the use of the kurtosis coefficient as a signal distortion measure, has been improved.

The practical significance of the results obtained in the dissertation is as follows:

1. The conditions for achieving high recognition accuracy in automatic speech recognition systems for signals distorted by noises of different nature and intensity, in the presence of different a priori information about the signal-to-noise ratio, have been established, which allows ensuring the robustness of the automatic recognition system through relatively simple configuration;
2. The conditions for achieving high recognition accuracy in automatic speech recognition systems for signals distorted by reverberation in rooms with different reverberation time values have been established, which allows for ensuring the robustness of the automatic recognition system through relatively simple configuration;
3. The efficiency and effectiveness of evaluating speech intelligibility by an indirect method, using a measure of signal quality in the form of Bark spectral distortion, which allows evaluating the intelligibility of speech distorted by reverberation, in the presence of a reference undistorted signal, has been established;
4. The dependence of speech intelligibility on the density of early sound reflections and reverberation time was obtained, using probabilistic RIR models, which allows substantiating the results of forecasting and evaluation of speech intelligibility at different points of the room;

5. The possibility of clipping detection automatization, the assessment of its degree, and the objective assessment of the quality of speech signals distorted by clipping have been obtained.

Keywords: automatic speech recognition, recognition accuracy, hidden Markov model, noise interference, reverberation interference, kurtosis.

List of applicant's publications

1. Prodeus A., Didkovska M., Kukharicheva K. Comparison of Speech Quality and Intelligibility Assessments in University Classrooms // International Journal of Architectural Engineering Technology. 2021. Vol. 8. P. 52–60. https://doi.org/10.15377/2409-9821.2021.08.5
2. Prodeus A., Didkovska M., Kukharicheva K, Motorniuk D. Two Simplified Models Of Early Sound Reflections In a Room // Electronics and Control Systems. 2020. Vol. 3, (65). https://doi.org/10.18372/1990-5548.65.14991
3. Prodeus A., Kotvytskyi I., Didkovska M., Kukharicheva K. Kurtosis and Normalized Variance as Measures of Speech Signals Clipping Value // Electronics and Control Systems, 2019. Vol. 4 (62). P.24-32. https://doi.org/10.18372/1990-5548.62.14378
4. Prodeus A., Kukharicheva K. Accuracy of Automatic Speech Recognition System Trained on Noised Speech // Electronics and Control Systems. 2016. Vol. 3 (49). https://doi.org/10.18372/1990-5548.49.11230
5. Prodeus A., Didkovska M., Kukharicheva K. Impact of University Classroom Size on the Relationship Between Speech Quality and Intelligibility // International Journal of Computing. 2022. Vol. 21 (3). https://doi.org/10.47839/ijc.21.3.2690
6. A. M. Prodeus, I. V. Kotvitskyi, M. V. Didkovska, V. S. Didkovskyi, K. A. Kukharicheva, D. E. Motorniuk, O. O. Dvornyk Method of detecting clipping of speech and music signals // Patent RU 144291 U, IPC G01R 23/20, publ. 09/25/2020

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

ART, All-Reverb Training – навчання на всіх зареверберованих сигналах

ASR, Automatic Speech Recognition – автоматичне розпізнавання мови

BSD, Bark Spectral Distortion – Барк-спектральне спотворення

CNN, Convolutional Neural Network – згорткова нейронна мережа

CPC, Contrasive Predictive Coding – контрастивне передбачення кодування

DMOS, Degradation Mean Opinion Score – шкала деградації середньої суб’єктивної оцінки

DNN, Deep Neural Network – глибинна нейронна мережа

EDT, Early Decay Time – час приходу ранніх відбиттів

FMT, Fully-Matched Training – навчання при повному співпадінні значень навчальної та тестувальної вибірок

FWSNR, Frequency-Weighted Segmental Signal-to-Noise Ratio – частотно-зважене сегментне відношення сигнал-завада

GMM, Gaussian Mixture Model – модель Гауссівської суміші

HTK, Hidden Markov Model Toolkit – інструментарій прихованої марковської моделі

LPC, Linear Predictive Coding – кодування з лінійним предиктором

LSD, Log-Spectral Distance – логарифмічна спектральна відстань

MER, Match Error Rate – показник помилки зіставлення

MFCC, Mel-frequency cepstral coefficients – Мел-частотні кепстральні коефіцієнти

MOS, Mean Opinion Score – усередненна оцінка розбірливості мовлення

PESQ, Perceptual Evaluation of Speech Quality – перцептуальна оцінка якості мовлення

PLP, Perceptual Linear Prediction – перцептивне лінійне прогнозування

RIR, Room Impulse Response – імпульсна характеристика приміщення

RMT, Reverb-Matched Training – навчання при співпадінні часу

реверберації навчальної та тестувальної вибірок

RNN, Reccurent Neural Network – рекурентна навчальна мережа

SNR, Signal-to-Noise Ratio – відношення сигнал-завада

SNRMT, Signal-to-Noise Ratio Matched Training – навчання при співпадінні значень відношення сигнал-завада навчальної та тестувальної вибірок

SSNR, Segmental Signal-to-Noise Ratio – сегментне відношення сигнал-завада

STI, Speech Transmission Index – індекс передачі мовлення

TDNN – Time-Delay Neural Network – нейронна мережа з часовою затримкою

WER, Word Error Rate – показник помилкових слів

WIL, Word Information Lost – втрачені слова

WIP, Word Information Preserved – збережені слова

АРМ – автоматичне розпізнавання мови

АЧХ – амплітудно-частотна характеристика

БІХ – бінауральна імпульсна характеристика

ГМ – гучномовець-мікрофон

ІХ – імпульсна характеристика

НМ – нейронна мережа

ПММ – прихована марковська модель

САР – система автоматичного розпізнавання

ВСТУП

Актуальність роботи. Дисертаційна робота направлена на підвищення робастності (стійкості) систем автоматичного розпізнавання мовлення (АРМ) до негативної дії таких завад як шум та реверберація. Суттєвим недоліком систем АРМ є чутливість до спотворень мовленнєвого сигналу, спричинена різницею властивостей завад в режимах навчання та роботи. Зрештою, це призводить до суттєвого зниження точності розпізнавання, аж до фактичної втрати працездатності систем АРМ. Цього недоліку можна частково позбутися, навчаючи систему АРМ на чистих (неспотворених) сигналах, а в режимі роботи відновлюючи спотворені сигнали перед подаванням їх на вхід систем АРМ. Альтернативним напрямом підвищення робастності систем АРМ є таке налаштування систем АРМ, що робить їх більш стійкими до дії спотворень. Якщо перший напрям, із двох вищевказаних, є досить очевидним та достатньо вивченим, цього не можна сказати про другий напрям, що можна пояснити існуванням низки проблем.

По-перше, суттєвою проблемою є різноманітність завад, таких як шум, реверберація, відлуння, амплітудні та фазові спотворення, похибки кодування, залишкове спотворення сигналів при використанні алгоритмів відновлення спотворених сигналів тощо.

По-друге, налаштування систем АРМ є дуже непростим завданням, що пояснюється надзвичайно високою складністю алгоритмів навчання та розпізнавання, що використовуються в систем АРМ. Така складність спричинена низкою причин. Це, зокрема, проблема вибору оптимальних в заданому сенсі класифікаційних ознак. Іншою проблемою є вибір елементів мовлення, що піддаються розпізнаванню – якщо в якості таких обрано фонеми, тоді кількість класів сягає кількох сотень, а при обранні складів в якості елементів мови кількість класів обчислюється тисячами. На точність розпізнавання також, зокрема, впливає такий чинник як спосіб апроксимації законів розподілу класифікаційних ознак. Зрештою, висока складність та

вартість існуючих систем систем АРМ робить актуальною розробку надійних систем систем АРМ, стійких до дії завад різної природи та відносно простих у налагодженні.

Значний внесок у дослідження та розв’язання задач підвищення робастності систем АРМ зробили такі іноземні та вітчизняні вчені, як Рабінер Л., Янг Б., Лі С., Фрай Д., Сузукі Д., Нагата К., Мартін Т., Нельсон А., Редді Д., Джелінек Ф., Ітакура А., Левенсон С., Розенберг А., Уілпон Д., Джуанг Б., Віртанен Т., Хуанг Х., Асіро А., Хон Х., Свєтоянскі П., Ванг Х., Рагні А., Повей Д., Баум Л., Вінцюк Т.К., Пилипенко В.В., Величко В.М., Загоруйко Н.Г. Проте аналіз праць, що стосуються адаптації системи АРМ до дії шумової та реверберaційної завад, виявив певні прогалини в дослідженні точності розпізнавання систем АРМ з точки зору зміни стилю навчання.

Таким чином, розробка нових та вдосконалення існуючих методів підвищення точності систем автоматичного розпізнавання мовлення, працездатність яких зберігається, незважаючи на негативну дію завад різної природи, й разом із тим відносно простих для практичної реалізації, є **актуальною** науково-технічною задачею автоматичного розпізнавання мовлення, що має важливе прикладне значення. Цим визначаються актуальність та практичне значення теми дисертаційного дослідження.

Мета і завдання дослідження. *Метою дисертаційної роботи* є розробка нових та вдосконалення відомих методів навчання систем автоматичного розпізнавання мовлення, а також методів оцінювання якості та розбірливості мовленнєвих сигналів, що забезпечують підвищення точності систем автоматичного розпізнавання мовлення без суттєвого ускладнення процедури налаштування таких систем.

Об’єктом дослідження є процес навчання систем автоматичного розпізнавання мовлення із врахуванням об’єму та характеру апріорної інформації про параметри шумової або реверберaційної завади.

Предметом дослідження є вплив об’єму та характеру апріорної інформації про параметри шумової або реверберaційної завади на точність автоматичного розпізнавання мовлення.

Для досягнення поставленої мети необхідно було вирішити такі завдання:

1. Виконати аналітичний огляд сучасних методів автоматичного розпізнавання мовлення, звернувши при цьому першочергову увагу на причини порушення робастності систем АРМ до дії шуму та реверберації, а також на перспективні шляхи відновлення такої робастності.

2. Встановити зв’язок між об’єктивними мірами розбірливості та якості мовленнєвих сигналів, спотворених реверберацією, а також виявити таку об’єктивну міру якості, яку можна було б використовувати як міру розбірливості в навчальних приміщеннях різного розміру.

3. Встановити зв’зок між розбірливістю мовлення, спотвореного реверберацією, та такими параметрами реверберaційної завади як час реверберації та щільність ранніх відбить звуку.

4. Дослідити потенційні можливості використання коефіцієнта ексцесу в якості міри ступеня кліпування мовленнєвого сигналу, а також в якості маркера наявності такого кліпування, що сприймається людською слуховою системою.

5. Отримати кількісні оцінки ступеню підвищення точності розпізнавання мовлення, спотвореного шумом різної природи та інтенсивності, шляхом навчання системи автоматичного розпізнавання на сигналах, спотворених шумом, із врахуванням об’єму та характеру апріорної інформації про шумову заваду.

6. Встановити принципову можливість підвищення робастності систем автоматичного розпізнавання мовлення до дії реверберації шляхом навчання системи автоматичного розпізнавання на сигналах, спотворених

реверберацією, із врахуванням об’єму та характеру апріорної інформації про реверберацiйну заваду.

Методи дослідження. Для досягнення поставленої мети використано методи аналітичного та комп’ютерного моделювання, а також методи експериментального оцінювання суб’єктивних та об’єктивних мір якості та розбірливості спотвореного мовлення. При оцінюванні ступеня спотворення сигналів кліпуванням застосовано статистичний метод опису законів розподілу випадкових процесів із використанням моментів цих процесів. Суб’єктивне оцінювання розбірливості спотвореного мовлення виконувалося артикуляційним методом, а суб’єктивне оцінювання якості спотвореного мовлення виконувалося інтрузивним методом. При об’єктивному оцінюванні розбірливості мовлення, спотвореного ранніми відбиттями, використано модуляційний метод. Імпульсні характеристики приміщень при цьому сформовано шляхом імітаційного моделювання із використанням методу Монте-Карло. При моделюванні системи автоматичного розпізнавання мовлення суттєво використано методи теорії ймовірностей.

Наукова новизна отриманих результатів.

У дисертації представлено наступні наукові результати:

1. Вперше для реальних мовленнєвих сигналів отримано кількісні оцінки ступеню підвищення точності розпізнавання мовлення, спотвореного шумом різної природи та інтенсивності, шляхом навчання системи автоматичного розпізнавання на спотворених шумом сигналах.
2. Вперше для реальних мовленнєвих сигналів отримано кількісні оцінки ступеню підвищення точності розпізнавання мовлення, спотвореного реверберацією, шляхом навчання системи автоматичного розпізнавання на спотворених реверберацією сигналах.

3. Вдосконалено метод оцінювання розбірливості мовлення непрямим методом, із використанням міри якості сигналів у вигляді барківського спектрального спотворення.
4. Уточнено висновки щодо залежності розбірливості мовлення від щільності відбить звуку та часу реверберації, із використанням імовірнісних моделей імпульсних характеристик приміщень.
5. Вдосконалено спосіб виявлення ефекту кліпування мовленнєвих сигналів та об’єктивного оцінювання якості мовленнєвих сигналів, спотворених кліпуванням, що базується на використанні коефіцієнта ексцесу як міри спотворення сигналів.

Особистий внесок здобувача. Усі результати, наведені у дисертаційній роботі і винесені на захист, отримано за активної участі здобувача та опубліковано у спеціалізованих фахових виданнях.

У роботі [1], опублікованій в співавторстві, здобувач особисто взяв участь в записі тестових сигналів в приміщеннях різного об’єму та виконав їх первинну обробку шляхом обчислення імпульсних характеристик приміщень в різних точках цих приміщень. У роботі [2], опублікованій в співавторстві, здобувач особисто розробив алгоритм та комп’ютерну програму імітаційного моделювання ранніх відбить звуку в імпульсній характеристиці приміщення, а також виконав аналіз результатів експериментальних досліджень. У роботі [3], опублікованій в співавторстві, здобувач особисто взяв участь в суб’єктивному оцінюванні якості спотворених мовленнєвих сигналів інтрузивним методом та виконав подальшу статистичну обробку отриманих результатів. У роботі [4], опублікованій в співавторстві, здобувачем особисто виконано комп’ютерне моделювання системи автоматичного розпізнавання мовлення та здійснено статистичну обробку великого масиву даних у вигляді спотворених мовленнєвих сигналів. У роботі [5], опублікованій в співавторстві, здобувачем особисто розроблено комп’терні програми для обробки великого об’єму даних

у вигляді оцінок імпульсних характеристик приміщень різного об’єму, а також виконано аналіз результатів експериментальних досліджень. У роботі [6], опублікованій в співавторстві, здобувачем особисто виконано наступне: взято участь в оформленні патенту на корисну модель.

Практичне значення отриманих результатів. Практичне значення отриманих результатів полягає у встановленні умов досягнення високої точності розпізнавання в системах автоматичного розпізнавання мовленнєвих сигналів, спотворених шумами різної природи та інтенсивності, за наявності різної апріорної інформації щодо відношення сигнал-шум, що дозволяє забезпечити робастність системи автоматичного розпізнавання шляхом використання відносно простих правил її налаштування; встановленні умов досягнення високої точності розпізнавання в системах автоматичного розпізнавання мовленнєвих сигналів, спотворених реверберацією приміщень із різним часом реверберації, за наявності різної апріорної інформації щодо часу реверберації приміщення, що дозволяє забезпечити робастність системи автоматичного розпізнавання шляхом використання певних правил її налаштування; визначенні працездатності та ефективності оцінювання розбірливості мовлення непрямим методом, із використанням міри якості сигналів у вигляді барківського спектрального спотворення, що дозволяє оцінювати розбірливість мовлення, спотвореного реверберацією, за наявності еталонного неспотвореного сигналу; визначенні залежності розбірливості мовлення від щільності ранніх відбить звуку та часу реверберації, із використанням імовірнісних моделей імпульсних характеристик приміщень, що дозволяє обґрунтувати експериментальні результати оцінювання розбірливості мовлення в різних точках приміщення; встановленні можливості автоматизації виявлення кліпування та об’єктивного оцінювання якості мовленнєвих сигналів, спотворених кліпуванням.

Зв'язок роботи з науковими планами, програмами, темами.

Викладені у дисертації нові теоретичні та практичні результати досліджень знайшли застосування у освітньому процесі кафедри акустичних та мультимедійних електронних систем за спеціальністю 171 Електроніка, в освітній програмі "Акустичні електронні системи та технології обробки акустичної інформації" Національного технічного університету України "Київський політехнічний інститут імені Ігоря Сікорського".

Апробація результатів дисертації. Матеріали дисертаційних досліджень обговорювалися на міжнародних конференціях:

1. IEEE 6th International Conference on Methods and Systems of Navigation and Motion Control (MSNMC), 20-23 жовтня 2020, Київ, Україна.
2. IEEE 5th International Conference "Actual Problems of Unmanned Aerial Vehicles Developments (APUAVD)", 22-24 жовтня 2019, Київ, Україна.
3. IEEE 2nd International Conference "Advanced Information and Communication Technologies (AICT)", 4-7 липня 2017, Львів, Україна.
4. IEEE 4th International Conference "Methods and Systems of Navigation and Motion Control (MSNMC)", 18-20 жовтня 2016, Київ, Україна.

Публікації. За результатами досліджень опубліковано 5 наукових публікацій (з них 3 статті в наукових фахових виданнях України, 1 стаття в періодичному науковому виданні інших держав, 1 стаття в періодичному науковому виданні**,** що входить до WoS або Scopus), 1 патент на корисну модель, 4 тези доповідей у збірниках матеріалів конференціїй.

Структура та обсяг дисертаційної роботи. Робота складається зі вступу, чотирьох розділів, списку використаних джерел із 85 найменувань та 5 додатків. Робота містить 55 рисунків та 33 таблиці. Загальний обсяг дисертаційної роботи складає 148 сторінок.

1 ТЕОРЕТИЧНІ ТА ПРАКТИЧНІ ЗАСАДИ РОЗРОБКИ СИСТЕМ АВТОМАТИЧНОГО РОЗПІЗНАВАННЯ МОВИ, РОБАСТНИХ ДО ДІЇ ЗАВАД

Даний розділ присвячено теоретичному огляду сучасних систем автоматичного розпізнавання мови, робастних до дії завад. Наведено літературний огляд способів підвищення стійкості системи АРМ до дії шумової та реверберaційної завади.

1.1 Моделювання систем АРМ

Постановка задачі підвищення робастності системи АРМ полягає в наступному. В процесі експлуатації системи АРМ мовленнєвий сигнал, що надходить на вхід її звукового тракту, може бути спотворений шумовою або реверберaційною (при експлуатації системи АРМ в приміщенні) завадою:

$$y(t) = x(t) \otimes h(t) + n(t), \tag{1.1}$$

де $\otimes$ - символ згортки; *x(t)*- неспотворений мовний сигнал; *h(t)* - імпульсна характеристика (ІХ) приміщення; *n(t)* - випадковий шумовий процес (шум навколишнього оточення). В самому звуковому тракті також можуть виникнути чинники, що знижують точність розпізнавання, зокрема кліпування, а отже і ефективність роботи такої системи. Схематично вплив спотворень на мовний сигнал показано на рис. 1.1 [11]. Задачею при моделюванні є підбір таких програмних та апаратних засобів, що допоможуть мінімізувати згубний ефект вищенаведених чинників.

Рис.1.1. Вплив завад на мовленнєвий сигнал [11]

Основними етапами роботи з системою АРМ є навчання та тестування (рис.1.2). На етапі навчання кожній структурній одиниці ставиться у відповідність набір ознак (features), відбувається формування шаблонів або зразків (patterns) на основі лексикону, акустичної та/або мовної моделей. Лексикон дозволяє записати слова зі словнику у вигляді фонем, тобто власне звуків, які використовуються для промовляння кожного слова. Задачею акустичної моделі є отримати послідовність фонем з аудіо-сигналу. Мовна модель дозволяє отримати з фонем та слів послідовність слів у вигляді речення. На етапі тестування в системі аналізу мови відбувається виокремлення ознак, після чого інформація передається на класифікатор, на якому відбувається порівняння отриманих даних з вже наявними після етапу навчання, об’єднання всіх отриманих даних з лексикону, акустичної та мовної моделей і відбувається прийняття рішення - ми отримуємо результат розпізнавання.

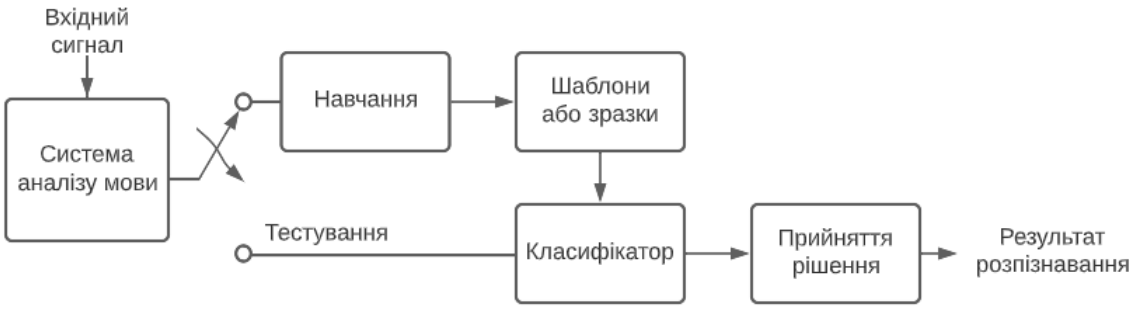


Рис.1.2. Принципова схема загального підходу до моделювання систем АРМ

Для параметризації сигналу, при виділенні ознак, важливо, щоб вони були стійкими до можливої мінливості голосу диктора, впливу завад і спотворень. З огляду на це поширеним є використання MFCC-ознак (Mel-frequency cepstral coefficients - Мел-частотні кепстральні коефіцієнти, ([12], [13]) та PLP-ознак (від Perceptual Linear Prediction - перцепційне лінійне прогнозування, [12], [14], [15]), які є ознаками, що базуються на основі системи сприйняття мовлення людиною [16]. До інших ознак, що можуть використовуватися для параметризації сигналу, відносяться такі, що базуються на приципі продукування мовлення: артикуляційні ознаки та LPC-коефіцієнти (від Linear Predictive Coding - кодування з лінійним предиктором) [12], а також динамічні та просодичні ознаки [16].

В системі АРМ НТК [17], що була використана в даній роботі, параметризація сигналу відбувається з використанням мел-частотних кепстральних коефіцієнтів. Їх формування відбувається наступним чином. Початковий сигнал розбивається на фрейми довжиною 25 мс, в більшості систем вони “перекриваються” для більш плавного переходу від фрейму до фрейму. До цих фреймів застосовується вікно, щоб виключити вплив різкої зміни параметрів сигналу по краях фрейму [12]. Після цього відбувається обчислення кепстру сигналу c , що являє собою перетворення Фур’є FT від логарифма спектру сигналу s:

$$c = FT^{-1}\left\{\lg\left|FT\left\{s\right\}\right|\right\} \tag{1.2}$$

Враховучи те, що людське вухо має різну частотну роздільну здатність в різних діапазонах частот, і оскільки MFCC-коефіцієнти формуються з урахуванням системи сприйняття мовлення людиною, перетворення сигналу в спектр має відбувається також нелінійно. Для цього використовується Мел-частотна шкала, апроксимуюча шкалу частот людського слуху, яка визначається наступним виразом [17], [18]:

$$Mel(f) = 2595\log(1 + \frac{f}{700}) \tag{1.3}$$

Програмно етапи навчання та тестування можуть бути реалізовані на основі прихованих марковських моделей (ПММ), нейронних мереж (НМ), гібридних або тандемних технік та їх комбінацій.

Прихована Марковська модель (ПММ) лежить в основі перших систем АРМ, запропонованих Баумом та ін. в 1960-х роках [19]. В основі підходу лежить наступний принцип. Мовленнєві сигнали представляються у вигляді часового ряду, тобто характеризуються послідовністю вимірювань $x_0, x_1, ...$, де послідовність відображає зміну в часі, а x_t є t-тим вимірюванням в ряду. Мовлення є нестаціонарним процесом, тобто його характеристики змінюються в часі. ПММ є статистичними моделями часового ряду. ПММ моделює часовий ряд як згенерований процесом, який проходить через ряд станів, слідуючих по ланцюгу Маркова. Кожний наступний стан, в який перейде процес з будь-якого теперішнього, визначається стохастично і залежить виключно від теперішнього стану. В кожний момент часу процес створює спостереження з розподілу імовірностей, пов’язане зі станом, в якому знаходиться в цей момент [20].

Математично ПММ описується як імовірнісна функція ланцюгу Маркова і є двічі стохастичною моделлю. Перший рівень моделі - ланцюг Маркова, що описується початковим станом розподілу імовірності π і матрицею переходів **A**. π визначає імовірність знаходження процесу в будь-якому стані в початковий момент часу. Позначивши послідовність станів, яких набуває процес, як $q_0, q_1, ...$, отримаємо математичний опис імовірності того, що в перший момент часу процес буде знаходитись в стані i :

$$\pi(i) = P(q_0 = i) \tag{1.4}$$

A є матрицею, для якої (*i*, *j*)-ті вхідні значення (1.5) описують імовірність, з якою процес перейде в стан *j* зі стану *i*:

$$a_{i,j} = P(q_{t+1} = j \mid q_t = i) \tag{1.5}$$

Таким чином, ланцюг Маркова імовірнісно описує спосіб переходу процесу з одного стану в інший.

Другий рівень моделі є рядом вихідних станів розподілів імовірностей, по одному для кожного стану. Якщо процес набуває стану *i* в момент часу *t*, створюєтся спостереження $\mathbf{x}_t$, яке добувається з розподілу вихідних станів $P_i(\mathbf{x})$. При застосуванні ПММ в системах розпізнавання мовлення розподіли вихідних станів зазвичай моделюються з застосуванням Гауссівської суміші, при цьому $P_i(\mathbf{x})$ описується виразом:

$$P_i(\mathbf{x}) = \sum_{k=1}^{K} \omega_{i,k} N(\mathbf{x}; \boldsymbol{\mu}_{i,k} \boldsymbol{\theta}_{i,k}), \tag{1.6}$$

де $N(\mathbf{x}; \boldsymbol{\mu}_{i,k} \boldsymbol{\theta}_{i,k})$ - багатоваріативна Гауссівська щільність з середнім вектором $\boldsymbol{\mu}$ та коваріантною матрицею $\boldsymbol{\theta}$; $\omega_{i,k}, \boldsymbol{\mu}_{i,k}, \boldsymbol{\theta}_{i,k}$ є ваговим коефіцієнтом суміші, середнім вектором та коваріантною матрицею *k*-того гауссіана в суміші для стану *i*. *K* – кількість гауссіанів в суміші [20].

Описані параметри ПММ використовуються в системі АРМ для обчислення імовірностей слідування певної послідовності станів [20], генерування послідовності спостережень з послідовності станів [20], застосування алгоритмів максимальної правдоподібності, прямого-зворотнього ходу, алгоритму Баума-Велча [21], алгоритму Вітербі [22] та ін. на різних етапах обчислень в залежності від задачі розпізнавання.

Підхід на основі нейронних мереж (НМ) засновано на моделюванні

організації та функціонування мереж нервових клітин людини, які відповідають за сприйняття мови. Будується “модель слухової системи”, в якій відбувається аналіз сигналу і надається спектральна інформація про сигнал, що є доступною для детекторів ознак [18]. На вхід штучної НМ подається послідовність коефіцієнтів, на виході системи ми отримуємо розподіл імовірності для кожного фрейму. Для задачі розпізнавання мови використовуються різні види НМ [23], такі як глибинні НМ (deep neural network, DNN [24], [25], [26], [27]), згорткові НМ (convolutional neural network, CNN [28], [29]), рекурентні НМ (recurrent neural network, RNN [30], [31]) , НМ з часовою затримкою (time-delay neural network TDNN [31], [32]), та ін.

Практичне застосування в індустрії знайшли так звані гібридна та тандемна техніки, за якими в одній системі АРМ використовуються обидва підходи: (і ПММ, і DNN) [33]. Так, в роботі [23] запропоновано двоетапний варіант навчання системи АРМ – використання гібридної моделі. На першому етапі ініціалізуються шари детекторів по одному за раз шляхом підлаштування генеративних моделей, кожна з яких має один шар прихованих змінних. Ці генеративні моделі навчаються без використання інформації про стани ПММ, які будуть розрізнятись акустичною моделлю. На другому етапі кожна модель в стеку використовується для ініціалізації одного шару прихованих одиниць в DNN, а потім вся мережа вибірково налаштовується для прогнозування цільових станів ПММ [23].

Вичерпний опис поєднання гібридної та тандемної технік наведено в [34]. Обидві моделі, - тандемна та гібридна, - мають схожу структуру акустичної моделі на основі ПММ. В тандемній моделі DNN застосовується для виокремлення ознак, які використовуються акустичною моделлю на основі моделі гауссівської суміші (Gaussian Mixture Model, GMM) на етапі *back-end*. В гібридній моделі DNN використовується в якості акустичної моделі та генерує апостеріорні стани ПММ. Акустичні лог-імовірності з тандемної GMM-моделі

та гібридної DNN лінійно пов’язуються як нові акустичні показники, що далі використовуються в алгоритмі декодування Вітербі. [34].

В останні роки активно досліджується відносно нова техніка *End-to-End* (E2E), що показує високу точність розпізнавання при дії завад на мовленнєвий сигнал. [35], [36]. За технікою E2E створюється єдина мережа, за допомогою якої відбувається пряме перетворення мовної послідовності в послідовність вихідних значень (результату розпізнавання). До переваг такого підходу відносять використання єдиної цільової функції, що є зручним при оптимізації системи АРМ, тоді як в гібридних моделях оптимізують компоненти системи окремо, що не завжди може гарантувати глобальний оптимум; також, оскільки відбувається пряме перетворення мовлення в результат розпізнавання, значно спрощеним є процес побудови АРМ, порівняно з традиційними гібридними моделями. Втім, досі є відкритими питання інтегрування лексичних особливостей в таку систему, збільшення обсягу словника та роботи з low-resource мовами, тому з практичних міркувань, таких як передача даних, затримки, можливість адаптації і т.д., для комерційних продуктів досі найчастіше застосовуються гібридні моделі. [35]

1.2 Підвищення робастності систем АРМ до дії завад

До підходів, що підвищують робастність системи АРМ в несприятливих шумових та реверберaційних умовах, відносяться два основні: попередньої корекції та адаптації системи. За методом попередньої корекції обробка мовного сигналу $y(t)$ відбувається в препроцесорі для отримання відновленого сигналу $x(t)$ до потрапляння мовленнєвого сигналу власне в блок виділення структурних одиниць та створення класифікаційних ознак (що, зокрема, досліджувалось в [37], [38], [39]). Метод адаптації системи передбачає зміни “всередині” системи АРМ шляхом адаптації параметрів системи до шумових $n(t)$ та реверберaційних $h(t)$ умов [11]. Оскільки в даній дисертації

досліджується метод адаптації параметрів системи до спотвореної завадою мови, далі буде наведено огляд варіантів саме цього методу.

При роботі з шумовою завадою адаптація системи може бути проведена за двома стратегіями: адаптація моделі та компенсація шуму в самій моделі. У підході на основі адаптації моделі з шумовою завадою працюють неявно шляхом коригування параметрів акустичної моделі мови, тоді як у підході на основі компенсації шуму відбувається точне моделювання шуму та його впливу на ознаки (features) зашумленої мови. Найпоширенішими підходами адаптації моделі є лінійна регресія максимальної правдоподібності (*maximum likelihood linear regression*; див. роботи [39], [40], [41]), максимальна апостеріорна адаптація (*maximum a posteriori adaptation*, [42], [43], [44]) та їх узагальнення [45], [46], [47]. Основою цих підходів є змінення акустичної моделі мовлення з урахуванням додаткових навчальних даних або даних тестування. Підхід адаптації моделі є дещо більш загальними, ніж підхід, заснований на компенсації шуму в моделі, оскільки може впливати на сигнал, який важко явно змоделювати, наприклад, нелінійне спотворення та зміни голосу в реакції на шум (ефект Ломбарда, [48]). Однак за наявності додаткового шуму неврахування взаємодії між мовою та шумом може погіршити продуктивність [20].

У підході на основі компенсації шуму для формування зашумленого мовного сигналу явно моделюються різні фактори, присутні в акустичному середовищі: мова, різні джерела акустичних завад та варіанти їх взаємодії. Моделюючи шум окремо від мовлення, ці моделі можуть продукувати комбінації звуків мови та шуму, які не входили в вибірку під час навчання. Така модель впливу акустичного середовища може бути використана для більш точного визначення послідовностей станів мовної моделі. Спільне виконання цих процесів дозволяє розпізнати різні можливі комбінації мовлення та завад [20].

Одним з перших прикладів застосування методу компенсації шуму є

дослідження, описане в роботі [51], де для навчання системи АРМ до мовних сигналів було додано білий шум. Було проведно два експерименти: зі співпадаючими та неспівпадаючими значеннями SNR навчальної та тестової вибірок, і показано, що навчання системи в першому випадку результує в нижчому показнику WER%. Втім, дослідниками було зазначено, що таке навчання потребує значних (на той час) обчислювальних потужностей. В [20] наводяться результати навчання системи на вибірці, до якої також додано білий шум; показано значне зниження значення WER%, особливо для низького значення відношення сигнал-шум.

При роботі з реверберaційною завадою найпоширенішими методами адаптації системи є техніки back-end. Такі підходи до розпізнавання мови, спотвореної реверберацією, базуються на основі модифікації акустичних моделей і законів декодування розпізнавача. Прямий підхід до отримання акустичної моделі, яка підходить для розпізнавання реверберaційної мови, полягає в навчанні розпізнавача за допомогою реверберaційних даних, записаних у цільовому середовищі. Якщо точні умови тестування невідомі під час навчання, ряд моделей для різних умов реверберації може бути навчений заздалегідь, і для розпізнавання обирається одна з таких попередньо навчених моделей [20], [49]. Оскільки запис мовлення в цільовому середовищі є дуже ресурсозатратним, альтернативно можна виміряти лише ІХ приміщення, а власне навчальну вибірку згенерувати шляхом згортання «чистих» мовних сигналів із виміряними ІХ [50]. Однак, в такому випадку не враховується часова дисперсія ІХ, що може призвести зниження ефективності акустичних моделей. Також, щоб заощадити ресурси на вимірюванні ІХ, можна використовувати штучно створені ІХ, якщо їхні параметри (такі як час реверберації) є наближеними до параметрів цільового середовища [20].

Оскільки з розвитком науки і техніки обчислювальні потужності стають більш доступними, цей метод з часом “еволюціонував” у використання інструментів *Data Augmentation* - засобів для збільшення обсягу бібліотеки

шляхом модифікації існуючих даних або шляхом синтетичного генерування нових зразків даних з використанням початкових зразків. При роботі з аудіо-зразками розширення бібліотеки відбувається шляхом додавання шумової завади або реверберації, зміни висоти тону диктора, швидкості вимови, та ін. В останні роки досить багато уваги приділяється вивченню різних технік для збільшення обсягу бібліотек. Так, в роботі [52] досліджувалась ефективність системи при навчанні на аудіо-сигналах, створених з додаванням змодельованих та/або реальних ІХ, а також синтетичних та реальних шумів (*far-field speech*) зі збільшенням обсягу бібліотеки втричі, при використанні акустичної моделі на основі TDNN. В [53] інструменти WavAugment використано для створення вибірок зокрема зі зміною висоти тону голосу, додаванню шумів, реверберaційної завади, пропусканням зразків мовлення через режекторний фільтр, які зрештою застосовувались для дослідження ефективності моделі на основі CPC (*Contrasive Predictive Coding* - контрастивне передбачення кодування). Також вплив засобу WavAugment на зниження показника WER% досліджено в [54] і показано, що такий засіб вирішує проблему перенавчання. Разом з дослідженням впливу зміни умов навчання на ефективність роботи системи АРМ, також проводяться порівняння різних програмних інструментів та їх комбінацій для збільшення обсягу бібліотек: в статті [55] - наводится порівняння інструментів SpecAugment та MixSpeech; в [56] - техніку додавання випадкового шуму поєднано з методом виключення, або дропауту (від англ. dropout).

Ще одним варіантом збільшення обсягу даних є штучна зміна швидкості мовлення без зміни висотності та огинаючої сигналу. В [32] запропоновано використовувати разом з оригінальними зразками також вповільнені та прискорені зразки (0,9 та 1,1 початкової швидкості, відповідно), акустичну модель було побудовано на основі DNN. Також, в [57] досліджувались можливості штучної зміни тону диктора (*speaker augmentation*) для збільшення обсягу бібліотеки та підвищення ефективності роботи системи АРМ.

Особливої актуальності підхід data augmentation набуває при роботі з так званими low-resource мовами - мовами, бази даних яких значно обмежені по кількості годин, що показано в роботах [57] та [58]. Ще одним варіантом підвищення точності розпізнавання при роботі з low-resource мовами може бути згадане вище поєднання систем АРМ різної побудови: різні системи АРМ можуть мати на вході одні і ті самі ознаки, але мати значні відмінності в принципах побудови (напр. акустичні та мовні моделі), що надає додаткові переваги, що показано в [34], де такі системи застосовували для підвищення ефективності системи при вирішенні задачі пошуку ключових слів.

Втім, результати згаданих вище робіт ілюструють загальну ефективність того чи іншого підходу в комбінації з певною архітектурою акустичної (та/або мовної) моделі або обсягу бібліотеки, але не показують деталізовано ефективність систем при конкретних шумових та реверберaційних умовах, в яких може використовуватись система АРМ. Беручи до уваги показану в [51], [20] тенденцію значного підвищення точності розпізнавання при навчанні системи на вибірках, зашумлених білим шумом, та підвищення ефективності системи АРМ при використанні методу адаптації системи для шумової та реверберaційної завад, в даній дисертаційній роботі за мету поставлено дослідити вплив різних варіантів навчання системи на точність розпізнавання при використанні різних комбінацій навчальних та реальних експлуатаційних умов. Даний напрямок дослідження може бути перспективним, оскільки дозволить зробити висновок про доцільність застосування того чи іншого методу з точки зору підбору оптимального обсягу бібліотек та стилю навчання при наявній апріорній інформації про умови експлуатації систем АРМ, що потенційно результуватиме у зниженні ресурсів, необхідних для побудови ефективної системи АРМ.

1.3 Міри оцінювання якості та розбірливості мовлення

При моделюванні систем АРМ та пошуку шляхів підвищення їх стійкості до дії шумової та реверберaційної завад важливо приділити увагу мірам оцінювання якості та розбірливості мовлення. Система АРМ під час розпізнавання буде підпадати під вплив реверберації приміщення, шумових завад та певних спотворень, спричинених спектральними характеристиками приміщення, так само як і людина під час слухання лектора. Перелічені фактори можуть суттєво знизити якість розпізнавання та сприйняття інформації.

Стосовно сприйняття інформації людиною, особливого дискомфорту будуть зазнавати студенти з порушеннями слуху [59] та учні початкової школи [60]. Незважаючи на значний розвиток даної сфери, в [61], [62] показано, що є ряд проблем, що потребують ретельного дослідження. Наприклад, на противагу концертним та конференц-залам або шкільним аудиторіям, для яких проведено багато досліджень та замірів [63], [64], аудиторії університетів були позбавлені такої уваги. Це можна пояснити тим, що університетські приміщення за розмірами займають проміжне становище між малими шкільними аудиторіями та великими концертними майданчиками [65]. Втім, розмір університетських аудиторій теж може значно варіюватися. За класифікацію, запропонованою в [66], аудиторії поділяють на малі (до 350 $м^3$), середні (350-650 $м^3$) та великі (більше 650 $м^3$), а за іншою класифікацією, наведеною в [67] малі, середні та великі аудиторії мають об’єм до 230 $м^3$, 230-350 $м^3$ та більше 350 $м^3$ відповідно [1].

Об’єм приміщення суттєво впливає на його акустичні властивості, що наявно проявляється у зміні форми та параметрів імпульсної характеристики приміщення [68]. Відомо, наприклад, що збільшення розмірів приміщення призводить, як правило, до збільшення часу реверберації, зменшення середньої щільності відбиттів звуку, зростання граничного моменту часу, що відокремлює ранні відбиття від пізньої реверберації [1], [67], [68].

Зрештою, об’єм приміщення впливає на такі показники сприйняття мовленнєвих сигналів як якість та розбірливість мовлення. Зазвичай ці показники пов’язують із часом реверберації T60 [63], що можна пояснити руйнівною дією пізньої реверберації, яка посилюється із зростанням часу реверберації. Почасти популярність часу реверберації як міри розбірливості мовлення можна пояснити відносною простотою вимірювання часу реверберації. Оскільки забезпечення якісного та розбірливого мовлення є актуальним для школярів молодших класів, школярів із вадами слуху та людей похилого віку, цілком зрозумілим є інтерес дослідників до пошуку шляхів оптимізації часу реверберації [1], [69], [70].

Кореляція між значеннями параметрів T30, EDT, з однієї сторони, та значеннями STI, з іншої сторони, в 5-6 локаціях університетських приміщень малого, середнього та великого розміру, досліджувалася в [59]. Було показано, що EDT, на відміну від T30, може виконувати роль міри розбірливості мови в різних локаціях певного приміщення. Разом із тим, було виявлено [59], що використання EDT як міри розбірливості мовлення може бути проблематичним в малих аудиторіях через недостатньо високий рівень кореляції (приблизно 0,5) значень EDT із значеннями STI в різних локаціях приміщення. Можливі причини виявленого явища в [59] не були вказані, хоча доречно було би припустити, що певну роль тут відіграють ранні відбиття звуку [1].

Взагалі, переважно вважається, що дія ранніх відбиттів звуку є корисною, оскільки забезпечує високу розбірливість мовлення [71], [72]. В [72] навіть зроблено висновок про еквівалентність впливу ранніх відбиттів та прямого звуку на розбірливість мовлення. Проте, наведені у [62] результати досліджень не узгоджуються із цим висновком й свідчать, що ранні відбиття відіграють менш важливу роль, порівняно із прямим звуком, в забезпеченні високої розбірливості мовлення. Зазначені результати не були ґрунтовно пояснені в [62], хоча насправді пояснення можуть бути не надто складними.

Наприклад, якщо врахувати, що утворення ранніх відбиттів звуку є подібним до утворення сигналу на виході гребінчастого фільтру, якість сигналу, створеного за участю ранніх відбиттів, має бути нижчою за якість прямого сигналу через нерівномірність частотного відгуку віртуального гребінчастого фільтру. Таким чином, наявність ранніх відбиттів звуку може бути причиною неузгодженості між розбірливістю мовлення та якістю мовлення, коли розбірливість мовлення підвищується, а якість мовлення спадає. Проте умови, за яких має місце така неузгодженість, або за яких воно порушується, досі залишаються недослідженими [1].

В роботах [73], [74] було проведено оцінювання показників розбірливості C_{50} та STI для середніх та великих аудиторій, проте не були розглянуті малі приміщення. Натомість в роботі [75] було обчислено показники розбірливості мовлення в малих та середніх аудиторіях артикуляційним та інструментальним методами, проте не було проаналізовано аудиторії великого об’єму; були обчислені показники EDT, T_{30} та C_{50}, але обчислення показника STI проведено не було [1].

В роботах [76], [77] було проведено порівняння показника STI в середині приміщення та біля дальньої стіни та показано, цей показник зростає у дальньої стіни приміщення на 7-14%, проте не було проаналізовано зв’язок між оцінками розбірливості мови розміром приміщення, а також зв’язок між показником STI та та об’єктивними мірами оцінювання якості мови. Частково ці прогалини було закрито в [65], однак питання зв’язку між показником STI та об’єктивними мірами оцінювання якості мови залишається відкритим, а отже неясним залишається зв’язок між якістю та розбірливістю мовлення в аудиторіях різного розміру [1].

Розбірливість мовлення тісно пов’язана із змістом повідомлення й тому може слугувати мірою кількості інформації, сприйнятої слухачем. Якість мовлення відображає емоційну реакцію людини на спотворення форми мовленнєвого сигналу безвідносно до його змісту [1].

Як правило, якісне мовлення є одночасно розбірливим. Прикладом винятку із цього правила може бути ситуація, коли якість мовлення оцінюється інтрузивним методом, тобто шляхом порівняння із еталонним сигналом, й при цьому еталонний сигнал з певних причин є суттєво спотвореним. Наприклад, при випробуваннях лінії зв'язку вимова диктора може бути поганою з точки зору слухача, тому суб'єктивна оцінка якості мовлення за шкалою DMOS буде низькою. Слід очікувати, що при цьому низькою буде також суб'єктивна оцінка розбірливості мовлення, виконана артикуляційним методом. Проте об'єктивне оцінювання якості мовлення інтрузивним методом призведе до високої оцінки за шкалою DMOS, якщо в лінії зв'язку відсутні суттєві спотворення сигналу. Інший приклад якісного, але нерозбірливого, мовлення наведено в [78] [1].

Разом із тим, розбірливе мовлення не обов'язково є якісним. Одним із прикладів є вокодерні лінії зв'язку, де значне спотворення форми сигналу визнається допустимим. Іншим прикладом є використання високочастотних фільтрів в системах автоматичного розпізнавання мовлення (АРМ) та в слухових апаратах [17]. Такі фільтри, посилюючи високочастотні форманти в мовленнєвих сигналах, спотворюють форму мовленнєвого сигналу, але саме завдяки такому спотворенню підвищується розбірливість мовлення на виході системи АРМ. Ще одним прикладом підвищення розбірливості мовлення завдяки підвищенню потужності високочастотних компонентів сигналу є кліпування, тобто, нелінійне спотворення сигналів, яке застосовується в слухових апаратах та кохлеарних системах [1].

1.4 Висновки до розділу 1

1. В даному розділі наведено огляд найбільш поширених алгоритмів моделювання систем АРМ: алгоритм на основі прихованих марковських моделей, алгоритм на основі нейронной мережі, та застосування їх комбінацій. Описано спосіб виокремлення класифікаційних ознак.

2. Описано методи підвищення робастності систем АРМ до дії завад. Показано, що ефективними є метод попередньої корекції сигналу та метод адаптації системи. Наведено огляд методів адаптації системи для шумової завади. Розглянуто методи адаптації системи для реверберацiїної завади, до яких відносяться навчання системи на реверберованих сигналах та використання моделі, що може бути адаптована до реверберацiйних умов приміщення.

3. Наведено огляд розвитку і використання методу підвищення ефективності системи АРМ – Data Augmentation, за яким відбувається збільшення обсягу бібліотеки для досягнення вищої точності розпізнавання.

4. Наведено огляд мір якості та розбірливості мовлення, використання яких може бути доцільним при моделюванні систем АРМ для підвищення її ефективності.

2 ВПЛИВ ЛІНІЙНИХ ТА НЕЛІНІЙНИХ СПОТВОРЕНЬ СИГНАЛУ НА ЯКІСТЬ ТА РОЗБІРЛИВІСТЬ МОВЛЕННЯ

В даному розділі наведено результати дослідження певних факторів, що можуть істотно впливати на ефективність роботи системи АРМ, а саме: реверберації та кліпування сигналу [1], [2], [3], [5], [6], [7]. При цьому розглянуто міри визначення величини кліпування, доцільність використання об'єктивних показників розбірливості мовлення та запропоновано способи моделювання реверберації в приміщенні, що є корисним при створенні системи АРМ, стійкої до дії завад.

2.1 Коефіцієнт ексцесу як міра якості сигналу, спотвореного кліпуванням

2.1.1 Використання коефіцієнту ексцесу для виявлення та оцінки ступеня кліпування

Одним з видів нелінійних спотворень сигналу є кліпування. Кліпування - це спотворення сигналу, при якому високі миттєві значення сигналу замінюються певною пороговою константою або значеннями, близькими до значення цієї константи (жорстке та м'яке кліпування відповідно). Аналітично жорстке кліпування можна описати виразом:

$$y(n) = \begin{cases} x(n), |x| < A, \\ A \cdot \mathrm{sgn}\left[x(n)\right], |x| \geq A, \end{cases} \tag{2.1}$$

де n - номер вибірки сигналу, A - поріг кліпування ($0 < A < C = \max|x(n)|$) [3], [7].

М'яке кліпування може виконуватися за різними алгоритмами. Один із варіантів такого алгоритму описується наступним аналітичним виразом:

$$y(n) = \begin{cases} x(n), |x| < B, \\ a(n) + b \cdot x(n), |x| \geq B, \end{cases} \tag{2.2}$$

де B - значення додаткового порогу, $0 < B < A$ [3], [7].

В наявній літературі в якості міри кліпування запропоновано використовувати коефіцієнт кліпування [79]:

$$R_{cl} = \frac{2\max(D_l, D_r)}{D}, \tag{2.3}$$

де D - різниця між максимальним та мінімальним значеннями сигналу. D_l та D_r - відстані між піковими значеннями оцінки функції щільності імовірності (PDF, Probaility Density Function), як показано на рис. 2.1 (а, б). Недоліком використання такої міри є те, що необхідно попередньо оцінювати функцію розподілу сигналу, що потребує додаткових обчислень [3], [7].

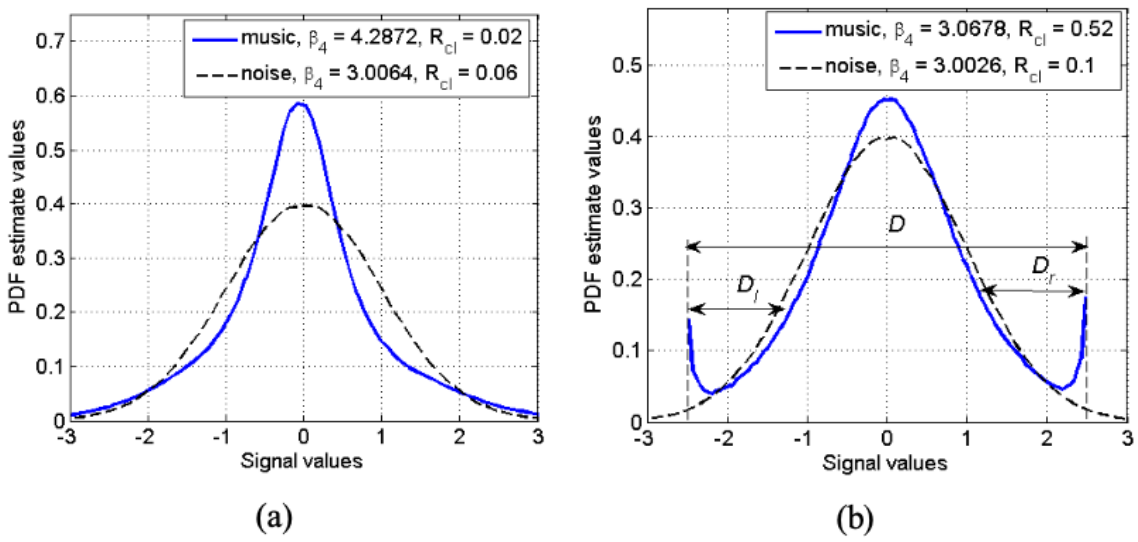


Рис. 2.1 Оцінки PDF, β_4 та R_{cl} для сигналів, не спотворених (а) та спотворених (б) кліпуванням [3]

В роботах [3], [7], [10] запропоновано використовувати коефіцієнт ексцесу - нормалізований центральний момент 4-го порядку:

$$\beta_4 = \frac{\mu_4}{(\mu_2)^2}, \tag{2.4}$$

де μ_k - центральний момент *k*-го порядку [80].

Порівнюючи (2.3) та (2.4), бачимо наочну перевагу використання коефіцієнту ексцесу - для визначення потрібно знати тільки значення сигналу, отже він є набагато простішим в розрахунках. В результаті, виключається проблема вибору типу оцінки функції щільності імовірності, а також і пов'язаним з нею питанням оптимізації параметрів такої оцінки [3], [7].

В той же час, параметр R_{cl} є скінченим в інтервалі можливих значень [0; 1], що є зручним для інженерного застосування, тоді як β_4 не є скінченим. Можна показати, що коефіцієнт ексцесу β_4 не може прийняти значення менше за одиницю, а значення $\beta_4 = 1$ відповідає ситуації кліпування. Дійсно, використавши:

$$Z = \left(X - E\{X\}\right) / \left(V\{X\}\right)^{\frac{1}{2}} \tag{2.5}$$

де $E\{\cdot\}$ - знак математичного очікування; *X* та *Z* - випадкові змінні; $V\{\cdot\}$ - знак дисперсії, отримаємо:

$$\beta_4 = V\left\{Z^2\right\} + 1 \tag{2.6}$$

Отже, коефіцієнт β_4 може набувати значень з діапазону $[1;\infty]$. Для реальних сигналів, що не потрапили під спотворення кліпуванням, значення

β_4 коливаються в межах 30-50 для музичних сигналів [81] та 7-12 для мовлення [80], що є не дуже зручним на практиці [3], [7].

Для інженерного застосування зручніше буде використати коефіцієнти (2.7) та (2.8):

$$\gamma_4 = \frac{1}{\beta_4} = \frac{(\mu_2)^2}{\mu_4} \tag{2.7}$$

$$\eta_4 = \frac{1}{\sqrt{\beta_4}} = \frac{\mu_2}{\sqrt{\mu_4}} \tag{2.8}$$

оскільки їх можливі значення лежать в діапазоні [0;1], причому ситуація $\gamma_4 = 0$ та $\eta_4 = 0$ відповідає ситуації відсутності кліпування. Зазначимо, що η_4 є нормованою дисперсією сигналу, нормалізація відбувається шляхом ділення на квадратний корінь з центрального моменту четвертого порядку. Коефіцієнт γ_4 можна представити як квадрат нормованої дисперсії сигналу, що аналізується. І хоча в [82] було запропоновано використовувати дисперсію сигналу для визначення міри кліпування, проте відсутність нормування робить цю пропозицію непридатною для практичного використання [3], [7].

Нижче наведено результати експериментальної перевірки припущення щодо доцільності використання параметрів η_4 та γ_4 в якості об’єктивних мір кліпування мовного сигналу. Оскільки переметри η_4 та γ_4 можна використовувати і як міру якості мовлення, спотвореного кліпуванням, було створено карти відповідності об’єктивних та суб’єктивних оцінок якості таких сигналів. Практичне значення таких карт відповідності полягає у можливості калібрування об’єктивних мір γ_4 та η_4 [3], [7].

2.1.2 Експериментальна перевірка доцільності використання коефіцієнту ексцесу як міри визначення ступеня кліпування

В даному експерименті використано фрагменти мовних сигналів 8 дикторів (4 чоловічі та 4 жіночі голоси), кожен тривалістю 15 секунд. Всі сигнали були записані в заглушеній кімнаті. Частота дискретизації - 22050 Гц, глибина квантування - 16 біт. Для встановлення величини кліпування використано параметр k [80]:

$$k = 20\lg\left(\max\left|x(n)\right| / A\right) \tag{2.9}$$

Було проведено суб'єктивне та об'єктивне оцінювання якості мовних сигналів, спотворених кліпуванням. Суб'єктивне оцінювання проведено з використанням шкали Degradation Mean Opinion Score (DMOS) [83], [80]. В експерименті взяли участь шість слухачів віком від 19 до 21 року без будь-яких порушень слуху. Об'єктивне оцінювання було проведено в середовищі MatLab з використанням мір (2.4), (2.7), (2.8). Обчислення коефіцієнту β_4 було проведено за наступними аналітичними виразами (2.10), (2.11), (2.12):

$$\bar{\beta}_4 = \frac{(N-1)\left((N+1)\bar{\beta}_4' - 3(N-1)\right)}{(N-2)(N-3)} + 3 \tag{2.10}$$

$$\bar{\beta}'_4 = \frac{\frac{1}{N}\sum_{n=1}^{N}\left(y(n) - \bar{m}_y\right)^4}{\left(\frac{1}{N}\sum_{n=1}^{N}\left(y(n) - \bar{m}_y\right)^2\right)^2} \tag{2.11}$$

$$\bar{m}_y = \frac{1}{N}\sum_{n=1}^{N} y(n) \tag{2.12}$$

Обчислення коефіцієнту R_{cl} було проведено згідно алгоритму, наведеного в [79]. Для цього було використано гістограмне представлення функції щільності імовірності зі 100 бінами (стовпчиками) гістограми [3], [7].

Результати суб'єктивного оцінювання якості мовлення, спотвореного кліпуванням (залежність DMOS(k)), наведено на рис. 2.2 [84]. Показано, що якість мовного сигналу є відносно високою (DMOS≥ 4) за умови $0 \leq k \leq 8$ dB.

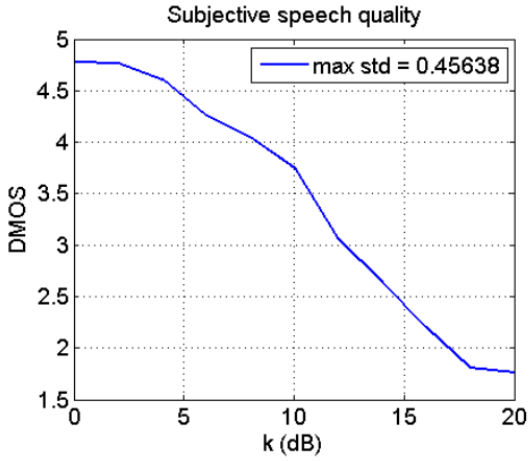


Рис. 2.2. Залежність DMOS(k)) [84]

Результати об'єктивного оцінювання за допомогою мір $\beta_4(k)$, $\gamma_4(k)$, $\eta_4(k)$ та $R_{cl}(k)$ наведено на рис. 2.3 [3], [7]. Можна побачити, що залежності $\beta_4(k)$, $\gamma_4(k)$, $\eta_4(k)$ змінюються повільно та монотонно в діапазоні $0 \leq k \leq 8$ dB. Така тенденція справедлива і для усереднених результатів, і окремо для кожного з сигналів, що аналізуються. В інтервалі значень $8 \leq k \leq 20$ dB швидкість зміни $\beta_4(k)$, $\gamma_4(k)$, η_4 (k) зростає, проте зберігає монотонність [3].

Поведінка міри $R_{cl}(k)$ відрізняється. На інтервалі $0 \leq k \leq 3$ dB значення $R_{cl}(k)$ зростає дуже повільно, потім стрімко збільшується в інтервалі $3 < k < 9$ dB і надалі знов повільно зростає на інтервалі $9 \leq k \leq 20$ dB. Також бачимо, що в діапазоні $0 \leq k \leq 8$ dB залежність не є монотонною, а залежність $R_{cl}(k)$ починає стрімко зростати на широкому інтервалі значень від $0 \leq k \leq 2$ dB до $0 \leq k \leq 6$ dB [3].

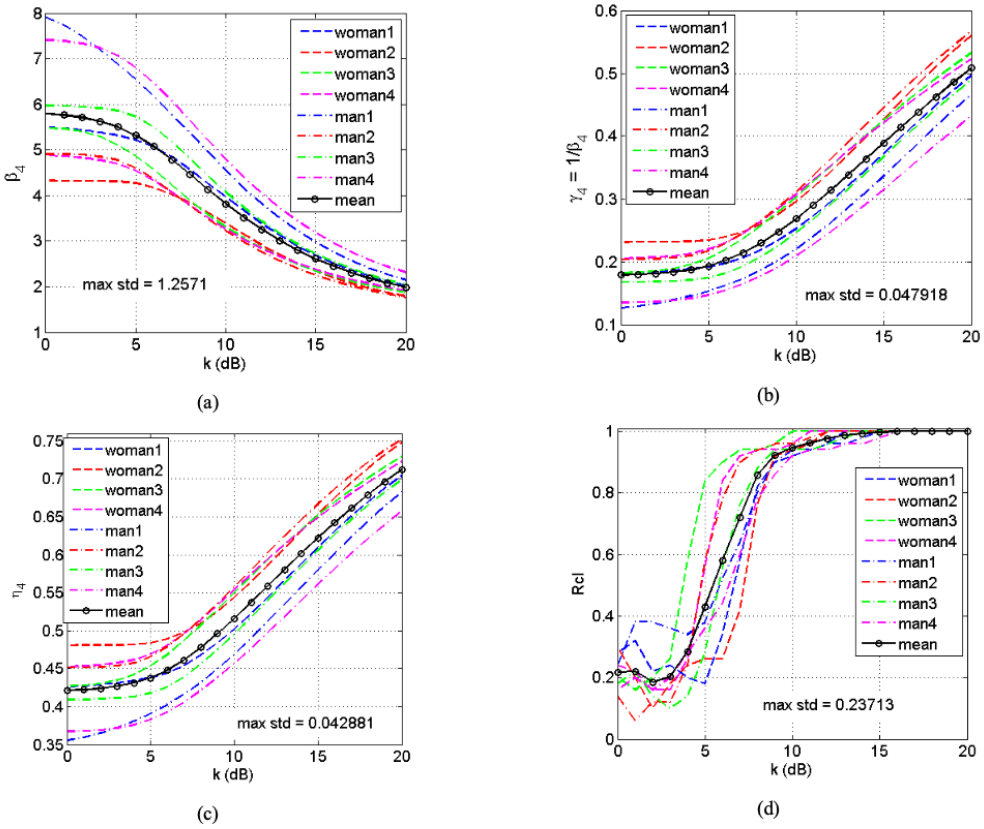


Рис. 2.3. Оцінки залежностей β_4 (k) (a), γ_4(k) (b), η_4(k) (c) та R_{cl}(k) (d) [3]

Переваги використання коефіцієнту ексцесу $\beta_4(k)$ та похідних від нього мір $\gamma_4(k)$, $\eta_4(k)$ також стали очевидними при побудові карт відповідності DMOS(β_4), DMOS(γ_4), DMOS(η_4) та DMOS(R_{cl}). Вони наведені на рис.2.4. Для побудови карт використано усереднені дані суб'єктивного оцінювання. В таблиці 2.1 наведено значення коефіцієнтів кореляції між DMOS та $\beta_4(k)$, $\gamma_4(k)$, $\eta_4(k)$ та $R_{cl}(k)$ [3].

Таблиця 2.1. Коефіцієнти кореляції [3]

Об'єктивна міра	β_4	γ_4	η_4	R_{cl}
Коефіцієнт кореляції	0,985	-0,987	-0,994	-0,885

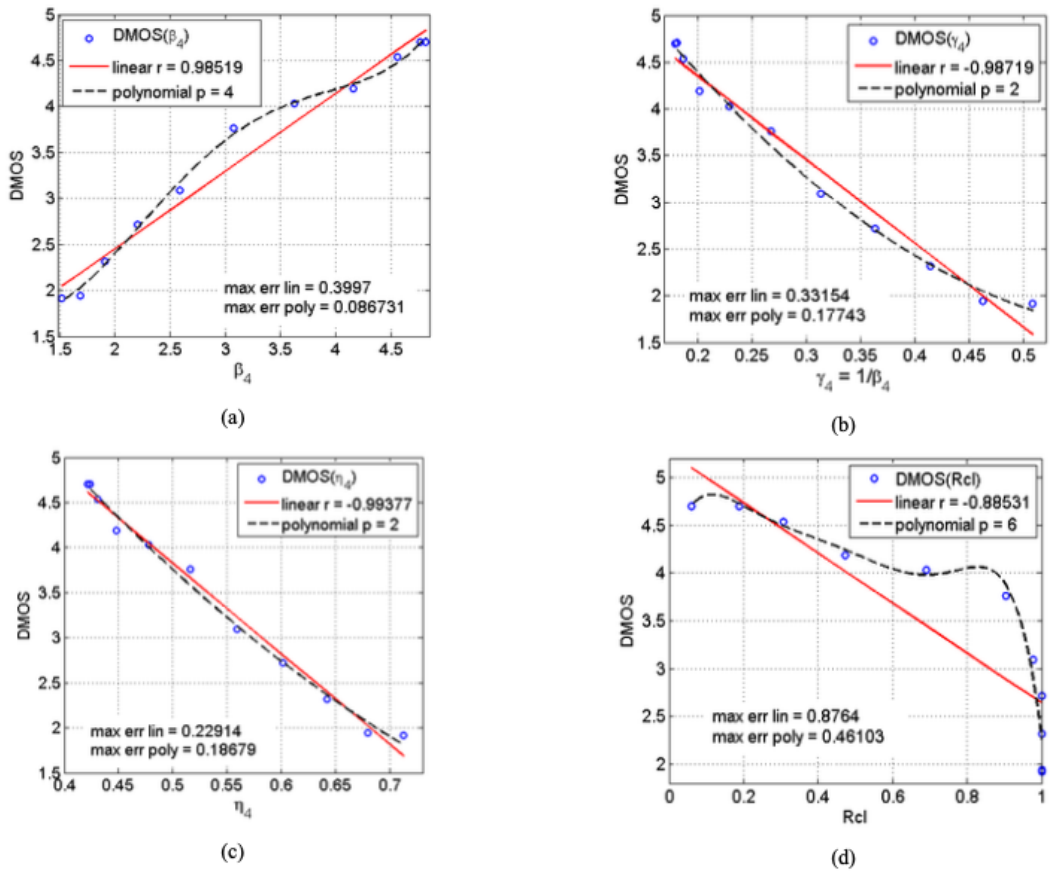


Рис. 2.4. Карти відповідності DMOS(β_4) (a), DMOS(γ_4) (b), DMOS(η_4) (c) та DMOS(R_{cl}) (d) [3]

Значення коефіцієнтів апроксимації поліномів:

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \ldots \qquad (2.13)$$

для DMOS(β_4), DMOS(γ_4), DMOS(η_4) та DMOS(R_{cl}) наведено в таблиці 2.2.

Можна побачити, що використання поліному 4-го порядку достатньо для досягнення малої похибки апроксимації для залежності DMOS(β_4), для залежностей DMOS(γ_4), DMOS(η_4) достатньо використання поліномів другого та першого порядку відповідно. В той час як для залежності DMOS(R_{cl}) малу похибку апроксимації можна отримати лише з використанням поліному 6-го порядку, при чому в даному випадку

апроксимацію не можна вважати задовільною через немонотонність функції на карті відповідності [3].

Таблиця 2.2. Коефіцієнти апроксимації поліномів [3]

Поліноміальний коефіцієнт	β_4 $p = 4$	γ_4 $p = 2$	η_4 $p = 1$	R_{cl} $p = 5$
a_0	5,72604	7,52257	8,85577	4,0868
a_1	-7,88317	-18,53578	-10,06244	16,62126
a_2	5,12563	14,48132	-	-132,09174
a_3	-1,20305	-	-	460,86927
a_4	0,09757	-	-	-834,49338
a_5	-	-	-	752,48433
a_6	-	-	-	-265,21887

Суттєво знизити похибку апроксимації для залежності DMOS(R_{cl}) можна, використовуючи кусково-лінійну апроксимацію (рис.2.5):

$$\mathrm{DMOS} = \begin{cases} 4.95 - 1.17 \cdot R_{cl}, \ 0 \le R_{cl} \le 0.9, \\ 13.6 - 10.72 \cdot R_{cl}, \ 0.9 \le R_{cl} \le 1. \end{cases} \tag{2.14}$$

Відношення (2.14) та рис 2.5 добре ілюструють висновки [79] стосовно низької застосовності міри R_{cl} як міри кліпування. Дійсно, значення R_{cl} належать до інтервалу (0,05; 0,85) та покривають 80% діапазону можливих значень [0; 1] для $\mathrm{DMOS} \ge 4$, тобто для високої якості кліпованого сигналу. Також бачимо стрімке падіння якості сигналу з 4 балів до 3,2 за шкалою DMOS, тобто від оцінки “гарно” до “задовільно”, на малому інтервалі значень R_{cl} - (0,85; 0,95). Тож, можемо зробити висновок, що шкала значень R_{cl} використовується не досить ефективно [3].

Залежності DMOS(β_4), DMOS(γ_4), DMOS(η_4) є гладкими кривими, це показує, що коефіцієнт ексцесу β_4, обернений коефіцієнт ексцесу γ_4, а також

квадратний корінь оберненого коефіцієнту ексцесу η_4 можуть бути застосованими в якості міри кліпування. Додатковою перевагою використання η_4 та γ_4 є те, що їх залежності DMOS(β_4), DMOS(η_4) є майже лінійними, що значно спрощує перерахунок значень (β_4) та (η_4) в значення шкали DMOS [3].

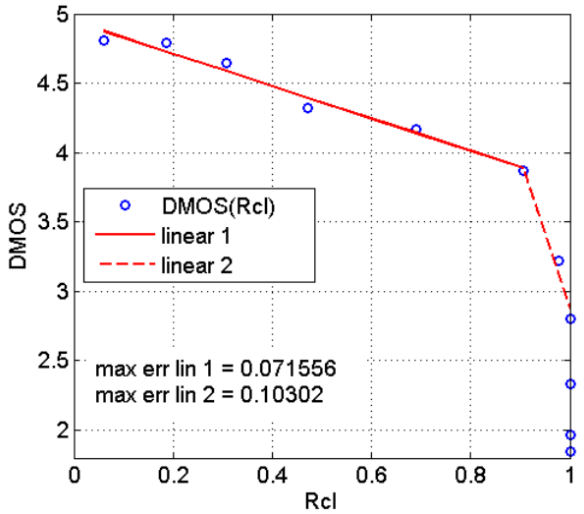


Рис.2.5. Кусково-лінійна апроксимація карти відповідності DMOS(R_{cl}) [3]

Як зазначено вище, коефіцієнт η_4 можна розглядати як нормовану дисперсію процесу, що аналізується, а нормалізація проводиться шляхом ділення на квадратний корінь центрального моменту четвертого порядку. Така інтерпретація коефіцієнту η_4 як міри кліпування співпадає з ідеєю, представленою в [82], де пропонувалося використовувати дисперсію процесу як міру кліпування, хоча відсутність нормування дисперсії принципово перешкоджало практичному втіленню цієї ідеї [3].

2.2 Порівняння мір оцінювання якості та розбірливості мовлення в аудиторіях університету

При моделюванні систем АРМ та пошуку шляхів підвищення їх стійкості до дії шумової та реверберaційної завад важливо приділити увагу мірам оцінювання якості та розбірливості мовлення. Система АРМ під час

розпізнавання буде підпадати під вплив реверберації приміщення, шумових завад та певних спотворень, спричинених спектральними характеристиками приміщення, так само як і людина під час слухання лектора. Перелічені фактори можуть суттєво знизити якість розпізнавання та сприйняття інформації. Отже, важливою задачею є визначення статистичних залежності між об’єктивними мірами оцінки якості мовлення та мірою оцінювання розбірливості мовлення.

2.2.1 Постановка та проведення експерименту для порівняння мір LSD, BSD, PESQ, SSNR та FWSNR та визначення ступеню кореляції з STI

Експериментальні дослідження зв’язку між якістю та розбірливістю мовлення проводилися у три етапи:

1. Для кожної локації кожного із приміщень оцінювалися міри якості мовлення, такі як сегментне відношення сигнал-шум (SSNR), логарифмічне спектральне спотворення (LSD), барківське спектральне спотворення (BSD), frequency-weighted segmental signal-to-noise ratio (FWSNR) та перцептуальна оцінка якості мовлення (PESQ);
2. Для кожної локації кожного із приміщень оцінювалися міри розбірливості мовлення speech transmission index (STI);
3. Для кожного із приміщень обчислювалися коефіцієнти кореляції Пірсона між масивами оцінок SSNR, LSD, BSD, FWSNR та PESQ, з однієї сторони, та STI, з іншої сторони, що відповідали множинам локацій кожної аудиторії [1].

Мовленнєві сигнали, що використовувались для проведення даного експерименту, це записи 8 дикторів (4 жіночих та 4 чоловічих голоси), які зачитували один і той же текст юридичного змісту. Довжина відрізків сигналів, що аналізувалися, становила 15 с, частота дискретизації записів

становила 22050 Гц, бітова глибина становила 16 біт. Зазначимо, що така довжина відрізків сигналів була обрана із урахуванням того, що для суб'єктивного оцінювання якості сигналів за шкалою DMOS достатньо мати відрізок сигналу довжиною 10-15 с [1].

Додатково розглянуто випадки іншого поєднання значень кількості дикторів N та довжини L відрізків сигналів:

- N = 8, L = 60 с;
- N = 12 (6 жінок та 6 чоловіків), L = 15 с.

Мета розгляду таких сполучень полягає в оцінюванні чутливості результатів досліджень до зміни умов досліджень [1].

Для проведення експерименту було використано бінауральні імпульсні характеристики приміщення (БІХ, binaural room impulse responses - BRIRs) малого (177 $м^3$), середнього (270 $м^3$) та великого (370 $м^3$) об'єму [1].

Схеми розташування вимірювальної апаратури в аудиторії №1 малого розміру та аудиторії №2 середнього розміру та аудиторії №3 великого розміру наведені на рис. № 2.6 a, b, c відповідно. Аудиторії №1 та №2 розташовані в навчальному корпусі №12 Національного Технічного університету України «Київський Політехнічний Інститут ім. Ігоря Сікорського». Для вимірювання БІХ цих приміщень використовувалась штучна голова власного виробництва та наступне звукове обладнання: зовнішній аудіо-інтерфейс Steinberg UR242, всенаправлені конденсаторні вимірювальні мікрофони Superlux ECM-999 та активний гучномовець Genius SP-HF 2.0 500. Бінауральні імпульсні характеристики аудиторії №3 (великого розміру) були запозичені з відкритої бібліотеки бінауральних імпульних характеристик бази даних Рейнсько-Вестфальського технічного університету Аахена [85]. Алгоритм оцінювання цих БІХ описано в [86] [1].

Згідно з ISO 3382-1 [63], джерело звуку було розміщено на висоті 1,5 м, а мікрофони - на висоті 1,2 м при вимірюванні на всіх площинах [1].

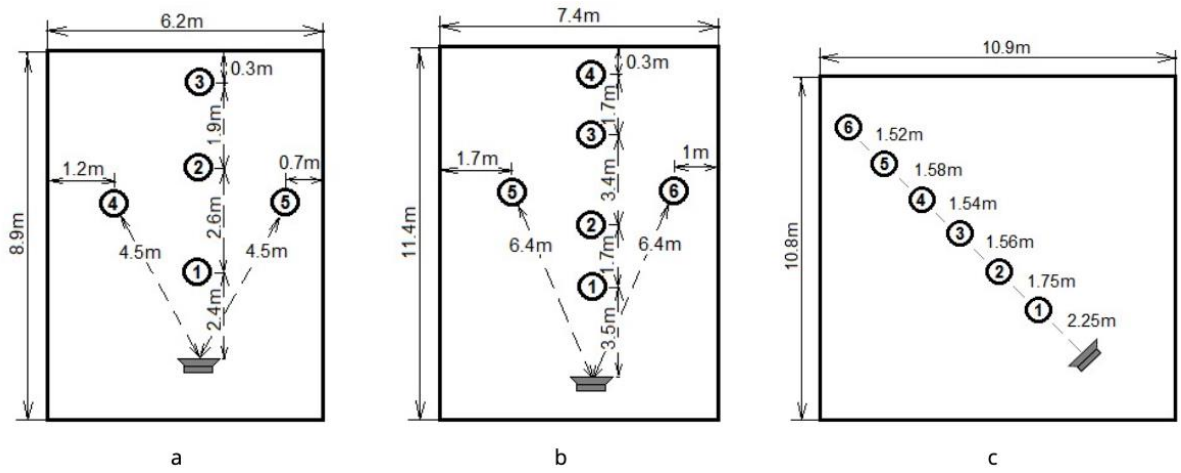


Рис.2.6. Схеми розташування вимірювального обладнання
в приміщеннях №№ 1 (a), 2 (b), 3 (c). [1]

Основою тестового сигналу слугувала mls-послідовність, що відповідає тривалості сигналу 1,49 с при частоті дискретизації 44100 Гц та 1,36 с при частоті дискретизації 48000 Гц. При випромінюванні дана mls-послідовність повторювалась 17 разів. При обчисленні БІХ приміщення 16 останніх сплесків усереднювались для підвищення відношення сигнал-шум на 12 дБ [1].

Запис тестових сигналів було проведено з частотою дискретизації 44100 Гц для приміщень №1 та №2 та 48000 для приміщення №3. Глибина квантування для всіх трьох аудиторій становить 24 біт. Під час запису сигналів мікрофони було розташовано в зонах вушних раковин штучної голови на відстані 1 см від вушного каналу [1].

Місця розташування штучної голови під час запису тестових сигналів наведено на рис.№2.6 пронумерованими кружечками. Детальні характеристики приміщень наведено в таблиці №2.3. Слід зазначити, що є певні відмінності в аудиторії №3, такі як діагональне розміщення столів в приміщенні, невелика кількість місць для сидіння та відсутність студентів під час проведення вимірювання БІХ [1].

При вимірюваннях БІХ в аудиторіях №1 та №2 тестовий сигнал *x(t)* випромінювали за допомогою гучномовця, розташованого на місці викладача. Відгук приміщення *y(t)* на стимул *x(t)* було записано за допомогою пари

мікрофонів, прикріплених до штучної голови, яку розміщували в різних точках приміщення (рис.2.6) [1].

Таблиця 2.3. Характеристики аудиторій [1]

Аудиторія	№1	№2	№3
Об’єм	177	270	370
Розміри	6,2× 8,9× 3,2	7,4× 11,4× 3,15	10,8× 10,9× 3,15
Кількість місць	37	59	25
Об’єм/чол.	4,8	4,6	14,8
Заповнення	12 чол. (32%)	10 чол. (17%)	Пусто (0%)
Дистанція ГМ	2,4 м, 5 м, 6,9 м, 4,5 м, 4,5 м	3,5 м, 5,2 м, 8,6 м, 10,3 м, 6,4 м, 6,4 м	2,25 м, 4 м, 5,56 м, 7,1 м, 8,68 м, 10,2 м
Стіни	3 вікна, цегла - штукатурка, задня стінка - верхня частина скло	4 вікна, цегла - штукатурка	3 вікна, бетон - штукатурка
Стеля	бетон - штукатурка	підвісна, акустичні плити	підвісна, акустичні плити
Підлога	паркет	паркет	паркет
Меблі та їх розташування	дерев’яні столи та стільці (3 ряди по 6)	дерев’яні столи та стільці (1 ряд по 9, 2 ряди по 10, 2 книжкових шафи та шафа-гардероб)	дерев’яні столи та стільці (діагональне розташування)

При вимірюваннях БІХ в аудиторіях №1 та №2 тестовий сигнал *x(t)* випромінювали за допомогою гучномовця, розташованого на місці викладача. Відгук приміщення *y(t)* на стимул *x(t)* було записано за допомогою пари мікрофонів, прикріплених до штучної голови, яку розміщували в різних точках приміщення (рис.2.6) [1].

У випадку широкосмугового тестового сигналу з рівномірним спектром, взаємна кореляційна функція $K_{xy}(\tau)$ сигналів *x(t)* та *y(t)* пропорційна БІХ з коефіцієнтом пропорційності k_0 (2.14) [68]:

$$K_{xy}(t) \approx k_0 \cdot h_r(t) \tag{2.15}$$

Враховуючи наявність спотворень в гучномовці та мікрофоні, замість БІХ приміщення натомість буде оцінено згортку:

$$h_{\Sigma}(t) = h_r(t) \otimes h_{lm}(t) \tag{2.16}$$

де $\otimes$ - символ згортки, $h_{lm}(t) = h_l(t) \otimes h_m(t)$ - імпульсна характеристика підсистеми «гучномовець - мікрофон» (ГМ), $h_l(t)$ - ІХ гучномовця та $h_m(t)$ - ІХ мікрофона. Тому при вимірюваннях в аудиторіях №1 та №2 оцінювання БІХ було проведено згідно з виразом:

$$h_r(t) = \mathrm{F}^{-1}\{H_r(f)\} = \mathrm{F}^{-1}\left\{\frac{H_{\Sigma}(f)}{H_{lm}(f)} \cdot M_R(f)\right\} \tag{2.17}$$

де $H_{\Sigma}(f) = \mathrm{F}\{h_{\Sigma}(t)\}$, $H_{lm}(f) = \mathrm{F}\{h_{lm}(t)\}$, F і F^{-1} - символи прямого та зворотнього перетворень Фур'є відповідно, |.| - символ модуля, $M_R(f)$ - регуляризуючий множник [87], що використовується для зниження дисперсії оцінки $h_r(t)$:

$$M_r(f) = \begin{cases} 0{,}5\left[1 + \cos(\pi f \,/\, \Delta F)\right], |f| \le \Delta F \\ 0, \ |f| > \Delta F \end{cases} \tag{2.18}$$

де ΔF - параметр регуляризації [87], прийнято ΔF =18 кГц [1].

Було використано алгоритм корекції частотної характеристики вимірювального тракту штучної голови, розроблений та описаний в [88]. Така кореція є необхідною, оскільки амплітудно-частотна характеристика гучномовця та мікрофону не повністю співпадає зі смугою частот, що аналізуються. В [88] показано, що таку корекцію можна провести за допомогою попередньо отриманої АЧХ системи «гучномовець - мікрофон». Певною проблемою таких обчислень є проведення операції ділення, оскільки АЧХ підсистеми ГМ може містити малі числові значення. Однак, очевидно, що при забезпеченні належного контролю над АЧХ підсистеми ГМ, така операція є практично застосовною. АЧХ підсистеми ГМ набуває найменших значень ближче до границь частотного діапазону, а дисперсія оцінки взаємного спектру системи «гучномовець - приміщення - мікрофон» є найбільшою біля правої границі частотного діапазону, отже доцільно застосувати метод регуляризації, щоб отримати необхідну точність обчислень. Роль регуляційного фактору відіграє вікно - вагова функція, значення якої є близькими до одиниці на низьких та середніх частотах та близькими до нуля на високих. В якості регуляційного фактору було використано вікно Ханнінга, ширина вікна - близько до 80% всього частотного діапазону, що аналізується. В результаті рівень бічних пелюсток імпульсної характеристики підсистеми ГМ було знижено до -44-45 дБ, що лише на 3-4 дБ вище за теоретично досяжний рівень [1].

Обробку отриманих сигналів було проведено наступним чином:

1. Виконано ресемплінг сигналу *x(t)* з частоти запису 22050 Гц до цільової частоти 44100 Гц або 48000 Гц
2. Виконано згортку сигналу *x(t)* з БІХ приміщення *h(t)*, в результаті отримано сигнал *y(t)*.
3. Обчислено значення мір LSD, BSD, PESQ, SSNR та FWSNR.

Обчислення міри LSD було проведено з використанням амплітудних спектрів сигналів *x(t)* та *y(t)* за формулами (2.19), (2.20), (2.21):

$$LSD = \frac{2}{RM} \sum_{m=1}^{M} \sum_{j=1}^{J} \left| G\{X(j,m)\} - G\{Y(j,m)\} \right|, \tag{2.19}$$

$$G\{X(j,m)\} = \max\{20\lg(|X(j,m)|), \delta\}, \tag{2.20}$$

$$\delta = \max_{l,k}\{20\lg(|X(j,m)|)\} - 50, \tag{2.21}$$

де $X(j,m)$ та $Y(j,m)$ - це дискретні перетворення Фур'є фреймів сигналів *x(n)* та *y(n)* відповідно, *j* - номер частотного «відліку» (семпла частоти), *J* - кількість «відліків» частотної вибірки [1].

Обчислення міри BSD було проведено за наступним виразом:

$$BSD = \frac{\sum_{m=1}^{M} \sum_{k=1}^{K} \left[B_x(k,m) - B_y(k,m) \right]^2}{\sum_{m=1}^{M} \sum_{k=1}^{K} \left[B_x(k,m) \right]^2}, \tag{2.22}$$

де $B_x(k,m)$ та $B_y(k,m)$ є барк-спектром *m*-того фрейму сигналу *x(t)* або *y(t)* відповідно, *k* - номер критичної смуги частот [1].

Міра PESQ обчислювалась за алгоритмом для широкосмугових сигналів, описаним в стандарті ITU-T P.862.2 [89] [1].

Обчислення міри STI було проведено за (2.23):

$$STI = \sum_{k=1}^{7} \alpha_k \cdot MT_k - \sum_{k=1}^{6} \beta_k \cdot \sqrt{MT_k \cdot MT_{k+1}}, \tag{2.23}$$

$$MT_k = \frac{1}{14}\sum_{i=1}^{14} T_{ki}, \qquad Tki = \begin{cases} 0, & \mathrm{E}_{ki} < -15; \\ \left(E_{ki} + 15\right)/30, & -15 \le E_{ki} \le +15; \\ 1, & \mathrm{E}_{ki} < -15, \end{cases}$$

$$E_{ki} = 10\lg\frac{m_{ki}}{1-m_{ki}}, \qquad m_{ki} = \left|\int_0^\infty h_{rk}{}^2(t)\exp(-2\pi F_i t)dt\right| \Big/ \int_0^\infty h_{rk}{}^2(t)dt,$$

де α_k - вагові коефіцієнти, β_k - коефіцієнти надмірності, $h_{rk}(t)$ - результат фільтрації сигналу $h_r(t)$ октавним фільтром k-тої смуги (7 октав з центральними частотами від 125 Гц до 8 кГц) [1].

Міру SSNR було обчислено за виразом [77]:

$$SSNR = \frac{1}{M}\sum_{m=1}^{M} 10\lg\left[\frac{\sum_{n=R(m-1)+1}^{Rm} x^2(n,m)}{\sum_{n=R(m-1)+1}^{Rm}\left[x(n,m)-y(n,m)\right]^2}\right], \qquad (2.24)$$

Де $x(n,m)$ та $y(n,m)$ - n-ті вибірки m-го фрейму чистого сигналу $x(n)$ та спотвореного сигналу $y(n)$, відповідно; M - кількість фреймів; R - кількість вибірок у фреймі [1].

Частотно-зважене сегментне відношення сигнал-шум FWSNR [77] також обчислювалося із використанням амплітудних спектрів фреймів сигналів:

$$fwSSNR = \frac{10}{M}\sum_{m=1}^{M}\frac{\sum_{k=1}^{K} W(k,m)\lg\frac{\left|X(k,m)\right|^2}{\left(\left|X(k,m)\right|-\left|Y(k,m)\right|\right)^2}}{\sum_{k=1}^{K} W(k,m)}, \qquad (2.25)$$

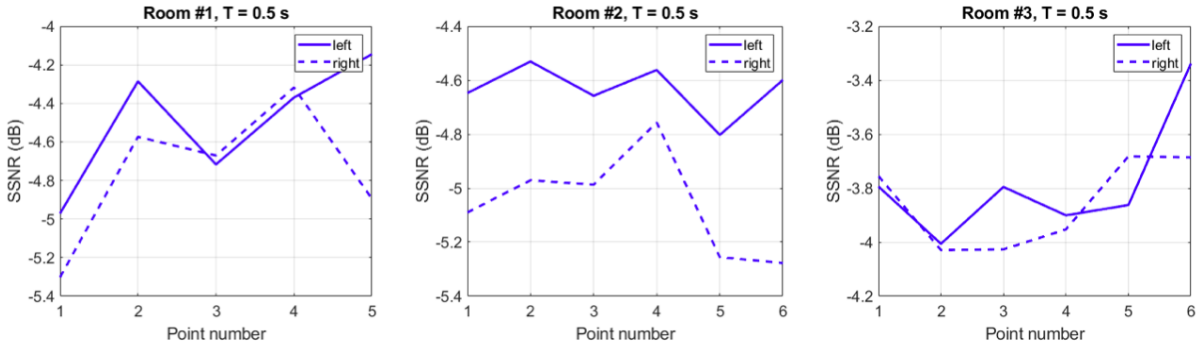
де k - номер критичної смуги; K - кількість критичних смуг; $\left|X(k,m)\right|$ - амплітудний спектр m-го фрейму чистого сигналу, обчислений із

використанням гаусівського вікна; $W(j,m) = |X(j,m)|^{0.2}$ - вагові коефіцієнти [1].

Коефіцієнти кореляції Пірсона між SSNR, LSD, BSD, FWSNR та PESQ, з однієї сторони, та STI, з іншої сторони, обчислювалися згідно стандартних правил [90]. Обчислення було проведено для БІХ приміщення відповідно до розташування штучної голови в кожній з кімнат [1].

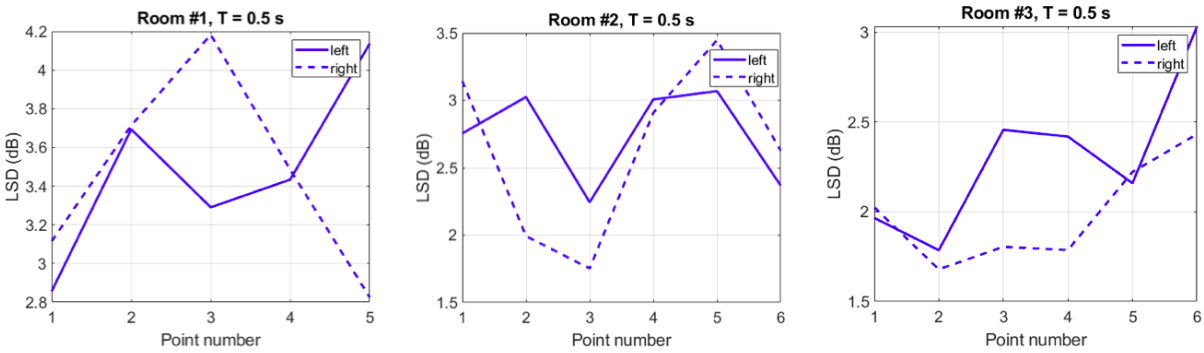
2.2.2 Результати експериментального дослідження - порівняння мір LSD, BSD, PESQ, SSNR, FWSNR та визначення ступеню кореляції з STI

Результати оцінювання параметрів SSNR, LSD, BSD, FWSNR та PESQ за умови, що довжина зразка мовленнєвого сигналу становила L = 15 с, а загальна кількість дикторів становила N = 8, наведено на рис.2.7, 2.8, 2.9, 2.10, 2.11 відповідно. Результати оцінювання STI наведено на рис. 2.12. Зазначені результати представляють собою графіки залежності оцінок вказаних вище параметрів від номерів точок, де розташовувалася штучна голова [1].



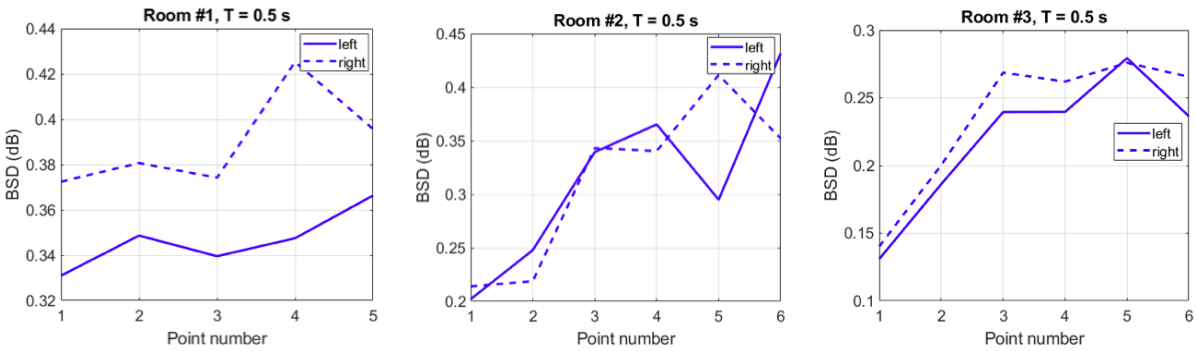
a b c

Рис. 2.7. Оцінки SSNR для приміщень малого (a), середнього (b) та великого (c) розміру [1]



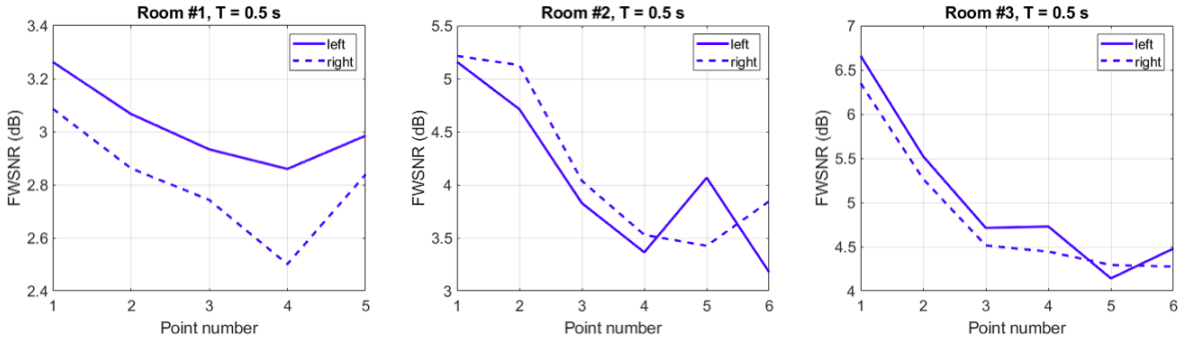
a b c

Рис. 2.8. Оцінки LSD для приміщень малого (a), середнього (b) та великого (c) розміру [1]



a b c

Рис. 2.9. Оцінки BSD для приміщень малого (a), середнього (b) та великого (c) розміру [1]



a b c

Рис. 2.10. Оцінки FWSNR для приміщень малого (a), середнього (b) та великого (c) розміру [1]

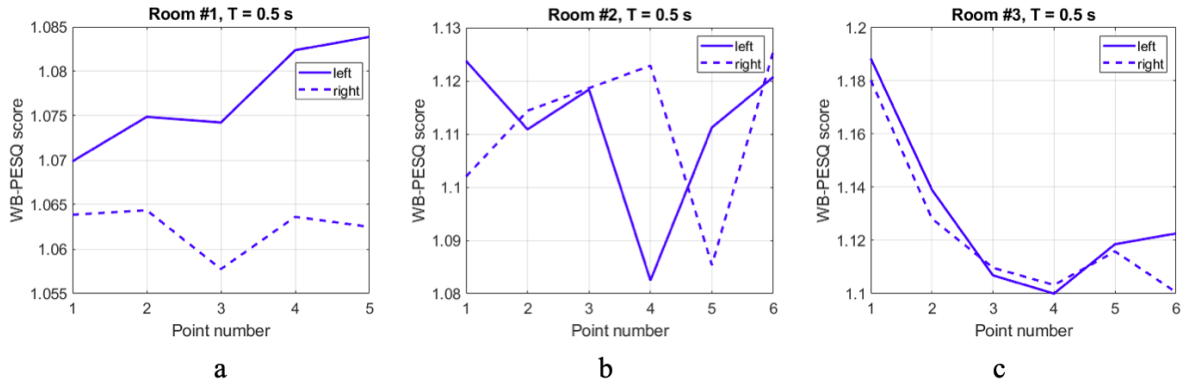


Рис. 2.11. Оцінки PESQ для приміщень малого (a), середнього (b) та великого (c) розміру [1]

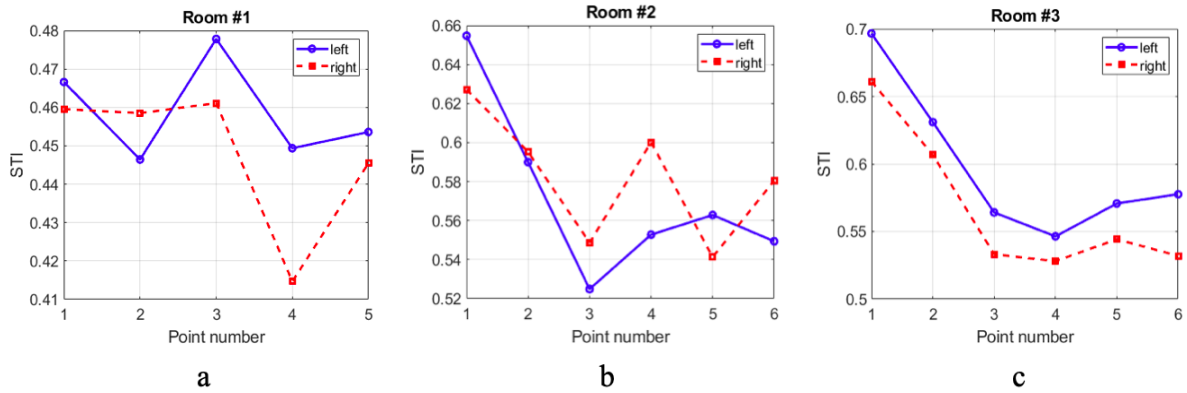


Рис. 2.12. Оцінки STI для приміщень малого (a), середнього (b) та великого (c) розміру [1]

Розглянемо детальніше наведені результати обчислення міри STI (рис. 2.12). Спільним для всіх результатів обчислення STI є зниження розбірливості мовлення в середині аудиторії, а також незначне підвищення розбірливості мовлення біля задньої стінки кімнати (приблизно 30 см). Зниження розбірливості мови в середині кімнати можна пояснити зменшенням співвідношення сигнал/шум зі збільшенням відстані між джерелом і приймачем звуку. Підвищення розбірливості мови на задній стінці кімнати можна пояснити додаванням енергій прямого звуку та звуку, відбитого від задньої стінки. Враховуючи зв’язок між якістю мовлення та розбірливістю, можна було б очікувати, що поведінка об’єктивних показників якості буде подібна до поведінки STI. Оскільки візуально розпізнати таку схожість важко,

доцільно використовувати коефіцієнт кореляції Пірсона як міру такої подібності [1].

Результати оцінювання коефіцієнтів кореляції *R* між SSNR, LSD, BSD, FWSNR та PESQ, з однієї сторони, та STI, з іншої сторони, наведено в табл. 2.4 та на рис. 2.13 [1].

Таблиця 2.4. Значення коефіцієнтів кореляції R [1]

	SSNR-STI		**LSD-STI**		**BSD-STI**		**FWSNR-STI**		**PESQ-STI**	
	L	**R**	**L**	**R**	**L**	**R**	**L**	**R**	**L**	**R**
Ауд. 1	-0,760	-0,604	-0,564	0,216	-0,569	-0,990	0,168	0,788	-0,542	-0,304
Ауд.2	0,055	0,351	0,367	0,146	-0,801	-0,795	0,841	0,651	0,302	0,250
Ауд.3	-0,136	0,050	-0,585	-0,198	-0,922	-0,983	0,929	0,977	0,986	0,958

Аналізуючи представлені результати коефіцієнтів кореляції (табл.2.4), можна побачити, що значення коефіцієнтів кореляції різні для різних мір якості та для каналів вимірювальної системи. Розбіжність значень коефіцієнтів кореляції в лівому і правому каналах вимірювальної системи можна пояснити екрануючим ефектом штучної голови.

Значно більший інтерес представляє поведінка значень коефіцієнта кореляції для різних мір якості в приміщеннях різної площі [1].

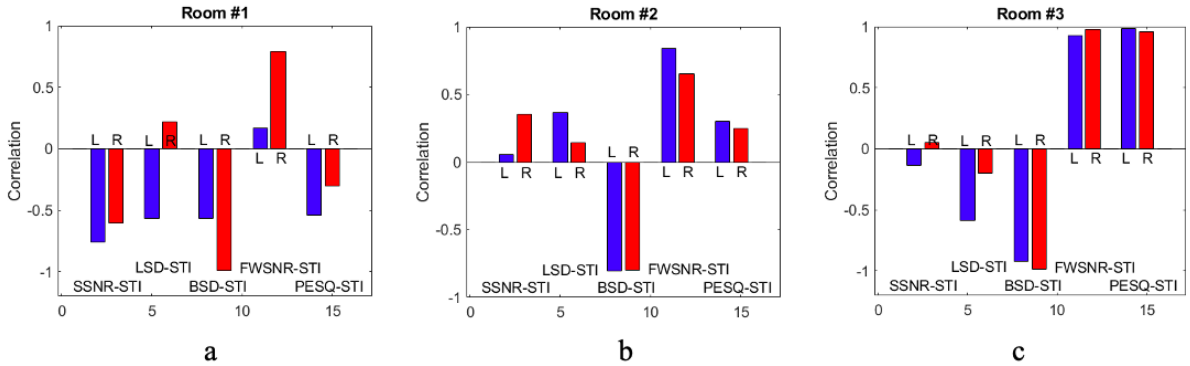


Рис. 2.13. Оцінки R для малих (a), середніх (b) та великих (c) аудиторій [1]

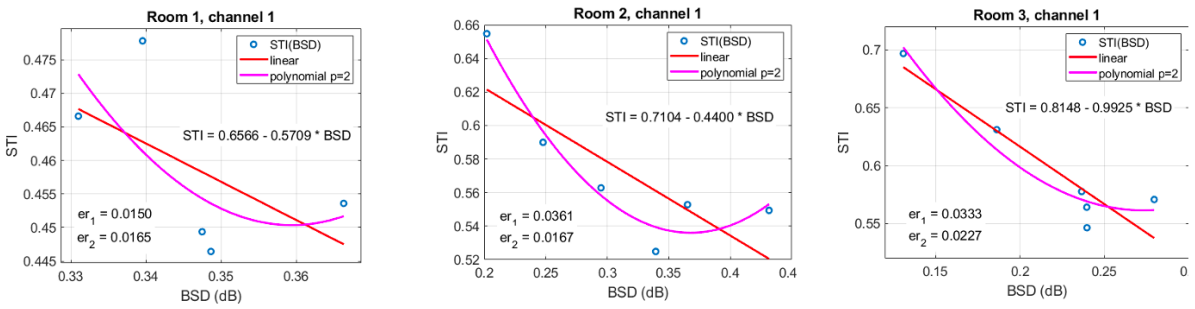
Результати обчислення кореляційного коефіцієнту *R*, наведені в табл.2.4, свідчать про високий рівень кореляції між показниками BSD та STI для приміщень всіх розмірів, при чому чим більший об'єм приміщення, тим вища кореляція. Для приміщення малого розміру *R* набуває значень від -0,57 для лівого каналу до -0,99 для правого каналу; для приміщення середнього розміру значення *R* близькі до -0,8; для великого приміщення коефіцієнт *R* досягає значень від -0,92 до -0,98. Що стосується низьких значень *R*, отриманих в приміщенні малого розміру, їх можна пояснити низькою здатністю задньої стінки аудиторії поглинати звукові хвилі, оскільки її верхня частина - скляна. В результаті формуються сильні перші перевідбиття, що призводить до низької розбірливості та низької якості отриманих мовленнєвих сигналів (рис. 2.8.а, 2.9.а, 2.11.а) [1].

Кореляційний зв'язок між FWSNR та STI є також високим, за винятком приміщення малого розміру [1].

Високу кореляцію між показниками PESQ та STI ми можемо бачити лише в приміщенні великого об'єму (рис. 2.11.с), а при аналізі показників SSNR та LSD бачимо відсутність кореляції з STI в усіх приміщеннях [1].

Аналіз карт зв'язку між значеннями оцінок BSD та STI (рис. 2.14) для лівого каналу (для правого каналу карти зв'язку є подібними) свідчить, що лінійна апроксимація зв'язку між BSD та STI є прийнятною для практичного застосування. Дійсно, похибка апроксимації er_2 залежності STI(BSD) поліномом другого порядку є величиною одного порядку із похибкою лінійної апроксимації er_1 (рис. 2.14). Якщо лінійну залежність STI(BSD) описати виразом $STI = a_0 + a_1 \cdot BSD$, коефіцієнт a_0 приймає значення 0,65, 0,71 та 0,81, а коефіцієнт a_1 приймає значення мінус 0,57, мінус 0,44 та мінус 1 для приміщень №1, №2 та №3, відповідно. Зрозуміло, що наведені значення є приблизними через невелику кількість локацій в кожному із розглянутих приміщень. Враховуючи обмежену кількість початкових та отриманих даних, отримані результати слід розглядати, як певну тенденцію, оскільки для

формулювання надійних тверджень про певні закономірності для їх причини слід збільшити обсяг початкових даних, та, відповідно, кількість обчислених значень [1].



a b c

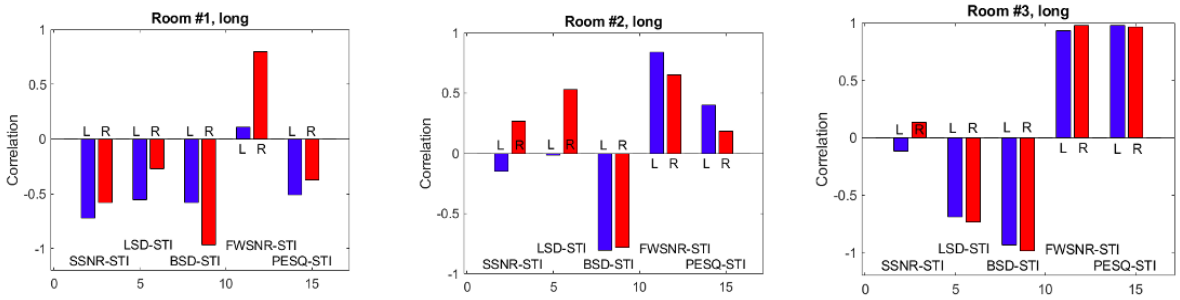
Рис. 2.14. Карти зв’язку між значеннями оцінок BSD та STI для приміщень малого (a), середнього (b) та великого (c) розміру [1]

Отримані результати підтверджують, що використання BSD як міри розбірливості мовлення в приміщеннях різного розміру є можливим і доцільним навіть там, де вплив реверберації є досить значним. Використання PESQ як міри розбірливості мовлення теж є можливим, проте лише у великих приміщеннях, а FWSNR не підійде для приміщень малого розміру. Показники SSNR та LSD не є застосовним для даної задачі [1].

Оскільки наведені вище оцінки R були отримані для певного обсягу даних, доцільно перевірити стійкість оцінок R до збільшення обсягу статистики [1].

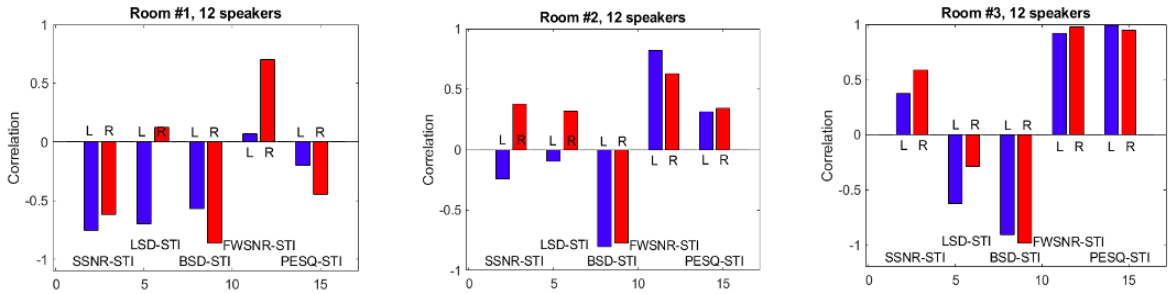
На рис. 2.15 наведено графіки оцінок R для умов N = 8, L = 60 с, а на рис. 2.16 – аналогічні графіки для умов N = 12, L = 15 с [1].

Порівнюючи між собою рис. 2.13, 2.15 та 2.16, не важко дійти висновку, що первинні висновки стосовно характеру та ступеня кореляційних зв’язків між об’єктивними оцінками якості та розбірливості мовлення залишаються незмінними [1]:



a b c

Рис. 2.15. Оцінки R для малих (a), середніх (b) та великих (c) аудиторій за умов N = 8, L = 60 с [1]



a b c

Рис. 2.16. Оцінки R для малих (a), середніх (b) та великих (c) аудиторій за умов N = 12, L = 15 с [1]

1) інтрузивні оцінки якості мовлення у вигляді середнього квадрату похибки у часовій (SSTI) або спектральній (LSD) області, практично не корелюють із оцінками розбірливості мовлення;
2) інтрузивні оцінки якості мовлення BSD та FWSNR, сформовані із врахуванням особливостей слухової системи людини (критичні смуги частот, закон Вебера-Фехнера тощо), корелюють із оцінками розбірливості мовлення, причому ступінь корельованості збільшується із зростанням розміру приміщення; оцінка якості мовлення PESQ корелює із оцінками розбірливості мовлення лише у випадку великих аудиторій [1].

В доповнення зазначимо, що така міра як «енергія модуляції мовлення реверберацією» (SRME), була запропонована в [91] для неінтрузивної діагностики якості реверберованої та дереверберованої мови. Такий показник

якості також є важливим, тому варто було б дослідити її зв’язок із розбірливістю мовлення. Однак такі дослідження виходять за рамки даної дисертаційної роботи.

2.3 Спрощені моделі ранніх відбиттів звуку в приміщенні

Ефективність систем голосового управління значно залежить від типу та інтенсивності звукових хвиль, що створюються у приміщенні, де знаходиться система автоматичного розпізнавання мови. Численні перевідбиття від стелі, підлоги, стін та інших поверхонь додаються до голосової команди, що надходить до мікрофону системи АРМ та негативно впливають на розбірливість мовлення та ефективність роботи системи АРМ [2].

Проблему впливу ранніх відбиттів на розбірливість мовлення досить широко досліджено. Наприклад, в роботі [92] показано, що вплив ранніх відбиттів є майже непомітним для людської слухової системи, якщо відбита хвиля має затримку не більше ніж 30-40 мс відносно прямого звуку. В роботі [93] доведено, що ранні відбиття з часом затримки до 95 мс мають позитивний вплив на розбірливість мовлення, тоді як пізні відбиття - негативний та сприймаються як шумова завада. Цей факт дозволяє використовувати показник C95 в якості міри розбірливості [94]. В роботі [95] розглянуто індекс чіткості C50 та показано, що він краще характеризує вплив реверберації на ефективність системи АРМ, ніж час реверберації T60. Також в [94] показано, що Індекс Передачі Мовлення - Speech Transmission Index (STI) є не менш ефективною мірою розбірливості мовлення, ніж C50 [2].

Оскільки різні частини імпульсної характеристики (ІХ) приміщення впливають на розбірливість мовлення по-різному, в дослідженнях часто «вирізається» потрібна частина та згортується з мовним сигналом. Наприклад, цей метод було застосовано в [96] при аналізі впливу ранніх та пізніх відбиттів на роботу системи АРМ. Також такий метод було використано в [59] при аналізі впливу реверберації на слухову систему людей з кохлеарними

імплантами, а також в [76] при аналізі сприйняття зареверберованої мови людьми без порушень слуху [2].

Проте, замість вирізання потрібної частини ІХ приміщення, можливо синтезувати цю частину за допомогою комп'ютерної симуляції [97], [98], [99]. Зазвичай для виконання такого підходу використовують фізичні [93] або математичні [72], [62] моделі реальних приміщень. В той же час, використання занадто точних та складних моделей реальних кімнат, як, приміром, в роботах [72], [62] потребує значних обчислювальних потужностей [2].

Отже, в рамках даного розділу дисертаційної роботи також запропоновано проаналізувати дві спрощені моделі ранніх відбиттів для аналізу впливу ранніх відбиттів на розбірливість мовлення в приміщенні [2].

2.3.1 Постановка та проведення експерименту

Широко використовуваною є модель спадання енергії звукових відбиттів звуку від поверхонь приміщення, котру в західній літературі іменують як Hybrid Energy Decay Curve [100]. Частиною цієї моделі є модель ІХ приміщення в вигляді потоку імпульсів за виразом:

$$h(t) = \left[\delta(t) + \sum_{n=1}^{N} a_n \delta(t - t_n) \right] e^{-\alpha t}, \tag{2.26}$$

де $\delta(t)$ - дельта-функція Дірака, a_n та t_n - випадкові змінні, N - кількість імпульсів в часовому інтервалі, $\alpha = \dfrac{\ln(10^3)}{T_{60}}$ - параметр, що визначає спад ІХ приміщення, T_{60} - час реверберації [2].

В приміщенні, в основі якого лежить паралелепіпед, щільність потоку імпульсів (2.26) d збільшується з часом за квадратичним законом [68]:

$$d(t) = 4\pi c^3 t^2 \ / V \tag{2.27}$$

В даній роботі запропоновано дві спрощені моделі ранніх відбиттів: перша - одиничне відбиття, друга - послідовність імпульсів сталої щільності. В якості міри розбірливості мовлення, спотвореного реверберацією, використано показник STI. Він має високу узгодженість з результатами суб'єктивного оцінювання розбірливості мовлення [94] [2].

Показник STI було обраховано згідно з модуляційним методом [101], [102] за формулою (2.23) [2].

Модель імпульсної характеристики з одиничним відбивачем може бути описано простим виразом:

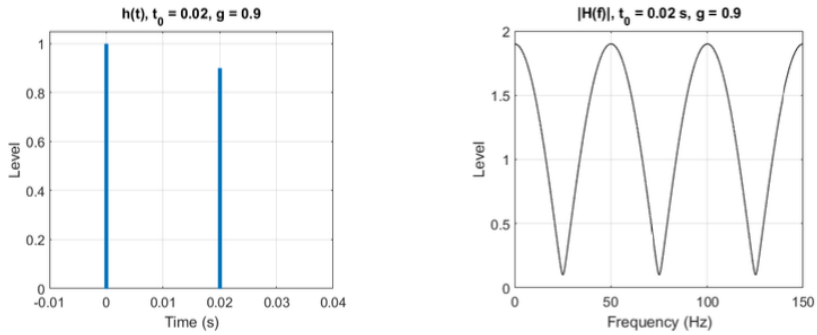
$$h(t) = \delta(t) + g\delta(t - t_0), \tag{2.28}$$

де t_0 - часова затримка, g - рівень відбиття, приймає значення в діапазоні [0;1]. Графік цієї функції та відповідну АЧХ (2.29) наведено на рис. 2.17 для значень t_0 = 0,02 с, g = 0,9.

$$|H(f)| = \sqrt{1 + g^2 + 2g\cos 2\pi f t_0} \tag{2.29}$$

Графіки другої моделі ІХ приміщення, описані (2.26) для T_{60} = 1 с та сталою щільністю відбиттів d = 160 Гц, наведені на рис.2.18 [2].

Графіки АЧХ, наведені на рис. 2.17 та рис.2.18 чітко вказують на те, що основна причина зниження якості мовного сигналу - значна нерівність частотної характеристики приміщення. Оскільки поняття якості мовлення та розбірливості мовлення не мають однозначного зв'язку, графіків рис. 2.17 та рис.2.18 недостатньо для пояснення причин зниження розбірливості, проте можливо використати (2.23) для розрахунку розбірливості мовлення [2].



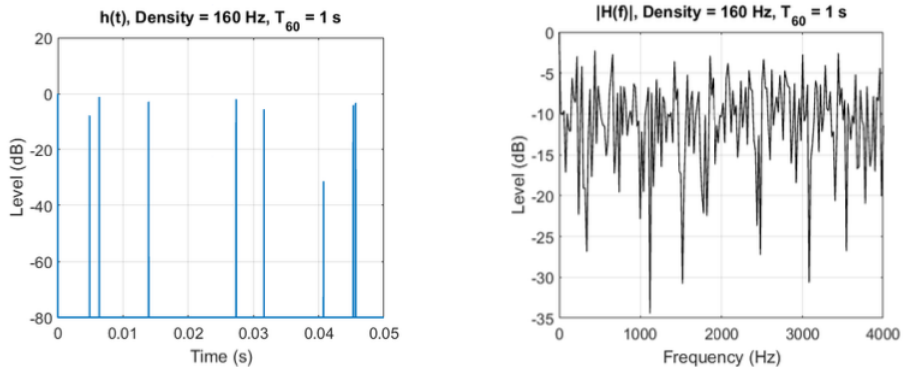
a) b)

Рис.2.17. ІХ (a) та АЧХ (b) першої моделі [2]

Для першої моделі є справедливим:

$$h_k(t) = \int_{-f_{k2}}^{-f_{k1}} H(f)e^{i2\pi f t}df + \int_{f_{k1}}^{f_{k2}} H(f)e^{i2\pi f t}df =$$
$$= 2\left[f_{k2}Sa(2\pi f_{k2}t) - f_{k1}Sa(2\pi f_{k1}t)\right] +$$
$$+2g\left[f_{k2}Sa(2\pi f_{k2}(t-t_0)) - f_{k1}Sa(2\pi f_{k1}(t-t_0))\right], \qquad (2.30)$$

де f_{k1} та f_{k1} - нижня та верхня частоти зрізу *k*-того смугового фільтру з АЧХ прямокутної форми *H(f)*, Sa(*x*) = sin *x/x*. Форму функції (2.30) для *k* = 1 та *k* = 4 для t_0 = 0,029 та *g* = 0,9 наведено на рис. 2.19 зі зміщеннями по осі часу на 0,1 с. Це зроблено для того, щоб можна було врахувати і ліву частину (до піку) функції Sa(x) [2].



a) b)

Рис.2.18 ІХ (a) та АЧХ (b) другої моделі

Для другої моделі є справедливим:

$$h_k(t) = 2\left[f_{k2} Sa(2\pi f_{k2} t) - f_{k1} Sa(2\pi f_{k1} t) \right] + \\ +2\sum_{n=1}^{N} a_n \left[f_{k2} Sa(2\pi f_{k2}(t - t_n)) - f_{k1} Sa(2\pi f_{k1}(t - t_n)) \right], \quad (2.31)$$

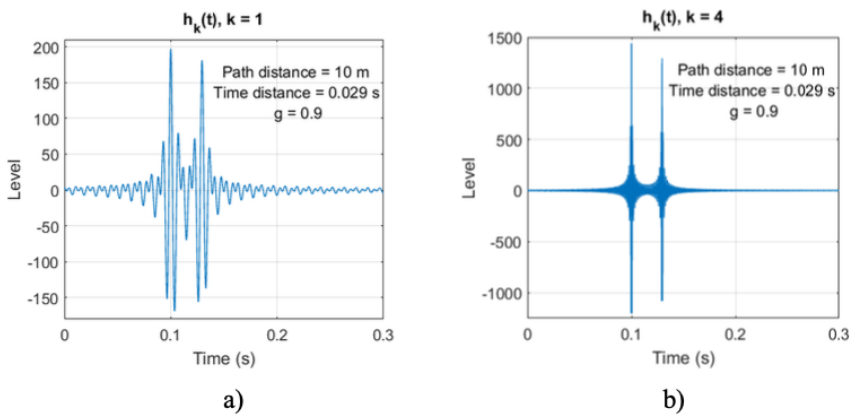


Рис.2.19. Форма функції (2.30) для *k*=1 (a) та *k*=4 (b)

Також варто зазначити, що обчислення STI є відносно простим, використовуючи формули (2.23), (2.28), (2.30) та (2.30) [2].

2.3.2 Результати моделювання

2.3.2.1 Модель з одиничним відбиттям

Результати змодельованої залежності STI від часової різниці прямого та відбитого сигналу для різних відносних рівнів відбитого звуку показано на рис. 2.20. Можна побачити, що у випадку g = 0,9, коли значення різниці в часовому інтервалі є близьким до 30 мс, значення показника STI спадає на одну позицію з «відмінно» до «добре» (табл. 2.5). Подальше зростання різниці в часі до 40 мс призводить до зниження розпізнавання ще на одну позицію - до рівня «задовільно». В той же час, значення STI є близьким до 1, коли значення часової різниці не перевищує 10-13 мс [2].

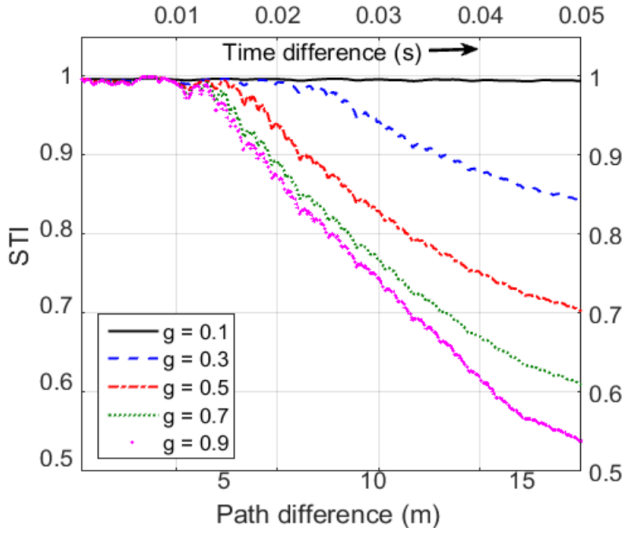


Рис.2.20. Оцінка STI першої моделі [2]

Таблиця 2.5. Значення STI та оцінка розбірливості [2]

STI	Оцінка розбірливості
> 0,75	Відмінно
0,60 - 0,75	Добре
0,45 - 0,60	Задовільно
0,30 - 0,45	Погано
< 0,30	Дуже погано

В випадку помірного відносного рівня відбиття g = 0,5 рівень розбірливості знижується з «відмінно» до «добре» для значень затримки в часі до 40 мс (шлях - близько 13,6 м) [2].

2.3.2.2 Модель з потоком імпульсів сталої щільності

Залежності середнього значення STI від щільності ранніх відбиттів в інтервалі 0–50 мс наведено на рис. 2.21. Усереднення проводили за 100 оцінками STI для значень часу реверберації 0,4 с, 0,7 с і 1 с. Кінці вертикальних ліній відповідають межам 95% довірчого інтервалу [2].

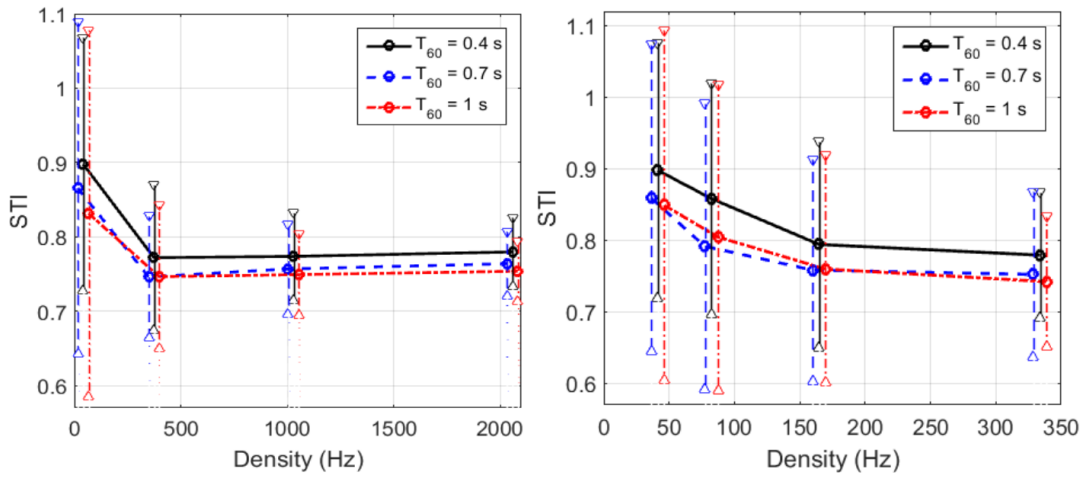


Рис. 2.21. Оцінки STI для другої моделі [2]

На рис. 2.21 розглянуто два випадки. У першому випадку (рис. 2.21.a) значення щільності були встановлені на 40 Гц, 375 Гц, 1030 Гц і 2060 Гц. Можна помітити, що середня розбірливість - мінімальна при щільності близько 375 Гц. Подібний графік для щільності від 40 Гц до 320 Гц, показаний на рис. 2.21.b, дозволяє зробити висновок, що насправді мінімум середньої розбірливості знаходиться біля $d \approx 375$ Гц [2].

Наведені на рис. 2.21 графіки можна аналізувати також з точки зору кількості відбиттів N у часовому інтервалі 0–50 мс. Дійсно, швидке зниження розбірливості зі збільшенням щільності від 40 Гц до 160 Гц (рис. 2..a) відповідає збільшенню середньої кількості відбиттів з 2 до 8. Показник розбірливості продовжує повільно знижуватись зі зростанням значень щільності до 360–375 Гц (18–19 відбиттів). Подальше збільшення N з 18 до 100 навіть дещо підвищує середню розбірливість, однак це збільшення настільки незначне, що можна сказати, що розбірливість мовлення стабілізувалася на мінімальному рівні [2].

З графіків на рис. 2.18 випливає, що збільшення T60 з 0,4 с до 1 с призводить до незначного погіршення розбірливості мовлення. Цей результат також можна пояснити властивостями першої моделі. Дійсно, збільшення T60

призводить до посилення впливу пізніх відбиттів, і, як наслідок, погіршується розбірливість мовлення [2].

Підсумовуючи результати, отримані для другої моделі, можна зробити наступні висновки:

- збільшення середньої кількості відбиттів з 2 до 8 призводить до стрімкого зниження середньої розбірливості мови з «відмінно» до «добре» для значень часу реверберації від 0,4 с до 1 с [2];

- існує мінімум середньої розбірливості мови, що близький до межі між «добре» та «задовільно», коли середня кількість відбиттів близька до 18;

- подальше збільшення середньої кількості відбиттів до 100 призводить до ледь помітного покращення розбірливості мовлення [2];

- для невеликої кількості ранніх відбиттів (від 2 до 8) розбірливість мови може змінюватися в широких межах; цей діапазон помітно звужується зі збільшенням кількості відбиттів [2];

- збільшення значення T60 з 0,4 с до 1 с призводить до незначного зниження розбірливості мови [2].

Отримані результати для першої моделі добре узгоджуються з відомим висновком [92] про те, що перше сильне відбиття несуттєво погіршує розбірливість мови за умови, що затримка відбитого звуку відносно прямого звуку не перевищує 30–40 мс. Крім того, ці результати узгоджуються з висновком [103], що шкідливий вплив першого відбиття збільшується зі збільшенням його амплітуди та часу затримки [2].

Для другої моделі добре підходить також мала (від 2 до 8) кількість відбиттів в інтервалі 0–50 мс. Ця кількість відбиттів близька до реальної ситуації, коли слухач у кімнаті прямокутної форми сприймає 6 первинних відбиттів. Подальше збільшення кількості відбиттів можливе шляхом додавання спеціальних відбивачів, а також шляхом додавання вторинних і третинних відбиттів [2].

Цікаво, що розкид значень розбірливості мови набагато більший для невеликої кількості відбиттів, ніж для великої. Цей результат можна пояснити, використовуючи результати першої моделі (рис. 2.20). Дійсно, при невеликій кількості відбиттів можливі екстремальні ситуації, коли більшість сильних відбиттів зосереджено або на початку інтервалу 0-50 мс (рис. 2.22a), або в його кінці (рис. 2.22.b). У першому випадку розбірливість висока, а в другому – низька. При великій кількості відбиттів такі ситуації малоймовірні, тому розкид значень розбірливості відносно малий [2].

Варто зауважити, що надзвичайно високі та низькі значення розбірливості мови також можна отримати при відносно рівномірному розподілі відбиттів в інтервалі 0-50 мс. У цьому випадку важливе місце розташування сильних відбиттів. Графік на рис. 2.22.c показує випадок, коли відображення розподілені відносно рівномірно, але сильні відбиття зосереджені в кінці інтервалу 0–50 мс, що призводить до зниження розбірливості мовлення [2].

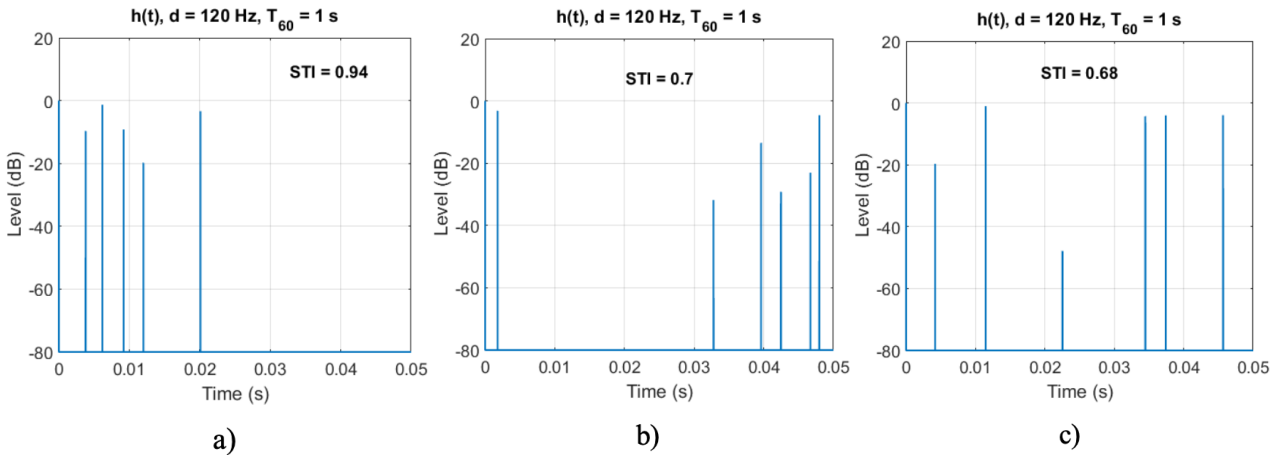


Рис.2.22. RIR для випадку малої кількості відображень: всі відображення зосереджені на початку інтервалу часу 0–50 мс (а); сильні відбиття зосереджені в кінці інтервалу часу (б); більш рівномірний розподіл відбиттів (в). [2]

2.4 Висновки до розділу 2

Даний розділ присвячено висвітленню питань впливу лінійних та нелінійних спотворень на якість та розбірливість мовних сигналів, оскільки вони є факторами, які необхідно враховувати при побудові системи АРМ. Розглянуто три питання:

1) виконано аналіз ефективності коефіцієнта ексцесу в якості міри оцінювання кліпування;
2) виконано аналіз зв'язку мір оцінювання якості та розбірливості мовлення в приміщеннях різного розміру;
3) запропоновано дві спрощені моделі ранніх відбиттів для аналізу впливу відбиттів на розбірливість мовлення в приміщенні.

Отримані результати дозволяють зробити наступні висновки.

1. Коефіцієнт ексцесу β_4 як міру оцінювання кліпування запропоновано на противагу використовуваному раніше коефіцієнту кліпування R_{cl}. Показано, що коефіцієнт ексцесу β_4, а також обернений коефіцієнт ексцесу γ_4 та квадратний корінь оберненого коефіцієнту ексцесу η_4 можуть бути застосовані для оцінювання кліпування мовного сигналу. Основною їх перевагою є значно спрощене обчислення (порівняно з R_{cl}), що суттєво знижує час та ресурси при розрахунках Аналіз карт відповідності між суб'єктивними та об'єктивними (з використанням запропонованих мір кліпування) оцінками якості мовлення, спотвореного кліпуванням свідчить про суттєву перевагу практичного використання міри η_4, з огляду на лінійний характер зв'язку із суб'єктивною оцінкою якості спотвореного сигналу.

2. Зв'язок між мірами оцінювання якості та розбірливості мовлення в приміщеннях різного розміру проаналізовано шляхом обчислення коефіцієнта кореляції Пірсона між мірами якості мовленнєвих сигналів SSNR, LSD, BSD, FWSNR та PESQ та оцінками розбірливості мовлення STI. Показано, що для пари BSD-STI кореляція є високою в аудиторіях всіх розмірів. Для пари

FWSNR-STI кореляція є високою для приміщень середнього та великого розміру. Для пари PESQ-STI кореляція є високою лише для аудиторій великого розміру. Щодо мір SSNR та LSD, їх оцінки практично не корелюють із оцінками STI.

3. Розглянуто дві моделі ранніх відбиттів звуку в кімнаті. Аналіз властивостей першої моделі, що передбачає одиничне відбиття в інтервалі часу 0–50 мс, дозволив отримати залежності STI від затримки та сили відбитого звуку. Залежності узгоджуються з результатами попередніх досліджень, які підтверджують валідність використаного методу. У другій моделі випадкові відбиття випадково, за рівномірним законом, розподіляються на інтервалі часу 0-50 мс. Аналіз властивостей цієї моделі показав, що для невеликої (до 8) кількості відображень розбірливість мовлення змінюється від відмінної ($STI \approx 1$) до задовільної ($STI \approx 0{,}6$). Високі значення STI можна отримати, коли найсильніші відбиття зосереджені на початку інтервалу 0–50 мс. І навпаки, найнижчі значення STI виникають, коли найсильніші відбиття знаходяться в кінці інтервалу.

3 ПІДВИЩЕННЯ РОБАСТНОСТІ СИСТЕМИ АВТОМАТИЧНОГО РОЗПІЗНАВАННЯ МОВИ ДО ДІЇ ШУМОВОЇ ЗАВАДИ

В даному розділі наведено результати експериментальних досліджень, спрямованих на підвищення робастності системи АРМ до дії шумової завади. Цим результатам передує короткий огляд мір якості систем автоматичного розпізнавання мови й, зокрема, мір якості, що використовуються при оцінюванні точності розпізнавання в програмному комплексі The Hidden Markov Model Toolkit (HTK) [17].

3.1 Оцінювання якості систем автоматичного розпізнавання мови

При моделюванні системи АРМ важливо надати оцінку якості її роботи. Ідея такого оцінювання є досить простою: потрібно подати на вхід системи АРМ певну кількість слів, а на виході підрахувати кількість правильно розпізнаних слів та кількість помилок розпізнавання [18].

Найбільш часто при оцінюванні якості системи АРМ використовують показник Word Error Rate (WER) – відносну помилку розпізнавання [17]:

$$WER = \frac{S + D + I}{N + H + S + D} \tag{3.1}$$

та таку величину як точність розпізнавання мови:

$$Acc = 1 - WER \tag{3.2}$$

де N – загальна кількість слів, що піддавалися розпізнаванню, H – кількість вірно розпізнаних слів, S – кількість замін (слово розпізнано невірно), D – кількість видалень (слово пропущене), I – кількість вставок (розпізнано певне

слово, хоча його насправді немає у вхідному сигналі; зазвичай таке трапляється в тривалих паузах між словами, де фоновий шум сприймається як тихе мовлення). За відсутності вставок *I* значення показника WER належать інтервалу [0, 1], проте при дуже великій кількості вставок показник *Acc* може набувати від'ємних значень, що з практичної точки зору незручно [17], [18].

Для усунення цього недоліку можливо не враховувати вставки у формулах (3.3) та (3.4):

$$\%Correct = \frac{H}{N} \cdot 100\% \tag{3.3}$$

$$P_e = \frac{S + D}{N} \tag{3.4}$$

проте і у цьому випадку при значній кількості вставок показники (3.3) та (3.4) не будуть адекватно відображати якість роботи системи [17], [18].

Для пошуку показників, альтернативних зазначеним вище, в [104] сформовано вимоги, яким мають відповідати нові критерії: по-перше, вони мають бути математично обґрунтованими, по-друге, вони повинні бути інтуїтивно зрозумілими, по-третє, – відносно простими для обчислення. В роботі [104] зазначено, що таким показником може бути «узгоджений процент помилок» - Match Error Rate (MER) [105], [18]:

$$MER = \frac{S + D + I}{N + H + S + D + I} = 1 - \frac{H}{N} \tag{3.5}$$

Окрім цього, в [104] запропоновано новий показник, що відображає «втрачену словесну інформацію» (Word Information Lost – WIL), а також пов'язаний з ним показник, що описує «збережену словесну інформацію» (Word Information Preserved - WIP):

$$WIP \cong \frac{I(X,Y)}{H(Y)} \cong \frac{(H - N_1 N_2 / nN)^2}{N_1 N_2} \tag{3.6}$$

$$WIL = 1 - WIP \tag{3.7}$$

де $I(X,Y)$ - повна взаємна інформація про набори слів X та Y на вході та виході системи АРМ відповідно; $H(Y)$ - ентропія послідовності слів Y; N_1 та N_2- кількість слів на вході та виході системи АРМ відповідно; N - кількість узгоджених за алгоритмом Вітербі пар вхідних та вихідних слів; H – кількість правильно розпізнаних слів; n – обсяг словника. Обидва показники – MER та WIL – в передбачених випадках набувають значення від 0 до 1 [18], [105].

Недоліком показників (3.6) та (3.7) є відносна складність їх аналітичного виведення, а також складність практичної перевірки їх справедливості; також користувачу системи необхідно вміти отримати з інструментарія системи вихідні дані для проведення обчислень згідно з (3.6) [18].

3.2 Міра точності розпізнавання мови у системі автоматичного розпізнавання мови The Hidden Markov Model Toolkit [35, 36]

Для досягнення мети даної дисертації моделювання системи автоматичного розпізнавання мови відбувалось в програмному комплексі The Hidden Markov Model Toolkit (HTK) [17], [18] для побудови та використання систем АРМ, основою яких є приховані марковські моделі.

Для аналізу якості роботи системи АРМ в програмному комплексі HTK здійснюється построкове співставлення вихідних послідовностей слів з відповідними файлами транскрипцій, засноване на алгоритмі динамічного програмування. Передбачено дві міри якості системи: *%Correct*, що враховує тільки ідентичні слова вхідних та вихідних послідовностей слів, та %Accuracy, у якому здійснюється порівняння цих же файлів на предмет наявності вставок, замін та видалень. Ці показники відповідно розраховуються за формулами (3.3) та (3.8) [17]:

$$\%Accuracy = \frac{H - I}{N} \cdot 100\% \qquad (3.8)$$

де H - кількість вірно розпізнаних слів, I - кількість вставок, N - загальна кількість слів, що розпізнаються [17], [18].

3.3 Дослідження робастності системи АРМ до дії шумової завади: постановка та проведення експерименту

В даному розділі наведено результати чотирьох експериментів: навчання системи АРМ за методами Fully-Matched Training (FMT), Noise-Matched Training (NMT, інша назва Spectrum Matched Training - SMT), Signal-to-Noise Ratio Matched Training (SNRMT) та Multistyle Training (MT). Відмінність методів полягає в співпадінні характеристик навчальної та тестової вибірок з точки зору типу шумової завади та відношення сигнал-завада. FMT передбачає повне співпадіння, NMT – співпадіння по типу шуму. SNMRT – співпадіння по відношенюю сигнал-завада, MT містить велику вибірку з усіма можливими варіантами поєднання цих характеристик.

Сигнали, що зашумлюються, - це попередньо записане мовлення одного диктора: 200 зразків 10 слів російської мови: числа від одного до десяти, кожне слово записане 20 разів з різною, наскільки це можливо, інтонацією. Фонемний словник складається з 22 фонем російської мови. Використовувались 39-мірні класифікаційні ознаки типу MFCC_0_D_A (мел-частотні кепстральні коефіцієнти). SNR навчальної вибірки становить 45 дБ. Параметри сигналів: частота дискретизації $f_d = 22050$ Гц, глибина квантування – 16 біт.

Для тестування системи АРМ була сформована адитивна суміш сигналу та шуму (рис. 3.1):

$$s(t) = k \cdot x(t) + n(t), \qquad (3.9)$$

де *x(t)* – мовний сигнал, *n(t)* – шум, *k* – поправочний коефіцієнт, що забезпечує необхідне відношення сигнал-шум SNR_0 та розраховується за формулою:

$$k = 10^{0.05(SNR_0 - SNR)} \tag{3.10}$$

де *SNR* – реальне відношення сигнал-шум записаного сигналу [106].

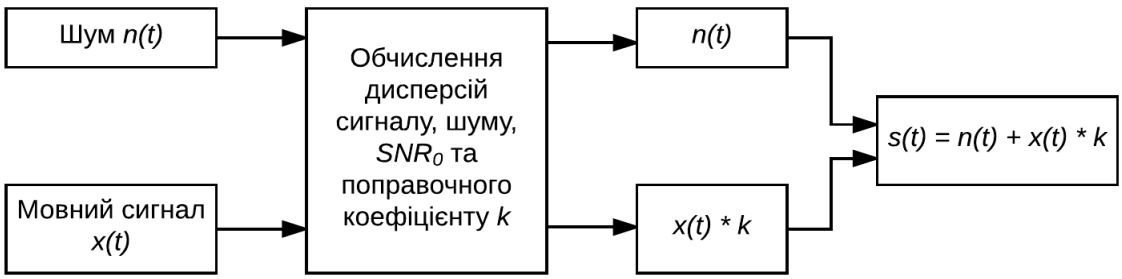


Рис. 3.1. Схема моделювання зашумленого сигналу [18]

Тестові сигнали являли собою шість зашумлених звукових файлів дискретної мови із записом усіх десяти слів, що використовувались при навчанні, з паузами між словами тривалістю 0,3 – 0,5 с та різними SNR від 0 до 45 дБ з кроком 5 дБ.

Загальна кількість файлів, створених для вирішення задачі даного розділу дисертаційної роботи, – 28840. Для пришвидшення проведення експерименту для кожного з етапів було написано комп’ютерні програми мовою Python.

Використано 14 шумів навколишнього оточення (таблиця 3.1), що належать до різних сфер людської діяльності та найчастіше можуть бути завадами високій якості розпізнавання. Такий вибір шумів зумовлено тим, що потрібно оцінити працездатність системи у реальних шумових умовах, тож було обрано ті, що зазвичай мають високий рівень та можуть істотно завадити роботі системи АРМ.

За методом Fully-Matched Training навчальна вибірка складалась виключно з файлів з одним спектром шуму та значенням SNR:

$$SNR_t = SNR_r,\ n_t(t) = n_r(t), \tag{3.11}$$

де SNR_t – відношення сигнал-завада навчальної вибірки, SNR_r – відношення сигнал-завада тестової вибірки, $n_t(t)$ – спектр шуму навчальної вибірки, $n_r(t)$ – спектр шуму тестової вибірки.

Таблиця 3.1. Шуми навколишнього оточення

	Сфера застосування	Шум
Acc,%	Транспорт	Вулиця, вкладена бруківкою
		Площа перед вокзалом
		Тролейбусна зупинка
		Вантажівки на пр. Перемоги
	Шуми приміщення: транспорт	Підземний перехід між вокзалами
		Поїзд метро під час розгону
		Фойє центрального вокзалу
		Фойє метро
		У троллейбусі
	Шуми приміщення: офіс та побутові	Аудиторія
		Мікрохвильова піч
		Комп'ютер
		Кавомолка
		Пральна машина

Експеримент було проведено для 140 ситуацій: по десять вибірок для кожного значення SNR для кожного з 14 шумів.

За методом Noise-Matched Training навчальна вибірка складається з сигналів, зашумленим одним певним шумом, але з усіма значеннями SNR:

$$SNR_t \neq SNR_r,\ n_t(t) = n_r(t) \tag{3.12}$$

В даній частині експерименту було сформовано 14 вибірок згідно з кількістю використаних шумів.

За методом Multistyle Training формується одна навчальна вибірка, яка складається з усіх можливих варіантів зашумленої мови, отже маємо різні спектри шумів та різний рівень зашумленості:

$$SNR_t \neq SNR_r,\ n_t(t) \neq n_r(t) \tag{3.13}$$

За методом Signal-to-Noise Ratio Matched Training навчальна вибірка складається з шумів одного і того самого рівня зашумленості, але різних спектрів:

$$SNR_t = SNR_r,\ n_t(t) \neq n_r(t) \tag{3.14}$$

Згідно з тим, що в рамках даної роботи відношення сигнал-завада може набувати десяти різних фіксованих значень (від 0 до 45 дБ з кроком 5 дБ), в даній частині експерименту проведено 10 дослідів.

Для наочності порівняння методів їх математичний опис також наведено в таблиці 3.2:

Таблиця 3.2. Методи навчання системи АРМ

Назва методу	*Співпадіння (Matching)*
Fully-Matched Training	$SNR_t = SNR_r,\ n_t(t) = n_r(t)$,
Noise-Matched Training	$SNR_t \neq SNR_r,\ n_t(t) = n_r(t)$
Multistyle Training	$SNR_t \neq SNR_r,\ n_t(t) \neq n_r(t)$
Signal-to-Noise Ratio Matched Training	$SNR_t = SNR_r,\ n_t(t) \neq n_r(t)$

3.4 Результати експериментальних досліджень системи АРМ, що навчалася на чистих сигналах

Для отримання початкової характеристики ефективності роботи системи АРМ, проведено її навчання на незашумлених сигналах. Отримані результати зведено в таблицю 3.2. та проілюстровано на рис.3.2.

Аналізуючи отриманий результат, можна зробити висновок, що для досягнення високої точності розпізнавання (95%) необхідно забезпечити відношення сигнал-шум тестового сигналу щонайменше 17 дБ, що в умовах реальної експлуатації зробити досить складно. Отже, система АРМ, навчена на чистих (не зашумлених) мовленнєвих сигналах, не є робастною (стійкою) до дії шумових завад [8], [9]. Крім того, із рис. 3.2 видно, що в окремих випадках потребується набагато вище значення SNR для забезпечення точності розпізнавання 95%. Так, найгіршими є перехід між вокзалами (майже 40 дБ), фойе метро та брукована вулиця (близько 30 дБ). Як це не дивно, але зашумлену аудиторію можна порівняти із тролейбусною зупинкою – в обох випадках точність розпізнавання *Acc* = 95% може бути досягнутою лише для значення SNR > 25 дБ.

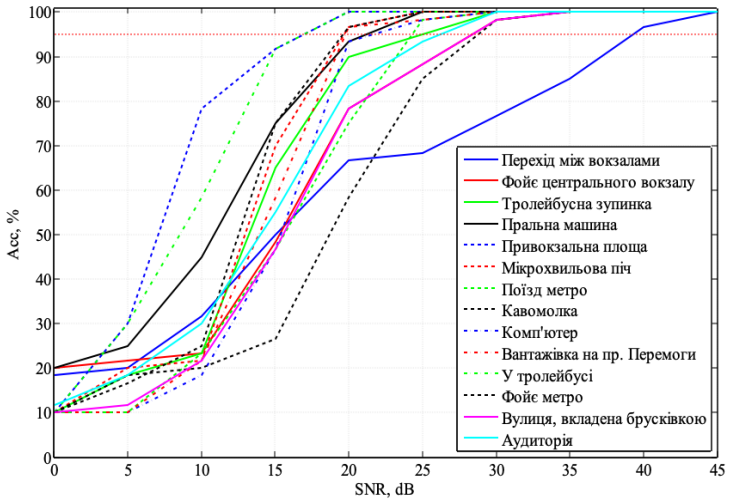


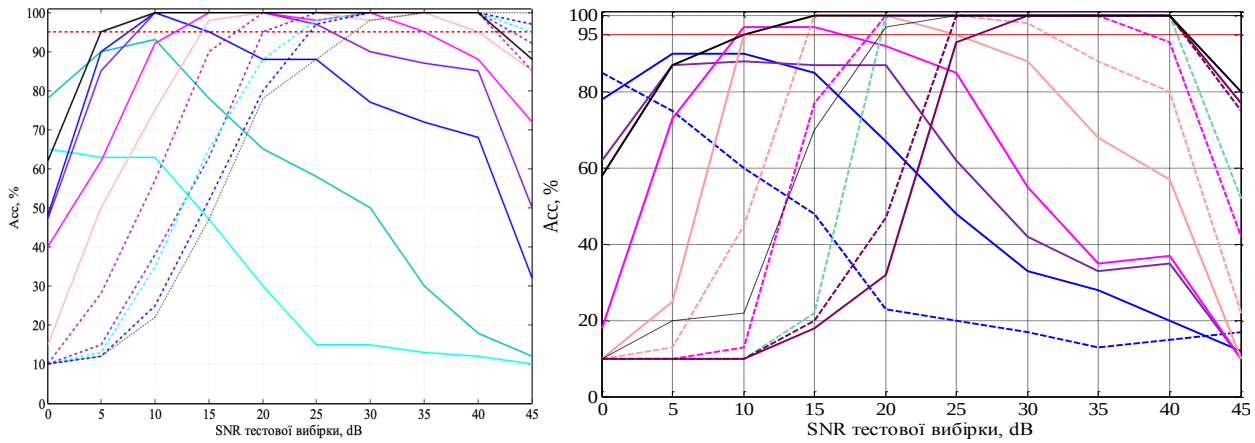
Рис. 3.2. Точність розпізнавання при навчанні методом clean training [9]

Таблиця 3.2. Точність розпізнавання мови для різних шумів при SNR навчальної вибірки 45 дБ [9]

	SNR	0	5	10	15	20	25	30	35	40
Acc,%	Вулиця, вкладена бруківкою	10	12	22	47	78	88	98	100	100
	Аудиторія	12	18	30	55	83	93	100	100	100
	Фойє метро	10	17	25	75	97	100	100	100	100
	У троллейбусі	10	30	58	92	100	100	100	100	100
	Вантажівки на пр. Перемоги	10	10	22	58	97	98	100	100	100
	Комп'ютер	10	10	18	47	93	98	100	100	100
	Кавомолка	10	18	20	27	58	85	98	100	100
	Підземний перехід між вокзалами	18	20	32	50	67	68	77	85	97
	Поїзд метро під час розгону	10	10	23	47	75	98	100	100	100
	Мікрохвильова піч	10	20	22	70	97	100	100	100	100
	Площа перед вокзалом	10	30	78	92	100	100	100	100	100
	Пральна машина	20	25	45	75	93	100	100	100	100
	Тролейбусна зупинка	10	18	23	65	90	95	100	100	100
	Фойє центрального вокзалу	20	22	23	48	78	88	98	100	100

3.5 Результати експериментальних досліджень системи АРМ, що навчалася за методом Fully-Matched Training

У цьому експерименті навчання системи АРМ виконувалося на зашумлених сигналах із SNR від 0 до 45 дБ з кроком 5 дб. Тестування також виконувалося на зашумлених сигналах із SNR від 0 до 45 дБ з кроком 5 дб. Метою цієї частини було встановити ефективність роботи системи, якщо навчати її на сигналах з такими самими шумовими умовами, у яких вона буде експлуатуватись. На рис. 3.3. відображено результати для шуму вулиці, вкладеної бруківкою (a) та шуму мікрохвильової печі (b). Графіки залежності точності розпізнавання від відношення сигнал-шум для інших шумів навколишнього оточення, а також табличне представлення результатів наведено в додатках Б та В відповідно.



(a) (b)

Рис. 3.3. Графік залежності точності розпізнавання від відношення сигнал-шум тестової вибірки для шуму вулиці, вкладеної бруківкою (a). та шуму мікрохвильової печі (b) SNR навчальних вибірок:

0 dB 5 dB 10 dB 15 dB 20 dB 25 dB 30 dB
35 dB 40 dB 45 dB Універсальна вибірка "Чисті" сигнали

За отриманими результатами можна зробити висновок, що найефективніше система працює у тому випадку, коли ступінь зашумленості навчальної та тестової вибірок збігається. Показано, що недоліком методу Fully-Matched Training є суттєва втрата ефективності, якщо шумові умови в режимі розпізнавання відрізняються від таких в режимі навчання [8], [9].

3.6 Результати експериментальних досліджень системи АРМ, що навчалася за методом Noise Matched Training

Метод Noise Matched Training полягає в навчанні системи на сигналах, зашумлених таким же шумом, який буде наявним в режимі розпізнавання. Метою його було визначення того, чи стане система більш завадостійкою, якщо навчальна вибірка міститиме усі можливі варіанти відношень сигнал-шум. Результати експерименту проілюстровано на рис. 3.4 та наведено в таблиці 3.3 [9].

Таблиця 3.3. Залежність точності розпізнавання від SNR тестових вибірок, дБ

	SNR тестових вибірок, дБ	0	5	10	15	20	25	30	35	40	45
	Фойє метро	50	88	98	100	100	100	98	98	100	85
	У троллейбусі	58	92	98	100	100	100	100	100	100	92
Acc, %	Вантажівки на пр. Перемоги	50	88	100	100	100	100	100	100	100	72
	Комп’ютер	50	85	97	100	100	100	100	100	100	60
	Кавомолка	58	87	100	100	98	97	95	98	98	98
	Поїзд метро під час розгону	42	45	65	80	92	97	98	100	100	85
	Мікрохвильова піч	58	87	95	100	100	100	100	100	100	80
	Привокзальна площа	72	90	93	95	95	97	100	100	100	100
	Підземний перехід	83	93	95	100	100	100	100	100	100	88
	Пральна машина	47	77	98	100	100	100	100	100	100	98
	Тролейбусна зупинка	60	92	97	100	100	100	100	100	100	92
	Фойє центрального вокзалу	55	93	100	100	100	100	100	100	100	97
	Вулиця, вкладена бруківкою	62	95	100	100	100	100	100	100	100	88
	Аудиторія з 13 людьми	47	67	82	87	90	95	98	98	98	80

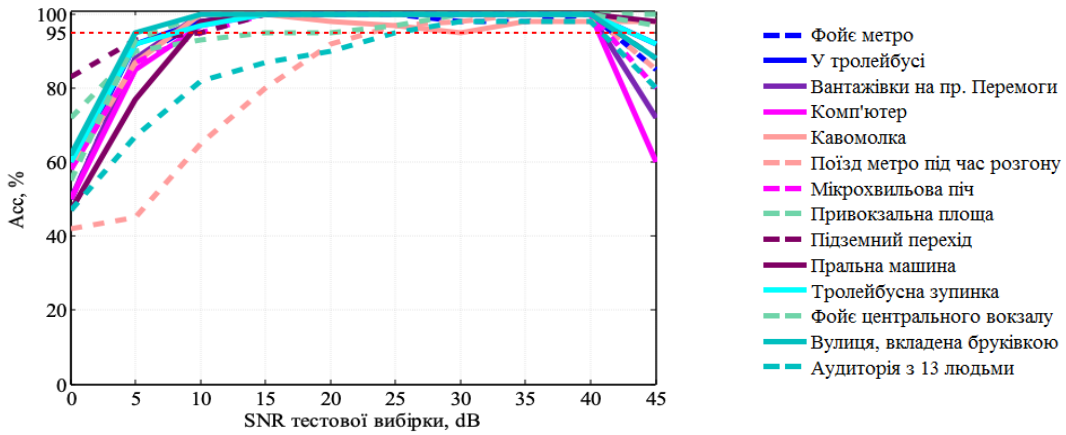


Рис. 3.4. Точність розпізнавання системою АРМ сигналів різної зашумленості (по горизонтальній вісі відкладені SNR тестових сигналів)

За результатами цього експерименту бачимо, що точність розпізнавання мовних сигналів суттєво збільшилась порівняно з методом FMT, особливо це стосується сигналів з низьким значенням SNR. У той час як при навчанні на «чистих» сигналах найкраща точність *Acc* для SNR = 3дБ дорівнювала 20%,

при навчанні системи АРМ на сигналах усіх рівнів зашумленості отримуємо мінімальне значення точності *Acc* = 42%; вдалось також досягти 95% точності для сигналів з SNR = 5 дБ, а для 12 з 14 шумів точність *Acc* = 95% і вище досягається при SNR = 10 дБ, що, порівняно з навчанням на чистих, є набагато кращим результатом [8].

Порівнюючи дані цього етапу з даними, отриманими при навчанні за методом FMT, бачимо, що певної закономірності підвищення чи зниження якості розпізнавання немає, для різних шумів ситуація різна. Отже, при встановленні такої системи АРМ на реальні прилади чи пристрої необхідно враховувати види шумів, за яких вона буде працювати, а також можливу мінливість рівня шуму. Навчаючи систему на вибірках з різними SNR послідовно, необхідно також оснастити її програмою вимірювання відношення сигнал-шум реального сигналу для подальшого вибору необхідних еталонних зразків для розпізнавання, що займає певний обсяг пам’яті та потребує часу. У випадку навчання на вибірці, що містить усі зашумлені сигнали, цього робити не потрібно, і хоч система АРМ, навчена на вибірці такого типу, для деяких шумів показує нижчу ефективність за низьких SNR (3 дБ), вона все ж таки є більш стійкою до завад, якщо відношення сигнал-шум в процесі мовлення буде змінюватись [8].

Слід зазначити ще одну особливість такого навчання: під час тестування вибіркою з SNR = 45 дБ відбулась певна втрата якості розпізнавання. Пояснити це можна наступним чином: під час навчання система формує один еталонний зразок, в якому присутні спільні для усіх навчальних зразків риси. Оскільки шум в більшій чи меншій мірі вплинув на спектр більшості навчальних сигналів, еталонний зразок було створено з урахуванням цих «спотворень», і при розпізнаванні сигналів з SNR = 45 дБ система робила помилки. Проте, у реальних умовах відношення сигнал-шум SNR ≥ 45 дБ майже не зустрічається, тож цим недоліком такого типу навчання можна знехтувати [8].

Порівнюючи працездатність системи АРМ в умовах різних шумів (табл.3.3), бачимо, що найгірше розпізнавання відбувається при зашумленні шумом поїзда, що розганяється, та шумом, записаним в аудиторії, в якій присутні 13 людей. У випадку шуму аудиторії зниження якості зумовлено відносно великою кількістю вставок (рис. 3.5), які виникають внаслідок того, що система помилково розпізнає шуми фонової мови як основний сигнал [8].

```
------------------------ Overall Results --------------------------
SENT: %Correct=0.00 [H=0, S=6, N=6]
WORD: %Corr=58.33, Acc=46.67 [H=35, D=1, S=24, I=7, N=60]
===================================================================
```

Рис. 3.5. Скріншот результатів оцінки точності розпізнавання при тестуванні вибіркою з SNR = 0 дБ (H – кількість вірно розпізнаних слів; I – кількість вставок)

3.7 Результати експериментальних досліджень системи АРМ, що навчалася за методом SNR-matched training

На даному етапі було проведено 10 дослідів: кожен дослід - навчання системи на вибірці шумів різних джерел, із рівнями відношення сигнал-завада SNR = 0, 5, 10… 45 дБ.

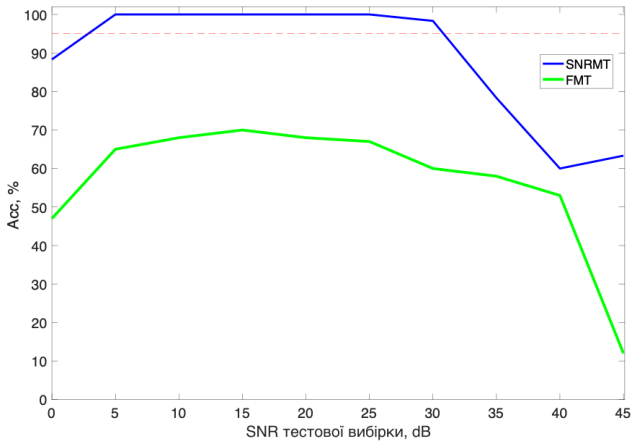


Рис. 3.6. Результати експерименту SNR-matched training, фойє центрального залізничного вокзалу [8]

На рис. 3.6 наведено залежність точності розпізнавання від відношення сигнал-шум тестової вибірки для навчання на синалах з рівнем SNR = 5 дБ (синій колір). Шум тестової вибірки – фойє центрального залізничного вокзалу. Для порівняння наведено також результати навчання за методом Fully-Matched training (FMT) [8], [9]. Схожість цих двох методів полягає в тому, що рівень зашумленості співпадає, проте спектри шумів різні: виключно шум фойє центрального залізничного вокзалу для FMT та усі спектри з використаних в даній роботі для SNRMT. Як видно з графіку, метод SNRMT значно перевищує FMT за ефективністю: 95% точність досягається вже при значеннях SNR>3 дБ, тоді як для методу FMT найвище значення точності сягає 70%. Слід однак зазначити, що ефективність методу SNRMT спадає при SNR>30 дБ. Це можна пояснити тим, що мовні сигнали, що подаються на вхід системи при навчанні, містять додаткові спектральні компоненти, наявні у шумі. При високих значеннях SNR тестового сигналу ці компоненти відсутні, що ускладнює розпізнавання. Оскільки в реальних умовах експлуатації такі досить високі значення SNR забезпечити складно, спосіб може бути застосовним для багатьох варіантів умов, проте важливим є дотримання певної стабільності рівня та спектру навколишнього шуму [8], [9].

3.8 Результати експериментальних досліджень системи АРМ, що навчалася за методом Multistyle Training

В даній частині експерименту було створено одну велику навчальну вибірку з усіма можливими в рамках даної роботи варіантами зашумлення: зі значеннями відношення сигнал-шум від 0 до 45 дБ та усіма чотирнадцятьма шумами навколишнього оточення.

Результати експерименту наведено на рис. 3.7 та у таблиці 3.2 [8].

Таблиця 3.2. Точність розпізнавання при навчанні методом Multistyle training [8]

Noises	SNR, (dB)								
	0	5	10	15	20	25	30	35	40
Вулиця, вкладена бруківкою	58	83	100	100	100	98	100	100	100
Вантажівки на пр. Перемоги	32	62	97	100	100	100	100	100	100
Тролейбусна зупинка	40	78	100	100	100	100	100	100	100
Поїзд метро під час розгону	25	50	92	97	100	100	100	100	100
Фойє метро	28	65	97	100	100	100	100	100	100
Фойє центрального вокзалу	40	78	100	100	100	100	100	100	100
Площа перед вокзалом	78	97	97	97	98	98	100	100	100
Аудиторія	28	65	97	100	100	100	100	100	100
У троллейбусі	52	87	98	100	100	98	98	100	100
Комп'ютер	18	57	93	100	100	100	100	100	100
Кавомолка	-2	15	60	88	92	93	92	92	97
Підземний перехід	83	93	95	100	100	100	100	100	100
Мікрохвильова піч	37	70	97	100	100	100	100	100	100
Пральна машина	47	72	93	100	100	100	100	100	100

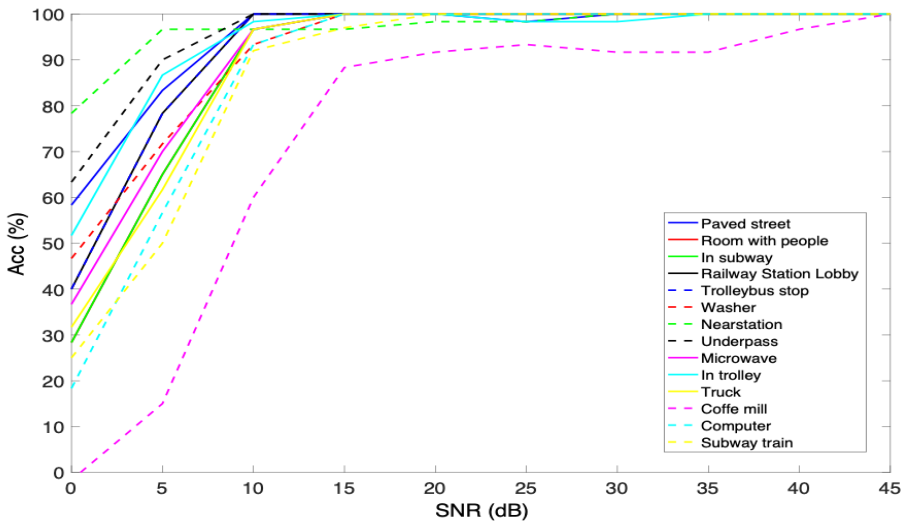
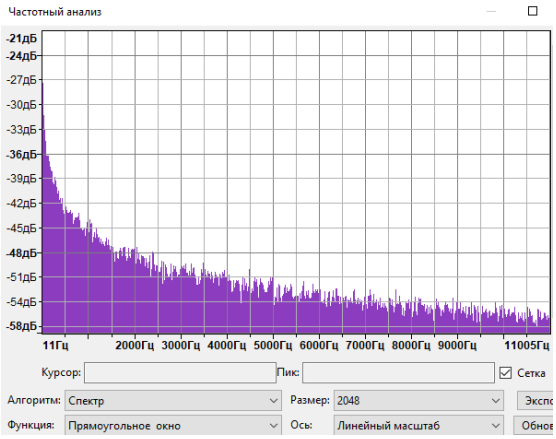


Рис. 3.7. Точність розпізнавання при навчанні методом Multistyle training [8]

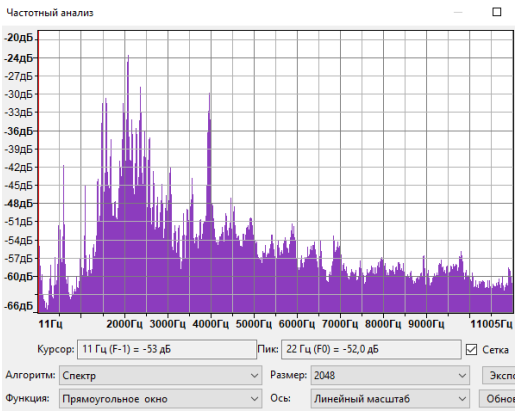
Бачимо, що такий варіант навчання добре підходить для умов, коли співвідношення сигнал-завада сягає або перевищує 10 дБ: тоді забезпечується дійсно висока точність розпізнавання. Перевагою такого виду навчання є

рівномірність результату незалежно від спектру шуму та рівня завади: навіть якщо ми на знаємо реальних умов експлуатації, можемо підготувати систему з достатньо високим рівнем якості розпізнавання. Однак, недоліком є низькі значення точності для мовних сигналів з відношенням сигнал-завада SNR < 10 дБ. З цим можна боротися за допомогою встановлення додаткових систем шумоподавлення або ввести обмеження на умови експлуатації такої системи [9].

Слід відзначити наявність винятку із наведеного вище правила, а саме: шум кавомолки виявився таким, що найліпше маскує мовний сигнал. Із літературних джерел відомо, що потужні маскувальні властивості властиві рожевому шуму, спектр якого спадає із швидкістю 3 дБ/октаву (рис. 3.8.а). Цікаво, що спектр шуму кофемолки (рис. 3.8.б) суттєво відрізняється від рожевого шуму (рис.3.8.а), оскільки містить низку періодичних дискретних сплесків. Обчислення кепстру шуму кавомолки (рис. 3.9) суттєво полегшує кількісний аналіз такої періодичності й дозволяє дійти висновку, що ця періодичність є трехкомпонентною – найбільшим є період 2 кГц (сплески на 2 кГц та 4 кГц), найменшим є період 100 Гц (часті сплески в околиці 2-2,5 кГц), а проміжним є період 500 Гц (це ширина однієї клітини на графіку кепстру).



а



б

Рис.3.8. Спектр рожевого шуму (а), спектр шуму кавомолки (б)

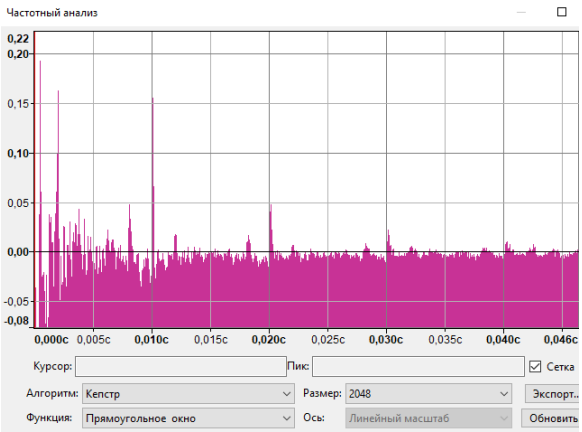


Рис.3.9. Кепстр шуму кавомолки

3.9 Порівняння результатів та вироблення рекомендацій

Підсумовуючи усі результати, включаючи [4], [8], [9], можемо виробити наступні рекомендації для вибору способу навчання системи АРМ.

Навчання на суто чистих сигналах не є ефективним, воно може бути корисним виключно для отримання якісних характеристик системи АРМ.

Навчання методом Fully Matched Training є найменш ефективним з усіх чотирьох, що порівнюються. Найвища точність забезпечується лише при співпадінні значень відношень сигнла-завада навчальної та тестової вибірок. Проте, перевагою такого способу є невелика кількість витрат ресурсів пам’яті системи: зберігається лише інформація для одного певного спектру та рівня шуму. Крім того, суттєво скорочується тривалість навчання системи АРМ

Навчання методом Noise Matced Training є в середньому більш ефективним, ніж методом FMT. Бачимо, що для забезпечення високої точності розпізнавання достатньо забезпечити відношення сигнал-завада вхідного сигналу не менше, ніж 10 дБ, при цьому для 12 з 14 шумів досягається точність *Acc* = 95% і вище, а для одного з шумів достатньо забезпечити SNR = 5 дБ. Такий спосіб є найбільш зручним для ситуації, коли умови експлуатації системи АРМ не є стабільними. Певним недоліком при цьому є ресурсозатратність системи, оскільки обсяг навчальної вибірки значно

більший, а отже і інформації зберігається більше. Крім того, тривалість навчання системи АРМ є помітно більшою, ніж для методу FMT.

Навчання методом SNR-matched training доцільно порівнювати з Fully Matched Training, оскільки для них є спільним основний фактор зниження якості розпізнавання – рівень шумової завади. Вище було сказано, що SNRMT є більш точним за FMT за рахунок більшого обсягу вибірки та ширшого різноманіття спектру шумів. Висока точність розпізнавання забезпечується починаючи від значення SNR = 3 дБ, що перевищує усі наведені в даній роботі результати. Варто зазначити, що за умов високих відношень сигнал-шум (SNR > 30 дБ) система буде втрачати у ефективності. Проте такі значення можуть зустрітись відносно рідко у реальних умовах експлуатації, тому метод SNRMT є одним з кращих за показниками точності, стабільності та обсягу необхідної пам’яті. Очевидним недоліком методу SNRMT є помітно більша, порівняно із методом FMT, тривалість навчання системи АРМ.

Навчання методом Multistyle Training має свої переваги по відношенню до інших: починаючи з певного порогового значення SNR (від 5 до 10 дБ для різних видів шумів) система АРМ забезпечує високу точність без “провалів” при підвищенні значення SNR (як у методах FMT та SNRMT). Також очевидною перевагою є універсальність цього методу: за умов зміни спектру шуму якість розпізнавання не знизиться або знизиться незначним чином. Разом із тим, очевидним недоліком методу MT найбільша, порівняно із іншими методами, тривалість навчання.

3.10 Висновки до розділу 3

При оцінюванні якості системи АРМ можливо використовувати різні міри якості, найпоширенішими з яких є показники *WER* та *Acc*, які є простими і зручними для обчислення. Недоліком цих показників є можливість набути від’ємних значень, якщо у вихідній послідовності слів буде багато вставок. Також використовуються показники *MER* та *WIL*, які позбавлені цього

недоліку, проте останній є складним для обчислення через складність вилучення з системи усієї необхідної інформації.

У системі АРМ, що використовувалась в даній роботі, – The Hidden Markov Model Toolkit [17] – є два показники якості системи: *%Correct* та *%Acc*. У даній роботі використовувався показник *%Acc*, оскільки *%Correct* у випадку великої кількості вставок не відображає дійсну ситуацію.

Було розглянуто 4 варіанти навчання системи АРМ на зашумлених сигналах: Fully-Matched Training, Noise-Matched Training, Signal-to-Noise Ratio Matched Training та Multistyle Training. Показано, що:

- метод Fully Matched Training доцільно застосовувати тільки для ситуації наперед визначених шумових умов, оскільки найвищу ефективність він показує за співпадіння значень SNR навчальної та тестово вибірок, а при неспівпаднні точність розпізнавання суттєво знижується;
- за методом Noise Matched Training для забезпечення високої точності розпізнавання достатньо забезпечити значення відношення сигнал-шум не менше ніж 5 дБ. Цей метод є досить універсальним, оскількі може бути застосовний у системах незалежно від місця іх встановлення та сфери застосування;
- метод SNR-matched training дозволяє розпізнавати мову з високою точністю для сигналів, значення SNR яких перевищує 3дБ, а отже для жорстких умов є оптимальним варіантом, проте система обов’язково має працювати лище в одному типі умов експлуатації з точки зору спектру шуму. Метод потребує забезпечення відносної стабільності умов експлуатації, оскільки для тестових зразків мови зі значеннями SNR, що суттєво відрізняються від навчальних, ефективність роботи системи АРМ знижується.
- метод Multistyle training є універсальним, оскільки надав високу точність розпізнавання і для високих, і для відносно низьких значень сигнал-

завада тестового сигналу. Цей метод доцільно застосовувати за умов зміни спектру або рівня завади шуму;

- вироблено рекомендації по вибору методу навчання системи АРМ залежно від умов експлуатації: спектру шуму, рівня завади та мінливості цих факторів, також до уваги приймається ресурсоємність системи.

4 ПІДВИЩЕННЯ РОБАСТНОСТІ СИСТЕМИ АВТОМАТИЧНОГО РОЗПІЗНАВАННЯ МОВИ ДО ДІЇ РЕВЕРБЕРАЦІЙНОЇ ЗАВАДИ

В даному розділі наведено результати експериментальних досліджень, спрямованих на підвищення робастності системи АРМ до дії реверберaційної завади: визначено ефективність роботи системи за різних варіантів навчання та тестування для сигналів, спотворених реверберaційною завадою.

4.1 Дослідження робастності системи АРМ до дії реверберaційної завади: постановка та проведення експерименту

В даному розділі наведено результати чотирьох експериментів:

1. *Clean training*: навчання системи проводиться на сигналах, не спотворених реверберaційною завадою (SNR = 45 dB, t_{rev} = 0 с), тестування проводиться на таких самих сигналах, та на сигналах, спотворених реверберацією - для визначення початкової ефективності роботи системи.
2. *All-reverb training* (ART): навчальна вибірка складається з сигналів, почергово згорнутих з імпульсними характеристиками приміщень. Тестування проводилось на “чистих” мовленнєвих сигналах та сигналах, спотворених реверберацією. Список та характеристики ІХ наведені нижче.
3. *Reverb-matched training* (RMT): навчальна вибірка складається з сигналів, згорнутих з імпульсним характеристиками з одним певним часом реверберації. Тестування проводилось на чистих сигналах та сигналах, спотворених реверберацією.
4. *Room-training*: досліджувалась ефективність роботи системи у випадку навчання її на сигналах, згорнутих з ІХ приміщення, записаними в різних точках цього приміщення для визначення доцільності

використання додаткових даних для підвищення точності розпізнавання системи АРМ.

Для наочності порівняння методів їх математичний опис також наведено в таблиці 4.1.

Таблиця 4.1. Математичний опис методів проведення експерименту на сигналах, спотворених реверберaційною завадою

Назва методу	*Співпадіння (matching)*
Clean training	$t_t = 0$ с, $t_r = \forall$
Reverb-matched training	$t_t = t_r$
All-reverb training	$t_t = \forall$
Room training	$d = d_n \ \vee \ d = \forall$

Мовні сигнали, що використовуються в даній частині експерименту, такі самі, як і при дослідженні ефективності роботи системи АРМ під дією шумової завади. Це попередньо записане мовлення одного диктора: 200 зразків 10 слів: числа від одного до десяти, кожне слово записане 20 разів з різною, наскільки це можливо, інтонацією. Фонемний словник складається з 22 фонем. Використовувались 39-мірні класифікаційні ознаки типу MFCC_0_D_A (мел-частотні кепстральні коефіцієнти). SNR навчальної вибірки становить 45 дБ. Параметри сигналів: частота дискретизації f_d = 22050 Гц, глибина квантування – 16 біт.

Для створення вибірок сигналів, спотворених реверберацією, використано згортку мовного сигналу з імпульсною характеристикою приміщення за допомогою швидкого перетворення Фур’є.

Імпульсні характеристики, використані у даному дослідженні, отримані з відкритої бібліотеки бінауральних імпульсних характеристик бази даних Рейнсько-Вестфальського технічного університету міста Аахен [85].

Загальна кількість ІХ, використаних в даному експерименті - 62 шт. Це ІХ, записані в різних точках приміщень декілька разів. Час реверберації та тип приміщення обрано таким чином, щоб охопити приміщення, в яких пристрої, оснащені системою АРМ, потенційно будуть використовуватися найчастіше: це навчальна аудиторія, офісне приміщення, кімната для нарад, сходовий майданчик. Відповідно, виокремлено 16 основних ІХ для даного дослідження зі значеннями t_{rev} = 0.3…1 с з кроком 0.1 с (по дві ІХ для кожного значення). В таблиці 4.2 наведено список використаних ІХ із зазначенням часу реверберації та основними характеристиками приміщень, в яких їх записано. Для зручності подальших описів результатів приміщення також пронумеровано. Повний перелік усіх ІХ з зазначенням основних характеристик наведено в додатку Г. В якості міри визначення часу реверберації обрано параметр T20, значення округлено до десятих.

До та після операції згортки усі сигнали було нормалізовано для уникнення впливу відмінностей в рівнях звукової енергії в різних аудіо-фрагментах.

Усього для даного розділу дисертаційної роботи було створено і оброблено 13020 аудіо-файлів.

Таблиця 4.2. Характеристики приміщень

Час реверберації T20, с	Номер, призначення приміщення та об’єм		
0.3	1	Кімната для нарад	малий об’єм
0.4	1	Кімната для нарад	малий об’єм
0.5	2	Офісне приміщення	малий об’єм
0.6	2	Офісне приміщення	малий об’єм
0.7	3	Лекційна аудиторія	великий об’єм
0.8	3	Лекційна аудиторія	великий об’єм
0.9	4	Сходовий майданчик	-
1.0	4	Сходовий майданчик	-

4.2 Результати експериментальних досліджень системи АРМ, що навчалась на чистих сигналах

В даній частині експерименту визначається початкова ефективність роботи системи АРМ. Навчання системи проводиться на мовленнєвих сигналах з SNR = 45 dB, t_{rev} = 0 с, тестування проводиться так само на сигналах з SNR = 45 dB, t_{rev} = 0 с та на сигналах, спотворених реверберацією. Для таких сигналів тестові вибірки містять зразки згортки з імпульсними характеристиками кожного з приміщень (табл.4.2). Отримані результати проілюстровано на рис. 4.1, табличне представлення наведено в додатку Ґ.

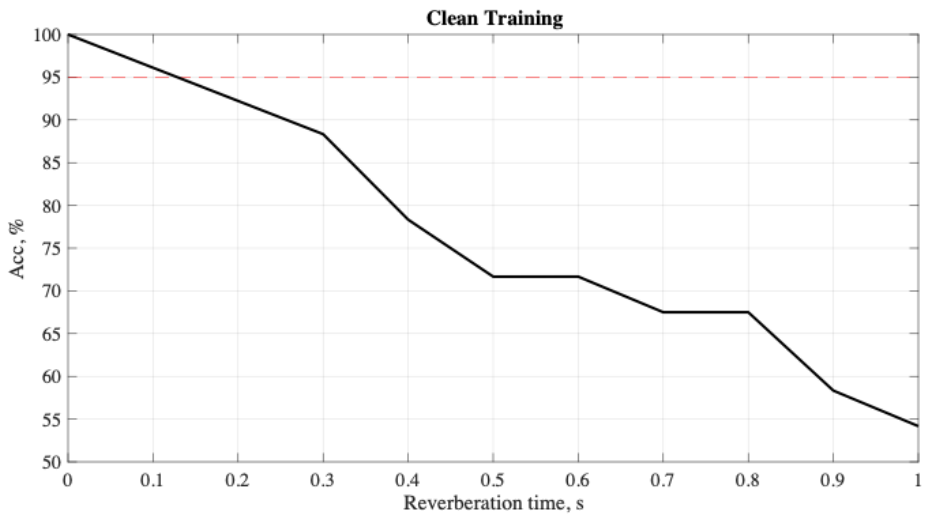


Рис. 4.1. Графік залежності точності розпізнавання від часу реверберації при навчанні на сигналах, не спотворених реверберaційною завадою.

Аналізуючи отримані дані, бачимо цілком передбачуваний спад точності розпізнавання зі зростанням часу реверберації T20. Досягнення високої точності розпізнавання можливе лише для чистих сигналів. При найнижчому з охоплених значень часу реверберації T20 > 0,3 с отримано значення нижче 95%, тоді як для значення T20 = 1 с точність розпізнавання доходить лише до 54 %. За даних умов навчання система АРМ не продукує текст, придатний для подальшого використання, а отже не є робастною до дії реверберaційної завади.

4.3 Результати експериментальних досліджень системи АРМ, що навчалась за методом all-reverb training

Експеримент для методу all-reverb training полягає в тому, що навчальна вибірка складається з сигналів, згорнутих з усіма з IX приміщень №№ 1, 2, 3, 4. Мета цієї частини - встановити, наскільки може підвищитись ефективність роботи системи АРМ, якщо збільшити вибірку зашумлених сигналів, тобто запропонувавши для навчання зімітовані різноманітні реальні умови. Тестування системи АРМ проводилось на тих самих вибірках, що і для попереднього методу.

Результати проілюстровано на рис. 4.2 та наведено в таблиці 4.3. В таблиці 4.3 також вказано різницю точності розпізнавання Acc% між двома методами (Clean training та All-reverb training), що порівнюються.

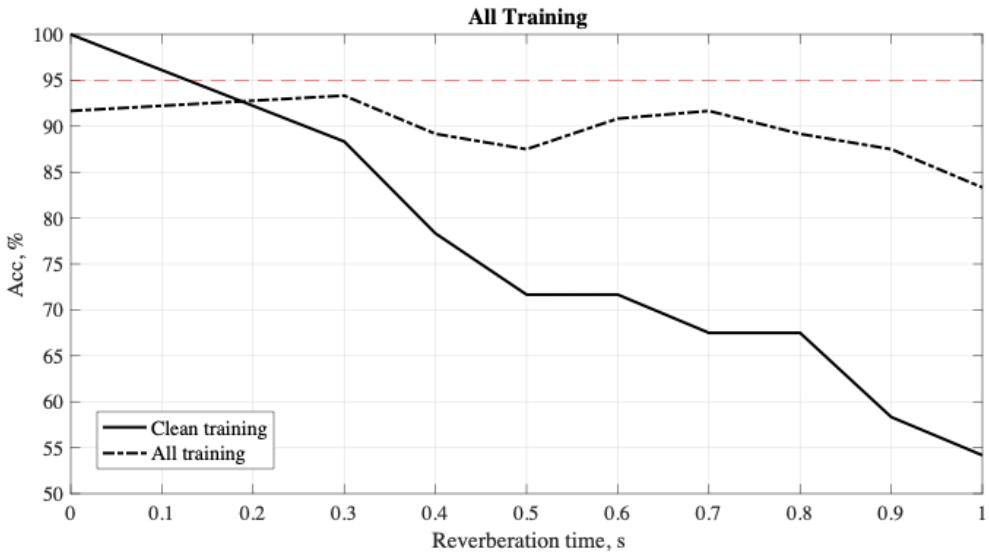


Рис. 4.2. Графік залежності точності розпізнавання від часу реверберації при навчанні за методом all-reverb training.

Оскільки метою даного дослідження є порівняння двох методів, на рис. 4.2. наведено дві криві: суцільною лінією для методу clean training та штриховою для all-reverb training. Можна побачити, що навчання на реверберованих сигналах показує вищу точність розпізнавання, починаючи з

часу реверберації 0,3 с - одного з найменших значень в вибірці. Мінімальне підвищення точності розпізнавання відбулось для значення часу реверберації T20 = 0.3 с (IX кімнати для наради) і становить 5%, максимальне - на 29,17% для значення T20 = 0,9 с (IX зі сходового майданчику). При цьому, для чистих сигналів точність розпізнавання стала нижчою на 8.3%, порівняно з результатами навчання за методом clean training. У випадку тестової вибірки з T20 ≈ 0 с це пояснюється тим, що при навчанні система отримувала сигнали з додатковими спектральними компонентами, які відсутні в даній тестовій вибірці. Оскільки такі умови можна отримати лише в акустично-заглушеній кімнаті, для переважної більшості випадків експлуатації цей результат не є суттєвим. Втім, незважаючи на суттєве зростання точності розпізнавання, за методом all-reverb training не вдається досягти високої точності розпізнавання (95%), результати для більшості тестувальних вибірок знаходяться в межах 84…94%, а для значення t_{rev} = 1 с значення точності Acc% = 83%.

Таблиця 4.3. Порівняння точності розпізнавання системи АРМ при навчанні за методами clean training та all training

t_{rev}, с	0	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
Acc_{clean}, **%**	100	88.34	78.34	71.67	71.67	67.5	67.5	58.34	54.17
Acc_{all}, **%**	91.67	93.34	89.17	87.5	90.83	91.67	89.17	87.5	83.33
Різниця, **%**	-8.33	**5**	10.84	15.84	19.16	24.17	21.67	**29.17**	29.16

4.4 Результати експериментальних досліджень системи АРМ, що навчалась за методом reverb-matched training

В цій частині експерименту навчальні та тестувальні вибірки складаються з сигналів, згорнутих з імпульсною характеристикою з певним часом реверберації з діапазону $t_{rev} = 0{,}3...1$ с з кроком 0,1 с. Навчання і,

відповідно, тестування системи АРМ було проведено для вибірок сигналів, згорнутих з кожною з ряду імпульсних характеристик із зазначеним вище часом реверберації. Дана частина експериментального дослідження має на меті визначити, чи достатньо навчати систему лише на тих умовах, за яких її будуть експлуатувати, щоб зменшити обсяг вибірки і знизити ресурси пам'яті та часові.

Результати даної частини експерименту наведено на рис. 4.3 та 4.4. Табличне представлення даних наведено в додатку Ґ.

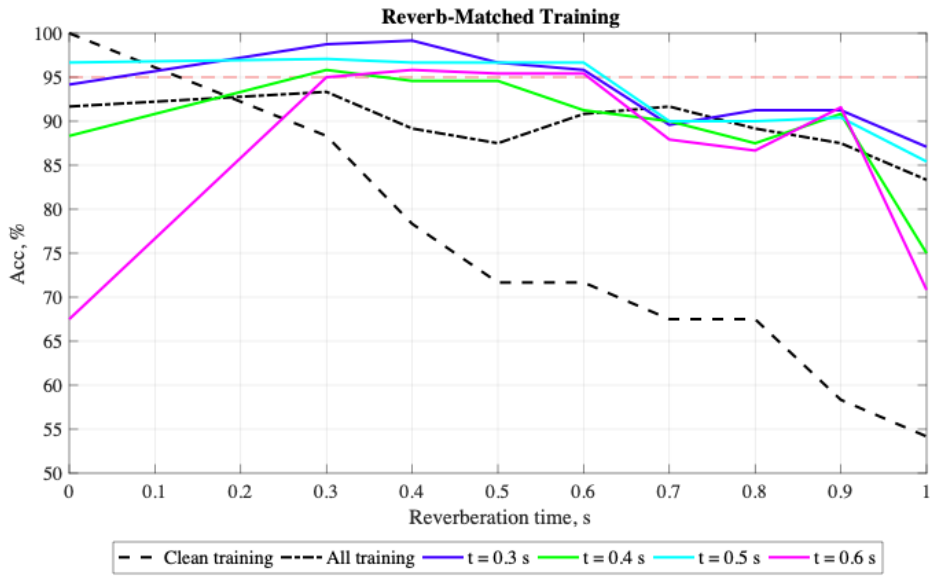


Рис. 4.3. Графік залежності точності розпізнавання від часу реверберації при навчанні за методом reverb-matched training для приміщень малого об'єму

Для зручності сприйняття графіки згруповано відповідно до часу реверберації в приміщеннях: на рис. 4.3 - приміщення малого об'єму (офісне та мітинг-рум, $t_{rev} \leq$ 0,6 с), на рис. 4.4 - приміщення великого об'єму (лекційна аудиторія) та сходовий майданчик ($t_{rev} \geq$ 0,7 с). На кожному з рисунків для порівняння методів наведено також результати попередніх дослідів: штриховою лінією для навчання за методом clean training та штрих-пунктирною для навчання за методом all-reverb training відповідно.

На рисунку 4.3 бачимо значно вищу точність розпізнавання, аніж у попередніх двох методах. За даним методом навчання вдалось досягти точності розпізнавання Acc% = 95% для тестувальних сигналів з часом реверберації від 0,3 с до 0,6 с. Найкращий результат отримано для навчання на вибірці зі значенням часу реверберації IX T20 = 0,3 с. Для інших значень часу реверберації T20 з приміщень малого об’єму також отримано суттєве зростання точності розпізнавання.

Загалом, показано, що в умовах, коли для навчання використовуються мовленнєві фрагменти, спотворені реверберацією, бачимо, що найкращу ефективність система АРМ показує, коли час реверберації навчальної та тестової вибірок співпадають або близькі за значеннями. Отже, для даного діапазону значень питання підвищення робастності системи АРМ до дії реверберaційної завади можна вважати вирішеним за допомогою зміни стилю навчання системи.

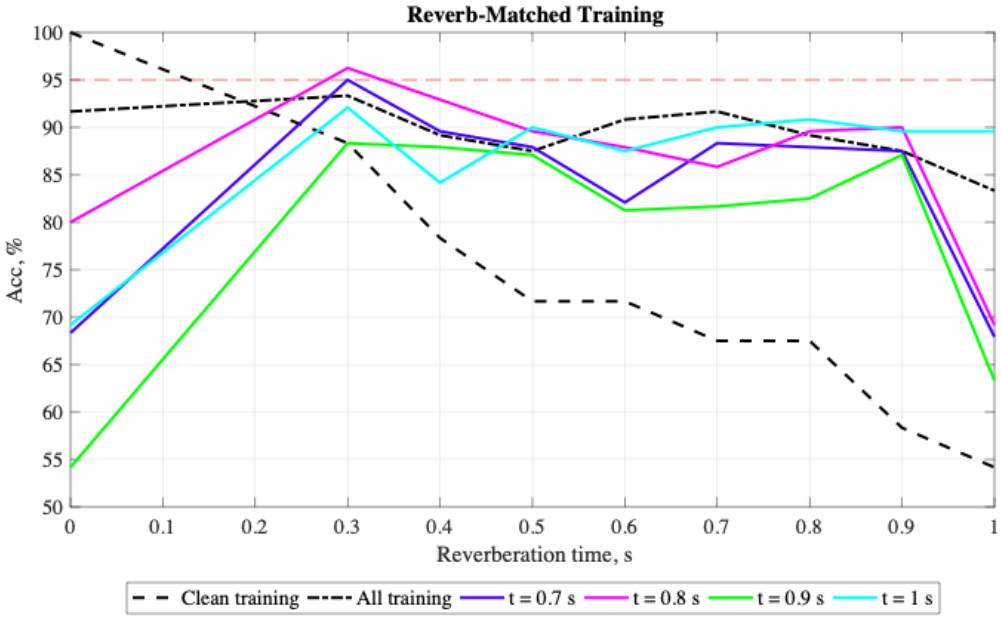


Рис. 4.4. Графік залежності точності розпізнавання від часу реверберації при навчанні за методом reverb-matched training для приміщення великого об’єму та сходового майданчику.

На рис. 4.4 наведено результати за навчання системи на сигналах, згорнутих з ІХ з часом реверберації від 0,7 с до 1 с з кроком 0,1. Тут також бачимо загальну тенденцію до підвищення ефективності роботи системи АРМ, порівняно з clean training, проте її не достатньо для забезпечення високої точності розпізнавання. При тестуванні на сигналах з t_{rev}= 0,3 с для часу реверберації навчальної вибірки t_{rev} = 0,7 та t_{rev} = 0,8 с маємо задовільний результат Acc% = 95%, проте для всіх інших випадків більш прийнятним є метод all-reverb training. Він не забезпечує необхідну точність, проте показує більш стабільний результат в межах 84…94%, а отже є більш передбачуваним. Виключенням є результат для навчальної вибірки з t_{rev} = 1 с, який дозволив отримати точність розпізнавання 89% для тестування на t_{rev} = 1 с проти 84% при навчанні за методом all-training.

Для випадків, за яких система АРМ буде експлуатуватись в приміщеннях великого об’єму або зі значеннями часу реверберації в приміщенні $t_{rev} \geq$ 0,7 с варто оснащувати систему пристроями придушення реверберації або встановлювати норми по близькості диктора до мікрофону, аби знизити вплив реверберaційної компоненти на якість роботи системи АРМ.

4.5 Результати експериментальних досліджень системи АРМ, що навчалась за методом room training

Дана частина експерименту має на меті перевірити, чи підвищиться точність розпізнавання системи АРМ, що працює в певному приміщенні, якщо навчати її на вибірці, що містить мовленнєві сигнали, згорнуті з ІХ, отриманими в декількох точках цього приміщення. Це допоможе визначитись, чи доцільно збільшувати обсяг вибірки, чи достатньо даних з однієї точки. Такі дані будуть актуальними у випадку, коли системою АРМ буде обладнаний прилад чи пристрій, що його передбачено використовувати в одних і тих самих заздалегідь визначених умовах. Тож, було проведено навчання системи АРМ на вибірках, де використано ІХ з усіх точок кожного з

приміщень №№ 1-4, а також навчання на вибірках, що містять сигнали, згорнуті з ІХ в окремих точках приміщень.

Для даної частини експерименту було використано усі наявні записи ІХ відповідно до таблиці у додатку Г. В кожній точці приміщень малого об'єму було зроблено 4 записи, великого об'єму та на сходовому майданчику - 2 записи. З них створено вибірки для навчання та тестування, результати усереднено.

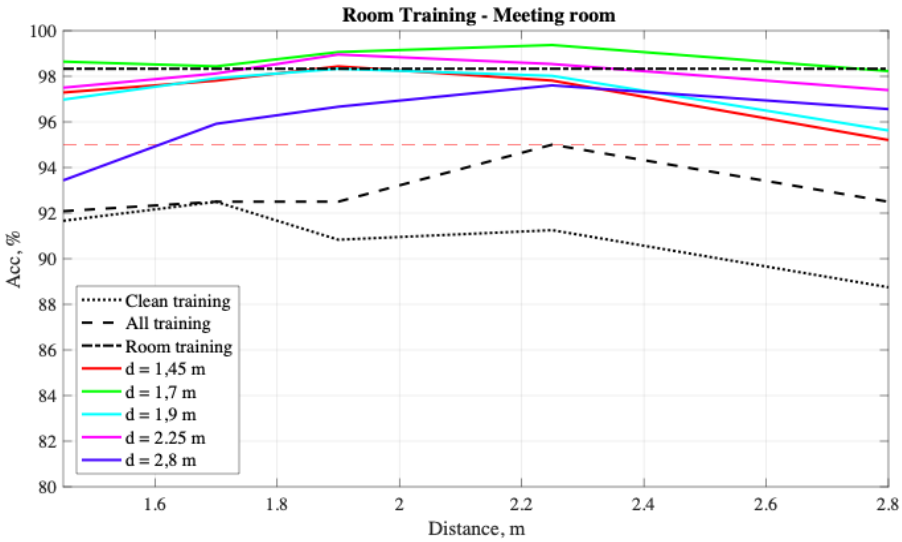


Рис. 4.5. Графік залежності точності розпізнавання від часу реверберації при навчанні за методом room-training (кімната для нарад)

Результати залежності точності розпізнавання *Acc%* від відстані між гучномовцем та мікрофоном для навчання за методом Room training наведено на рис. 4.5, 4.6, 4.7, 4.8 (кімната для нарад, офісне приміщення, сходовий майданчик, лекційна аудиторія відповідно). На кожному з графіків також наведено результати дослідження за методами clean training (лінія крапочками) та all-reverb training (штриховою лінією) для зручності порівняння.

На рис. 4.5 проілюстровано залежність точності розпізнавання від відстані між джерелом звуку та мікрофоном для приміщення №1 - кімната для нарад. По осі *x* відкладено відстань для тренувальних вибірок. Суцільні лінії

відповідають результатам навчання для вибірок з відстанню між джерелом та мікрофоном 1,45 м, 1,7 м, 1,9 м, 2,25 м, та 2,8 м. Штрих-пунктирна лінія ілюструє результат для методу room-training, за яким було сформовано велику навчальну вибірку, що містила дані з усіх точок. Отримані результати свідчать, що метод room-training для даного типу приміщення суттєво підвищує точність розпізнавання. Результат за методом room-training є рівномірним, на всіх відстанях забезпечується точність Acc% - 98,33%. Для усіх навчальних вибірок вдалось досягти точності Acc% > 95%, окрім єдиного значення - при навчанні на вибірці з відстанню між джерелом звуку та мікрофоном d = 2,8 м на меншій відстані тестування d = 1,4 м маємо результат 93%. Це пов'язано з тим, що система при навчанні отримала сигнали з більшим впливом реверберaційної компоненти, ніж при тестуванні, що збільшило кількість помилок. Схожу тенденцію можна спостерігати і для інших приміщень (рис. 4.6, 4.7, 4.8).

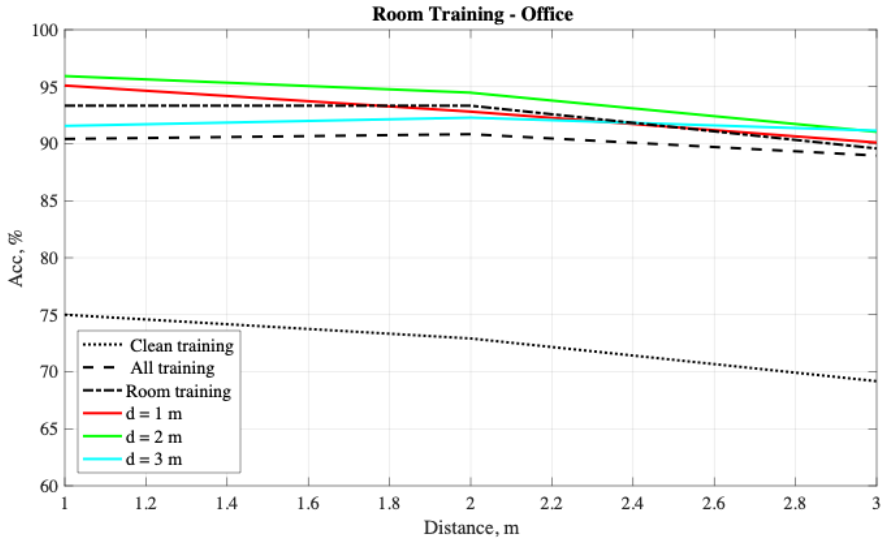


Рис. 4.6. Графік залежності точності розпізнавання від часу реверберації при навчанні за методом room-training (офісне приміщення)

Результати для приміщення №2 малого об'єму наведено на рис. 4.6. Тут бачимо, що результат навчання за методом room training являє собою дещо "середнє" відносно значень результатів для навчання на окремих вибірках.

Точність розпізнавання Acc% лежить в діапазоні 90…96%, що є значно кращим результатом за результат навчання методом clean training. Також даний метод є на 2-4% ефективнішим за метод all-reverb training, але при цьому обсяг вибірки для навчання є значно меншим, оскільки містить лише дані одного приміщення, а отже з точки зору ресурсів більш прийнятним. Схожий висновок можна зробити з рис. 4.7 - сходовий майданчик. За цими даними також можна спостерігати тенденцію до зниження точності зі збільшенням відстані між джерелом звуку та мікрофоном для навчання за методом all-training, при цьому результати за методом room-training залишаються стабільними.

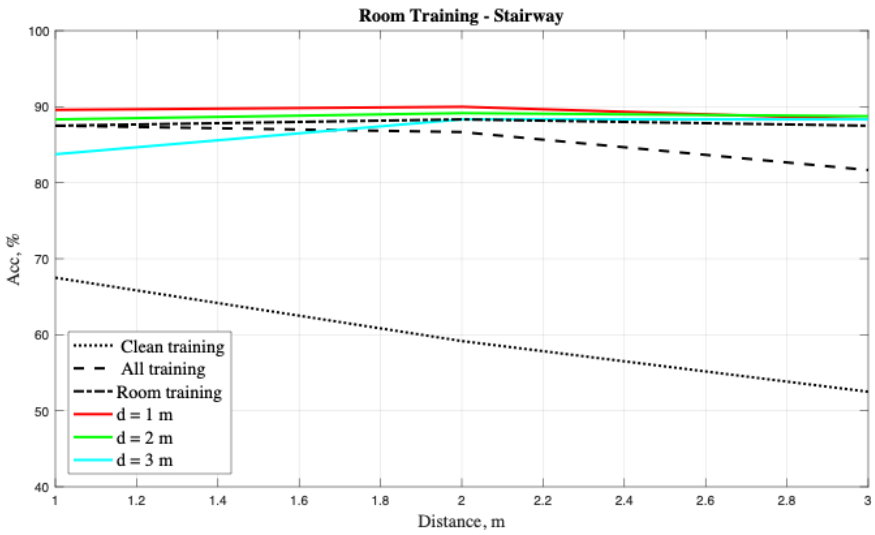


Рис. 4.7. Графік залежності точності розпізнавання від часу реверберації при навчанні за методом room-training (сходовий майданчик)

Дещо відрізняються результати для лекційної аудиторії - приміщення великого об’єму. В даному випадку точність розпізнавання для навчання в окремих точках (окрім d = 10.2 м) та за методом all-training є вищою, ніж за навчання методом room training та знаходиться в межах 89 - 95%, в той час як точність розпізнавання власне за методом room-training коливається в межах 85-85%. Втім, варто відмітити, що порівняно з іншими результатами крива

методу room-training є більш гладкою, тобто результат є більш стабільним та передбачуваним. Це зауваження справедливе для усіх 4х приміщень, що аналізуються.

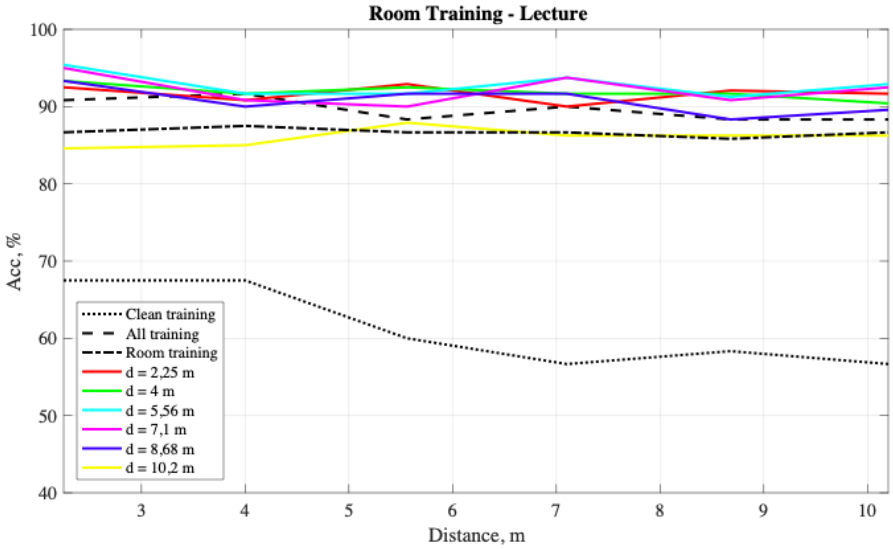


Рис. 4.8. Графік залежності точності розпізнавання від часу реверберації при навчанні за методом room-training (лекційна аудиторія)

4.6 Порівняння результатів та вироблення рекомендацій

За отриманими результатами можемо виробити наступні рекомендації для вибору способу навчання системи АРМ, більш стійкої до дії реверберaційної завади.

Навчання на чистих сигналах (за методом clean training) не є доцільним для вирішення прикладних задач, а може бути застосованим лише для визначення початкових якісних характеристик системи АРМ.

Навчання за методом all-training показало вищу ефективність за навчання на чистих сигналах, проте не є наочною залежність ефективності роботи системи АРМ від часу реверберації, а отже важко передбачити стабільний результат в разі, якщо реверберaційні умови будуть змінюватись (при переміщенні пристрою, оснащеного системою АРМ або при зміні характеристик приміщення). Також навчальна вибірка має містити значні

обсяги даних, що є затратним з точки зору потужності комп'ютера або мікропроцесора, а також ресурсів фінансових та часових. Втім, певною перевагою є те, що метод універсальний і не потребує додаткових налаштувань при зміні умов, а отже за власне мінливих умов може бути прийнятним з урахуванням вищенаведених недоліків (оснащення входу системи АРМ пристроєм придушення реверберації / закладення додаткових годин роботи транскрайберів тощо).

Навчання за методом Reverb-Matched Training є досить ефективним для приміщень малого об'єму або для приміщень з часом реверберації $t_{rev} \leq 0,6$ с. Для таких сигналів вдалось досягти високої точності розпізнавання (~95%). Маємо найвищу ефективність роботи системи АРМ при співпадінні або наближеності значень часу реверберації навчальної та тестувальної вибірок. На зазначеному проміжку значень метод Reverb-Matched Training вирішує завдання підвищення робастності системи АРМ до дії реверберaційної завади. Втім, для приміщення великого об'єму або з часом реверберації $t_{rev} \geq 0,7$ с даний метод поступається методу all-training, з єдиним виключенням для значення часу реверберації $t_{rev} = 1$ с. Перевагою даного методу є невеликий обсяг навчальної вибірки, а отже відносно малі затрати часу та ресурсів на її створення, та висока точність розпізнавання на певному діапазоні значень t_{rev}. До недоліків відноситься обмеженість умов, в яких доцільно застосовувати таким чином змодельовану систему АРМ та неуніверсальність такої системи.

Навчання за методом Room-Training доцільно застосовувати, коли заздалегідь відомо конкретне приміщення, в якому буде обладнано пристрій системою АРМ, та всі його параметри, що впливають на час реверберації. Для приміщень малого об'єму він показав високу ефективність, для усіх приміщень - практично незалежність результату від відстані між джерелом мовленнєвого сигналу та входом системи АРМ, а отже, від впливу реверберaційної компоненти. Це дає підстави вважати даний метод доцільним для портативних пристроїв, які можна вільно переміщувати в приміщенні

залежно від задач. Разом з тим слід врахувати, що для високих значень часу реверберації слід так само, як і для методу all-training, врахувати необхідність залучення додаткових системних елементів для придушення реверберації або додатковий час на роботу транскрайбера. Також перевагою є відносно невеликий обсяг навчальної вибірки, а недоліком - обмеженість умов використання.

4.7 Висновки до розділу 4

1. Проведено дослідження способів підвищення робастності системи АРМ до дії реверберaційної завади. Показано, що навчання системи на чистих сигналах не є ефективним, а його результати можна використовувати лише як початкову характеристику при моделюванні системи АРМ.

2. Проаналізовано три методи навчання системи АРМ, за яких навчальні вибірки містять мовленнєві сигнали, спотворені реверберацією: all-reverb training, reverb-matched training, room-training. Проведено порівняння з точки зору ефективності роботи системи, витрат ресурсів по обчислювальних потужностях, часу та фінансовій складові. Показано, що для заздалегідь визначених умов експлуатації доцільно використовувати методи room-training або reverb-matched training для певного діапазону значень часу реверберації в приміщенні. Також показано, що універсальним методом є all-reverb training, проте його варто використовувати в комбінації з придушенням реверберації до певного рівня або для певного ряду задаз з урахуванням додаткового робочого часу, необхідного на вдосконалення результатів.

ВИСНОВКИ

У дисертаційній роботі представлено результати дослідження шляхів підвищення робастності системи АРМ до дії завад, таких як шум та реверберація. Крім того, представлено результати досліджень впливу лінійних та нелінійних спотворень сигналів на якість та розбірливість мовлення. Виконано огляд досліджень за темою, проведено комп'ютерне моделювання системи АРМ та експериментальні дослідження. Отримано наступні результати:

1. Виконано огляд існуючих підходів до підвищення точності розпізнавання систем АРМ на тлі дії шумової та реверберaційної завад. Описано метод моделювання систем АРМ, що базується на використанні прихованих марковських моделей, для задачі розпізнавання ізольованих слів.
2. Детально розглянуто методи адаптації систем АРМ до дії завад, на основі чого обгрунтовано необхідність їх вдосконалення та розширення обсягу наявних експериментальних досліджень.
3. Вперше для реальних мовленнєвих сигналів отримано кількісні оцінки ступеня підвищення точності розпізнавання мовлення, спотвореного шумом різної природи та інтенсивності, шляхом навчання системи автоматичного розпізнавання на спотворених шумом мовних сигналах.
4. Проведено порівняння чотирьох стилей навчання системи АРМ на сигналах, спотворених шумовою завадою: Fully-Matched Training, Noise-Matched Training, Signal-to-Noise Ratio Matched Training та Multistyle Training. Визначено, що навчання на зашумлених сигналах дозволяє практично гарантовано досягти точності розпізнавання 95% за наступних умов:
   - за методом Fully-Matched Training - при забезпеченні відношення сигнал-шум тестового сигналу SNR від 10 до 25 дБ, в той час як при

навчанні на «чистих сигналах» необхідне значення SNR коливається від 20 до 40 дБ;

- за методом Noise-Matched Training - при забезпеченні відношення сигнал-шум тестового сигналу SNR > 10 дБ;
- за методом Signal-to-Noise Ratio Matched Training - при значеннях SNR > 5 дБ, що є вищим, ніж за використання Fully-Matched Training;
- за методом Multistyle Training - при значеннях SNR ≥ 10 дБ.

5. Вперше для реальних мовленнєвих сигналів отримано кількісні оцінки ступеня підвищення точності розпізнавання мовлення, спотвореного реверберацією, шляхом навчання системи автоматичного розпізнавання на спотворених реверберацією мовних сигналах.
6. Проведено порівняння трьох стилей навчання системи АРМ на сигналах, спотворених реверберaційною завадою: All-Reverb Training, Reverb-Matched Training, Room-Training. Визначено, що навчання на реверберованих сигналах дозволяє практично гарантовано підвищити точність розпізнавання за наступних умов:
   - За методом All-Reverb Training точність розпізнавання можна підвищити на 5-29% порівняно з навчанням на чистих сигналах;
   - За методом Reverb-Matched Training можна досягти точності розпізнавання 95% в приміщеннях малого об’єму для значень часу ревеберації до T20 ≤ 0,6 с при співпадінні або схожості навчальних та тестувальних умов;
   - За методом Room Training точність розпізнавання може сягати 85-88% для приміщень великого об’єму та 89 - 98% для приміщень малого об’єму, при цьому майже відсутня залежність від відстані між гучномовцем та мікрофоном (коливання в межах 3%), що дозволяє отримати передбачуваний та стабільний результат при зменшеному обсязі навчальної вибірки.

7. Вдосконалено метод оцінювання розбірливості мовлення непрямим методом, із використанням міри якості сигналів у вигляді барківського спектрального спотворення, при цьому показано високу кореляцію показників BSD та STI для приміщень усіх розмірів, на противагу показникам FWSNR, придатним лише для приміщень середнього та великого розміру, та PESQ, придатним лише для великих приміщень.
8. Уточнено висновки щодо залежності розбірливості мовлення від щільності відбить звуку та часу реверберації, із використанням двох імовірнісних моделей імпульсних характеристик приміщень: моделі з одиничним відбиттям в інтервалі часу 0-50 мс та моделі, в якій вібиття розподіляються випадково за рівномірним законом на інтервалі часу 0-50 мс. Для першої моделі отримано залежності показника STI від затримки та сили відбитого сигналу. Для другої моделі показано, що найвищі значення STI можна отримати за умови зосередження найсильніших відбиттів на початку часового інтервалу 0-50мс.
9. Вдосконалено спосіб об’єктивного виявлення ефекту кліпування мовленнєвих сигналів та об’єктивного оцінювання якості мовленнєвих сигналів, спотворених кліпуванням, що базується на використанні коефіцієнта ексцесу та його функціональних перетворень як мір спотворення сигналів. Показано, що квадратний корінь оберненого коефіцієнту ексцесу є більш зручним для практичного використання, ніж коефіцієнт ексцесу, оскільки є лінійно пов’язаним із суб’єктивною оцінкою якості сигналу.

Практичне значення отриманих результатів полягає у встановленні умов досягнення високої точності розпізнавання в системах автоматичного розпізнавання мовленнєвих сигналів, спотворених шумами різної природи та інтенсивності, за наявності різної апріорної інформації щодо відношення сигнал-шум, що дозволяє забезпечити робастність системи автоматичного розпізнавання шляхом використання відносно простих правил її

налаштування; встановленні умов досягнення високої точності розпізнавання в системах автоматичного розпізнавання мовленнєвих сигналів, спотворених реверберацією приміщень із різним часом реверберації, за наявності різної апріорної інформації щодо часу реверберації приміщення, що дозволяє забезпечити робастність системи автоматичного розпізнавання шляхом використання певних правил її налаштування; встановленні працездатності та ефективності оцінювання розбірливості мовлення непрямим методом, із використанням міри якості сигналів у вигляді барківського спектрального спотворення, що дозволяє оцінювати розбірливість мовлення, спотвореного реверберацією, за наявності еталонного неспотвореного сигналу; визначенні залежності розбірливості мовлення від щільності ранніх відбить звуку та часу реверберації, із використанням імовірнісних моделей імпульсних характеристик приміщень, що дозволяє обґрунтувати експериментальні результати оцінювання розбірливості мовлення в різних точках приміщення; встановленні можливості автоматизації виявлення кліпування та об’єктивного оцінювання якості мовленнєвих сигналів, спотворених кліпуванням.

Результати, представлені в дисертації, можуть бути використані при розробці та експлуатації систем автоматичного розпізнавання мовлення. На основі проведених експериментальних досліджень сформовано рекомендації стосовно доцільності використання різних способів навчання системи АРМ на сигналах, спотворених шумовою або реверберaційною завадою в залежності від наявності або відсутності попередньої інформації про потенційні шумові або реверберaційні умови експулатації системи АРМ; врахування наданих рекомендацій дозволить підвищити точність розпізнавання систем АРМ, не залучаючи додаткових часових, обчислювальних та фінансових ресурсів.

Отримані результати вже застосовано в освітньому процесі кафедри акустичних та мультимедійних електронних систем за спеціальністю 171 Електроніка, зокрема в освітні науковій програмі «Електроніка», а також в освітній професійній програмі “Акустичні електронні системи та технології

обробки акустичної інформації” Національного технічного університету України “Київський політехнічний інститут імені Ігоря Сікорського”.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Prodeus A., Didkovska M., Kukharicheva K. Comparison of Speech Quality and Intelligibility Assessments in University Classrooms. International Journal of Architectural Engineering Technology. 2021. Vol. 8. P. 52–60. https://doi.org/10.15377/2409-9821.2021.08.5
2. Prodeus A., Didkovska M., Kukharicheva K, Motorniuk D. Two Simplified Models Of Early Sound Reflections In a Room // Electronics and Control Systems. 2020. Vol. 3, (65). https://doi.org/10.18372/1990-5548.65.14991
3. Prodeus A., Kotvytskyi I., Didkovska M., Kukharicheva K. Kurtosis and Normalized Variance as Measures of Speech Signals Clipping Value. Electronics and Control Systems, 2019. Vol. 4 (62). https://doi.org/10.18372/1990-5548.62.14378
4. Prodeus A., Kukharicheva K. Accuracy of Automatic Speech Recognition System Trained on Noised Speech. Electronics and Control Systems. 2016. Vol. 3 (49). https://doi.org/10.18372/1990-5548.49.11230
5. Prodeus A., Didkovska M., Kukharicheva K. Impact of University Classroom Size on the Relationship Between Speech Quality and Intelligibility. International Journal of Computing. 2022. Vol. 21 (3). https://doi.org/10.47839/ijc.21.3.2690
6. Prodeus A., Didkovska M., Kukharicheva K., Motorniuk D. Modeling the Influence of Early Sound Reflections on Speech Intelligibility // Proceedings of 2020 IEEE 6th International Conference on Methods and Systems of Navigation and Motion Control (MSNMC), Kyiv, 20-23 Oct. 2020. Kyiv, Ukraine. https://doi.org/10.1109/MSNMC50359.2020.9255657
7. Prodeus A., Kotvytskyi I., Didkovska M., Kukharicheva K. Kurtosis and Its Transformations as Objective Measures of Clipping Value and Speech Quality // Proceedings of IEEE 5th International Conference on Actual Problems of

Unmanned Aerial Vehicles Developments (APUAVD), Kyiv, 22-24 October 2019. Kyiv, Ukraine. https://doi.org/10.1109/APUAVD47061.2019.8943880

8. Prodeus A., Kukharicheva K. Automatic Speech Recognition Performance for Training on Noised Speech // Proceedings of 2017 IEEE 2nd International Conference on Advanced Information and Communication Technologies (AICT), Lviv, 4-7 July 2017. Lviv, Ukraine. https://doi.org/10.1109/AIACT.2017.8020068
9. Prodeus A., Kukharicheva K. Training of automatic speech recognition system on noised speech // Proceedings of 2016 IEEE 4th Int. Conf. "Methods and Systems of Navigation and Motion Control (MSNMC)», Kyiv, October 18-20, 2016. Kyiv, Ukraine. P. 221-223. https://doi.org/10.1109/MSNMC.2016.7783147
10. А. М. Продеус, І. В. Котвицький, М. В. Дідковська, В. С. Дідковський, К. А. Кухарічева, Д. Є. Моторнюк, О. О. Дворник Спосіб виявлення кліпування мовного та музичного сигналів // Патент UA 144291 U, МПК G01R 23/20, опубл. 25.09.2020.
11. Продеус А.Н., Дидковский В.С., Дидковская М.В. Акустическая экспертиза и коррекция коммуникационных каналов. Киев: Lambert Academic Publishing, 2017.
12. Namrata D. Feature extraction methods LPC, PLP, and MFCC in speech recognition // International Journal For Advance Research in Engineering And Technology (ISSN 2320-6802). 2013. Vol.1.
13. Brunet K., Taam K., Cherrier E., Faye N., Rosenberger C. Speaker Recognition for Mobile User Authentication: An Android Solution. *8ème Conférence sur la Sécurité des Architectures Réseaux et Systèmes d'Information (SAR SSI)*. 2013. URL: https://www.researchgate.net/publication/257365356_Speaker_Recognition_for_Mobile_User_Authentication_An_Android_Solution

14. Vachhani B.B., Patil H.A. Use of PLP Cepstral Features for Phonetic Segmentation. *2013 International Conference on Asian Language Processing*. Urumqi, China. 2013. P.143-146.
15. Hermansky H. Perceptual linear predictive (PLP) analysis of speech. Journal of the Acoustical Society of America. 1990. Apr, Vol.87(4):1738-52.
16. Trabelsi I., Ben Ayed D. On the Use of Different Feature Extraction Methods for Linear and Non Linear kernels // 2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT). Sousse, Tunisia, 21-24 March 2012.
17. The HTK Book (for HTK Version 3.4) / S. Young et al. Cambridge: Cambridge University Engineering Department, 2006. 359 p.
18. Кухарічева К. А. Системи автоматичного розпізнавання зашумленої мови : магістерська дисертація на здобуття ступеня магістра : 171. Київ, 2018. 63 с.
19. Baum L. E., Eagon J.A. An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bulletin of the American Mathematical Society*. 1967. № 73. С. 360–363.
20. Virtanen T., Singh R., Raj B. Techniques for Noise Robustness in Automatic Speech Recognition, First Edition. John Wiley & Sons, Ltd., 2013. 496 p.
21. Kanter I. The Baum–Welch algorithm for estimating a Hidden Markov Model. URL: https://web.archive.org/web/20070929091507/http://www.ph.biu.ac.il/faculty/kanter/BW.pdf
22. Forney Jr. G.D. The Viterbi Algorithm: A Personal History. *Viterbi Conference*. University of Southern California, Los Angeles, USA. March 8, 2005. URL: https://doi.org/10.48550/arXiv.cs/0504020
23. Deep Neural Networks for Acoustic Modeling in Speech Recognition. Four research groups share their views. *IEEE Signal Processing Magazine*. Nov 2012. URL: https://www.cs.toronto.edu/~hinton/absps/DNN-2012-proof.pdf

24. Jin J. Liu P., Zhao G. Speech recognition with DNN-LAS. URL: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2761080.pdf.
25. Fohr D., Mella O., Illina I. New Paradigm in Speech Recognition: Deep Neural Networks. *IEEE International Conference on Information Systems and Economic Intelligence*. Apr 2017. URL: https://hal.science/hal-01484447
26. Kaur G., Shirastava M., Kumar A. Speaker and Speech Recognition using Deep Neural Network. *International Journal of Emerging Research in Management & Technology*. 2018. P. 118 – 123. URL: https://www.researchgate.net/publication/326046387_Speaker_and_Speech_Recognition_using_Deep_Neural_Network
27. Pan J., Liu C., Wang Z., Hu Y., Jiang H. Investigation of deep neural networks (DNN) for large vocabulary continuous speech recognition: Why DNN surpasses GMMS in acoustic modeling. *8th International Symposium on Chinese Spoken Language Processing*. 2012. P. 301-305. URL: https://doi.org/10.1109/ISCSLP.2012.6423452.
28. Palaz D., Magimai.-Doss M., Collobert R. Convolutional Neural Networks-based continuous speech recognition using raw speech signal. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. South Brisbane, QLD, Australia, 2015. P. 4295-4299. URL: https://doi.org/10.1109/ICASSP.2015.7178781.
29. Abdel-Hamid O., Mohamed A.-r., Jiang H., Deng L., Penn G., Yu D. Convolutional Neural Networks for Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2014. Vol. 22, No 10. P. 1533-1545.
30. Graves A., Mohamed A., Hinton G. Speech Recognition With Deep Recurrent Neural Networks. *38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. URL: https://doi.org/10.48550/arXiv.1303.5778

31. Peddinti, V., Povey, D., Khudanpur, S. A time delay neural network architecture for efficient modeling of long temporal contexts. *Interspeech 2015*. 2015. P. 3214-3218. https://doi.org/10.21437/Interspeech.2015-647
32. Ko, T., Peddinti, V., Povey, D., Khudanpur, S. Audio augmentation for speech recognition. *Interspeech 2015*. 2015. P. 3586-3589. https://doi.org/10.21437/Interspeech.2015-711
33. Swietojanski P., Ghoshal A., Renals S. Revisiting Hybrid and GMM-HMM system combination techniques. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing :* Proceedings. 2013.
34. Wang H., Ragni A., Gales M.J., Knill K., Woodland P.C., Zhang C. Joint decoding of tandem and hybrid systems for improved keyword spotting on low resource languages. *Interspeech.* 2015. https://doi.org/10.21437/Interspeech.2015-726
35. Li J. Recent Advances in End-to-End Automatic Speech Recognition. *APSIPA Transactions on Signal and Information Processing.* 2021. https://doi.org/10.48550/arXiv.2111.01690
36. Hadian H., Sameti H., Povey D., Khudanpur S. End-to-end Speech Recognition Using Lattice-free MMI. *Interspeech.* 2018. https://doi.org/10.21437/Interspeech.2018-1423
37. Ладошко О. М. Підвищення робастності систем автоматичного розпізнавання мовлення методами обробки сигналів : дис. канд. техн. наук : 05.09.08 / Ладошко Ольга Миколаївна – Київ, 2016. – 185 с.
38. Zhu Q., Zhang J., Zhang Z., Dai L. Joint Training of Speech Enhancement and Self-supervised Model for Noise-robust ASR. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2022. https://doi.org/10.48550/arXiv.2205.13293
39. Leggetter J., Woodland P.C. Maximum likelihood linear regression for speaker adaptation of HMMs. *Computer Speech and Language.* 1995. Vol. 9. P. 171–186.

40. Font M.F. Maximum-likelihood linear regression coefficients as features for speaker recognition // Thesis. Paris, Nov 2009. URL: https://theses.hal.science/tel-00616673/document
41. Ganitkevitch J. Speaker Adaptation using Maximum Likelihood Linear Regression // Seminar Automatic Speech Recognition, Rheinisch-Westf˙alische Technische Hochschule Aachen. Aachen, Germany, 2005. URL: https://www.cs.jhu.edu/~juri/pdf/mllr-rwth-2005.pdf.
42. Gauvain J.-L., Lee C.-H. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing.* April 1994. Vol. 2. No 2. P. 291–298.
43. Huang Z., Siniscalchi S.M., Chen I., Li J., Wu J., Lee C. Maximum a posteriori adaptation of network parameters in deep models. *Interspeech*. 2015. https://doi.org/10.48550/arXiv.1503.02108
44. Lu L., Ghosal A., Renals S. Maximum a posteriori adaptation of subspace Gaussian mixture models for cross-lingual speech recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* : Proceedings. 2012. P. 4877 – 4880.
45. Digalakis V.V., Rtischev D., Neumeyer L.G. Speaker adaptation using constrained estimation of Gaussian mixtures. *IEEE Transactions on Speech and Audio Processing. 1995.* Vol. 3. No. 5. P. 357–366.
46. Gales M.J.F. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech & Language.* 1998. Vol. 12. P. 75–98.
47. Kim K., Gales M.J.F. Noisy constrained maximum likelihood linear regression for noise robust speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing.* 2011. Vol. 19. No 2. P. 315–325.
48. Lane H., Tranel B. The Lombard sign and the role of hearing in speech. *Journal of Speech and Hearing Research. 1971.* Vol. 14. No 4. P. 677.

49. Couvreur L., Couvreur C. Blind model selection for automatic speech recognition in reverberant environments. *Journal of VLSI Signal Processing.* 2004. Vol. 36. No 2/3. P. 189–203.
50. Stahl V., Fischer A., Bippus R. Acoustic synthesis of training data for speech recognition in living room environments. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* : Proceedings. 2001. Vol. 1. P. 21–24.
51. Huang X., Acero A., Hon H.-W. Spoken Language Processing. A Guide to Theory, Algorithm, and System Development. New Jersey: Prentice-Hall, 2001. 980 p.
52. Ko T., Peddinti V., Povey D., Seltzer M.L., Khudanpur S. A study on data augmentation of reverberant speech for robust speech recognition. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* : Proceedings. 2017. P. 5220-5224.
53. Kharitonov E., Rivière M., Synnaeve G., Wolf L., Mazar'e P., Douze M., Dupoux E. Data Augmenting Contrastive Learning of Speech Representations in the Time Domain. *IEEE Spoken Language Technology Workshop (SLT).* Shenzhen, China, 2020. P. 215-222.
54. Park D., Chan W., Zhang Yu., Chiu C.-C., Zoph B., Cubuk E., Le Q. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *Interspeech 2019*. Graz, Austria, 2019. P. 2613-2617.
55. Meng L., Xu J., Tan X., Wang J., Qin T., Xu B. MixSpeech: Data Augmentation for Low-Resource Automatic Speech Recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021. P. 7008-7012. https://doi.org/10.1109/ICASSP39728.2021.9414483
56. Tóth L., Kovács G., Compernolle D.V. A Perceptually Inspired Data Augmentation Method for Noise Robust CNN Acoustic Models. *20th*

International Conference on Speech and Computer (SPECOM 2018) : Proceedings. September 18–22, 2018. Leipzig, Germany, 2018. P. 697-706

57. C. Du C., Yu K. Speaker Augmentation for Low Resource Speech Recognition. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* Barcelona, Spain, 2020. P. 7719-7723.
58. Ragni A., Knill K., Rath S.P., Gales M.J. Data augmentation for low resource languages. *Interspeech*. 2014. URL: https://www.isca-speech.org/archive_v0/archive_papers/interspeech_2014/i14_0810.pdf
59. Hu Y, Kokkinakis K. Effects of early and late reflections on intelligibility of reverberated speech by cochlear implant listeners. *Journal of the Acoustical Society of America*. 2013. Vol. 135 (1). https://doi.org/10.1121/1.4834455
60. Yang W, Bradley J. Effects of room acoustics on the intelligibility of speech in classrooms. *Journal of the Acoustical Society of America*. 2009. Vol. 125 (2). P. 1-12. https://doi.org/10.1121/1.3058900
61. Bradley J. Review of objective room acoustics measures and future needs. *Applied Acoustics*. 2011. Vol. 72(10). P. 713-720. https://doi.org/10.1016/j.apacoust.2011.04.004
62. Arweiler I, Buchholz J, Dau T. Speech intelligibility enhancement by early reflections. *ISAAR 2009: Binaural Processing and Spatial Hearing, 2nd International Symposium on Auditory and Audiological Research* : Proceedings. 2009. Elsinore, Denmark 2009.
63. ISO 3382-1:2009(en). Acoustics – Measurement of room acoustic parameters – Part 1: Performance spaces. URL: https://www.iso.org/obp/ui/#iso:std:iso:3382:-1:ed-1:v1:en
64. Acoustic design of schools: performance standards. Building bulletin 93. 2015. UK Department for Education, UK Education Funding Agency. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/400784/BB93_February_2015.pdf

65. Prodeus A, Didkovska M. Assessment of speech intelligibility in university lecture rooms of different sizes using objective and subjective methods. *Eastern-European Journal of Enterprise Technologies*. 2021. Vol. 35(111). P. 47-56. https://doi.org/10.15587/1729-4061.2021.228405
66. Leccese F, Rocca M, Salvadori G. Fast estimation of Speech Transmission Index using the Reverberation Time: Comparison between predictive equations for educational rooms of different sizes. *Applied Acoustics*. 2018. Vol. 140. P. 143-149. https://doi.org/10.1016/j.apacoust.2018.05.019
67. Nestoras C, Dance S. The Interrelationship Between Room Acoustics Parameters as Measured in University Classrooms Using Four Source Configurations. *Building Acoustics*. 2013. Vol. 20(1). P. 43-54. https://doi.org/10.1260/1351-010X.20.1.43
68. Kuttruff, H. Room Acoustics. London: CRC Press, 2009. 392 p.
69. Eggenschwiler, K. Lecture Halls – Room Acoustics and Sound Reinforcement. *Conference: 4th European congress on acoustics (Forum Acusticum 2005)* : Proceedings. 2005. Budapest, 2005.
70. Bradley J. S., Reich R. D., Norcross S. G. On the combined effects of signal-to-noise ratio and room acoustics on speech intelligibility. *The Journal of the Acoustical Society of America*. 1999. Vol. 106 (4). P. 1820–1828. https://doi.org/10.1121/1.427932
71. Lochner J. P. A., Burger J. F. The influence of reflections on auditorium acoustics. *Journal of Sound and Vibration*. 1964. Vol. 1 (4). P. 426–454. https://doi.org/10.1016/0022-460x(64)90057-4
72. Bradley J. S., Sato H., Picard M. On the importance of early re- flections for speech in rooms. *The Journal of the Acoustical Society of America*. 2003. Vol. 113 (6). P. 3233. https://doi.org/10.1121/1.1570439
73. Eldakdoky, S. Optimizing acoustic conditions for two lecture rooms in Faculty of Agriculture, Cairo University. *Ain Shams Engineering Journal*. 2017. Vol. 8. P. 481–490. http://doi.org/10.1016/j.asej.2016.08.013

74. Duran S, Ausiello L, Battaner-Moro J. Acoustic Design Criteria for Higher-Education Learning Environments. *Reproduced Sound 2019: Creating Engagement in Sound* : Proceeding. Bristol, UK, 2019. Vol. 41(3). P. 1-12.
75. Choi Y-J. The intelligibility of speech in university classrooms during lectures. *Applied Acoustics*. 2020. https://doi.org/10.1016/j.apacoust.2020.107211
76. Prodeus A, Didkovska M, Motorniuk D, Dvornyk O. The Effects of Noise, Early and Late Reflactions on Speech Intelligibility. *IEEE 40th Int. Conf. on Electronics and Nanotechnology (ELNANO`2020)* : Proceedings. Kyiv, Ukraine, 2020. P. 488-492.
77. Prodeus A, Didkovska M. Objective assessment of speech intelligibility in small and medium-sized classrooms. *IEEE International Scientific-Practical Conf. on Problems of Infocommunications, Science, and Technology (PIC S&T`2020)* : Proceedings. Kharkiv, Ukraine 2020.
78. Loizou P. Speech Enhancement: Theory and Practice. Second Edition. Boca Raton: CRC Press, Taylor & Francis Group, 2013
79. Aleinik S.V., Matveev Yu. N., Sholokhov A.V. Detection of clipped fragments in acoustic signals. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics.* 2014. No 4 (92). P. 91–97.
80. Prodeus A., Kotvytskyi I., Ditiashov A. Clipped Speech Signals Quality Estimation. *5th Int. Conf. "Methods and Systems of Navigation and Motion Control" (MSNMC-2018)* : Proceedings. October 16-18, 2018, Kyiv, Ukraine. P. 151–155.
81. Arora V., Kumar R., Probability distribution estimation of music signals in time and frequency. *19th Int. Conf. on Digital Signal Processing (DSP-2014)* : Proceedings. 20-23 August 2014, Hong Kong. P. 409–414.
82. Avanesjan G.R. Method and device for estimating and indicating distortions of output signal of audio frequency amplifiers (overload indication). Patent RU 2274868 C2, Int.Cl. G01R 23/20, G01R 19/165. 2006

83. Cote N. Integral and diagnostic intrusive prediction of speech. Berlin, Heidelberg: Springer-Verlag, 2011. 248 p.
84. Prodeus A., Kotvytskyi I., Ditiashov A. Assessment of clipped speech quality. *Electronics and Control Systems*. 2018. No 4(58). P. 11–18. https://doi.org/:10.18372/1990-5548.58.13504
85. Aachen Impulse Response Database. URL: https://www.iks.rwth-aachen.de/en/research/tools-downloads/databases/aachen- impulse-response-database/
86. Jeub M, Schäfer M, Vary P. A binaural room impulse response database for the evaluation of dereverberation algorithms. *Int. Conf. Digital Signal Processing (DSP)* : Proceedings. Santorini, Greece, 2009. https://doi.org/10.1109/ICDSP.2009.5201259
87. Тихонов, А. Н. О некорректных задачах линейной алгебры и устойчивом методе их решения. Доклады Академии наук СССР. 1965. № 163 (3). С. 591–594.
88. Dvornyk O, Motorniuk D, Didkovska M, Prodeus A. Artificial Software Complex "Artificial Head". Part 1. Adjusting the Frequency Response of the Path," *Microsystems, Electronics and Acoustics.* 2020. Vol. 22(1). https://doi.org/10.20535/2523-4455.mea.198431
89. Perceptual Evaluation of Speech Quality (PESQ) ITU-T Recommendations P.862, P.862.1, P.862.2. Version 2.0. October 2005.
90. Cox, D.R., Hinkley, D.V. Theoretical Statistics. Chapman & Hall. Appendix 3. 1974.
91. Falk T, Zheng C, Chan W-Y. A Non-Intrusive Quality and Intelligibility Measure of Reverberant and Dereverberated Speech. *IEEE Transactions on Audio, Speech, and Language Processing.* 2010. Vol. 18(7), https://doi.org/10.1109/TASL.2010.2052247
92. Haas H. Uber den Einfluss eines Einfachechos auf die Horsamkeit von Sprache. *Acustica.* 1951. Vol. 1. P. 49–58.

93. Lochner J., Burger J. The influence of reflections on auditorium acoustics. *Journal of Sound and Vibration*. 1964. Vol. 1. P. 426–454. 1964. https://doi.org/10.1016/0022-460X(64)90057-4
94. Jacob K. Correlation of Speech Intelligibility Tests in Reverberant Rooms with Three Predictive Algorithms. *85th Convention of the Audio Engineering Society* : Proceedings. 1988. Los Angeles, Nov. 3-6, 1988. P. 1020–1030.
95. Xiong F., Goetze S., Meyer B. Estimating Room Acoustic Parameters for Speech Recognizer Adaptation and Combination in Reverberant Environments. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing* : Proceedings. 2014. https://doi.org/10.1109/ICASSP.2014.6854659
96. Golzer H., Kleinschmidt M. Importance of early and late reflections for automatic speech recognition in reverberant environments. *Proc. Elektronische Sprachsignalverarbeitung (ESSV).* 2003.
97. Schroeder M.R., Atal B. S., Bird C. Digital computers in room acoustics. *4th International Congress on Acoustics (ICA)* : Proceedings. Copenhagen, 1962. No 21.
98. Rindel J. The Use of Computer Modeling in Room Acoustics. *Journal of Vibroengineearing*. 2000. No. 3(4). P. 219–224.
99. Vorlander M. Computer simulations in room acoustics: Concepts and uncertainties. *The Journal of the Acoustical Society of America.* 2013. Vol. 133 (3). P. 1203–1213. https://doi.org/10.1121/1.4788978
100. Jacob K. D., Birkle T. K., Ickler C. B. Accurate Prediction of Speech Intelligibility without the Use of In-Room Measurements. *Journal of the Audio Engineering Society.* 1991. Vol. 39. No 4. P. 232–242.
101. Steeneken H. J. M., Houtgast T. Validation of the revised STIr method. *Elsevier Speech Communication.* 2002. Vol. 38. P. 26–37. https://doi.org/10.1016/S0167-6393(02)00010-9
102. Steeneken H. Forty years of speech intelligibility assessment (and some history). *Proc. of the Institute of Acoustics.* 2014. Vol. 36. Pt. 3.

103. Warzybok A., Rennies J., Doclo S., Kollmeier B. Influence of early reflections on speech intelligibility under different noise conditions. *Forum Acusticum 2011* : Proceedings. June 27-July 1 2011. Aalborg, Denmark, 2011.
104. Mansour D., Juang B.H. A family of distortion measures based upon projection operation for robust speech recognition // IEEE Transactions on Acoustics, Speech, and Signal Processing. Nov 1989. Vol. 37. Issue 11. P. 1659-1671
105. Morris A.C., Maier V., Green P.D. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. *Interspeech*. *8th International Conference on Spoken Language Processing*. October 4-8, 2004. ICC Jeju, Jeju Island, Korea, 2004.
106. Дідковський В.С., Дідковська М.В., Продеус А.М. Комп’ютерна обробка акустичних сигналів: Навчальний посібник. Київ, 2010. 430 с.

ДОДАТОК А

Список опублікованих праць за темою дисертації

Статті в періодичних наукових виданнях інших держав або прирівняні до таких (публікації у виданнях категорії “А” або в закордонних виданнях, проіндексованих у базах даних Web of Science Core Collection та/або Scopus):

1. Prodeus A., Didkovska M., Kukharicheva K. Impact of University Classroom Size on the Relationship Between Speech Quality and Intelligibility // International Journal of Computing. 2022. Vol. 21 (3). https://doi.org/10.47839/ijc.21.3.2690
2. Prodeus A, Kukharicheva K, Didkovska M. Comparison of Speech Quality and Intelligibility Assessments in University Classrooms // International Journal of Architectural Engineering Technology. 2021. Vol. 8. P. 52–60. http://dx.doi.org/10.15377/2409-9821.2021.08.5

Статті в наукових виданнях, включених до переліку наукових фахових видань України з присвоєнням категорії “Б”:

1. Prodeus A., Didkovska M., Kukharicheva K, Motorniuk D. Two Simplified Models Of Early Sound Reflections In a Room // Electronics and Control Systems. 2020. Vol. 3, (65). ISSN 1990-5548. https://doi.org/10.18372/1990-5548.65.14991
2. Prodeus A., Kotvytskyi I., Didkovska M., Kukharicheva K. Kurtosis and Normalized Variance as Measures of Speech Signals Clipping Value // Electronics and Control Systems, 2019. Vol. 4 (62). P.24-32. https://doi.org/10.18372/1990-5548.62.14378

3. Prodeus A., Kukharicheva K. Accuracy of Automatic Speech Recognition System Trained on Noised Speech // Electronics and Control Systems. 2016. Vol. 3 (49). https://doi.org/10.18372/1990-5548.49.11230

Доповіді на конференціях:

1. Prodeus A., Didkovska M., Kukharicheva K., Motorniuk D. Modeling the Influence of Early Sound Reflections on Speech Intelligibility // Proceedings of 2020 IEEE 6th International Conference on Methods and Systems of Navigation and Motion Control (MSNMC), Kyiv, 20-23 Oct. 2020. Kyiv, Ukraine. https://doi.org/10.1109/MSNMC50359.2020.9255657
2. Prodeus A., Kotvytskyi I., Didkovska M., Kukharicheva K. Kurtosis and Its Transformations as Objective Measures of Clipping Value and Speech Quality // Proceedings of IEEE 5th International Conference on Actual Problems of Unmanned Aerial Vehicles Developments (APUAVD), Kyiv, 22-24 October 2019. Kyiv, Ukraine. https://doi.org/10.1109/APUAVD47061.2019.8943880
3. Prodeus A., Kukharicheva K. Automatic Speech Recognition Performance for Training on Noised Speech // Proceedings of 2017 IEEE 2nd International Conference on Advanced Information and Communication Technologies (AICT), Lviv, 4-7 July 2017. Lviv, Ukraine. https://doi.org/10.1109/AIACT.2017.8020068
4. Prodeus A., Kukharicheva K. Training of automatic speech recognition system on noised speech // Proceedings of 2016 IEEE 4th Int. Conf. "Methods and Systems of Navigation and Motion Control (MSNMC)», Kyiv, October 18-20, 2016. Kyiv, Ukraine. P. 221-223. https://doi.org/10.1109/MSNMC.2016.7783147

А. М. Продеус, І. В. Котвицький, М. В. Дідковська, В. С. Дідковський, К. А. Кухарічева, Д. Є. Моторнюк, О. О. Дворник Спосіб виявлення кліпування

мовного та музичного сигналів // **Патент** UA 144291 U, МПК G01R 23/20, опубл. 25.09.2020.

ДОДАТОК Б

Результати дослідження точності розпізнавання системи АРМ при методах навчання з використанням зашумлених сигналів (графічне представлення)

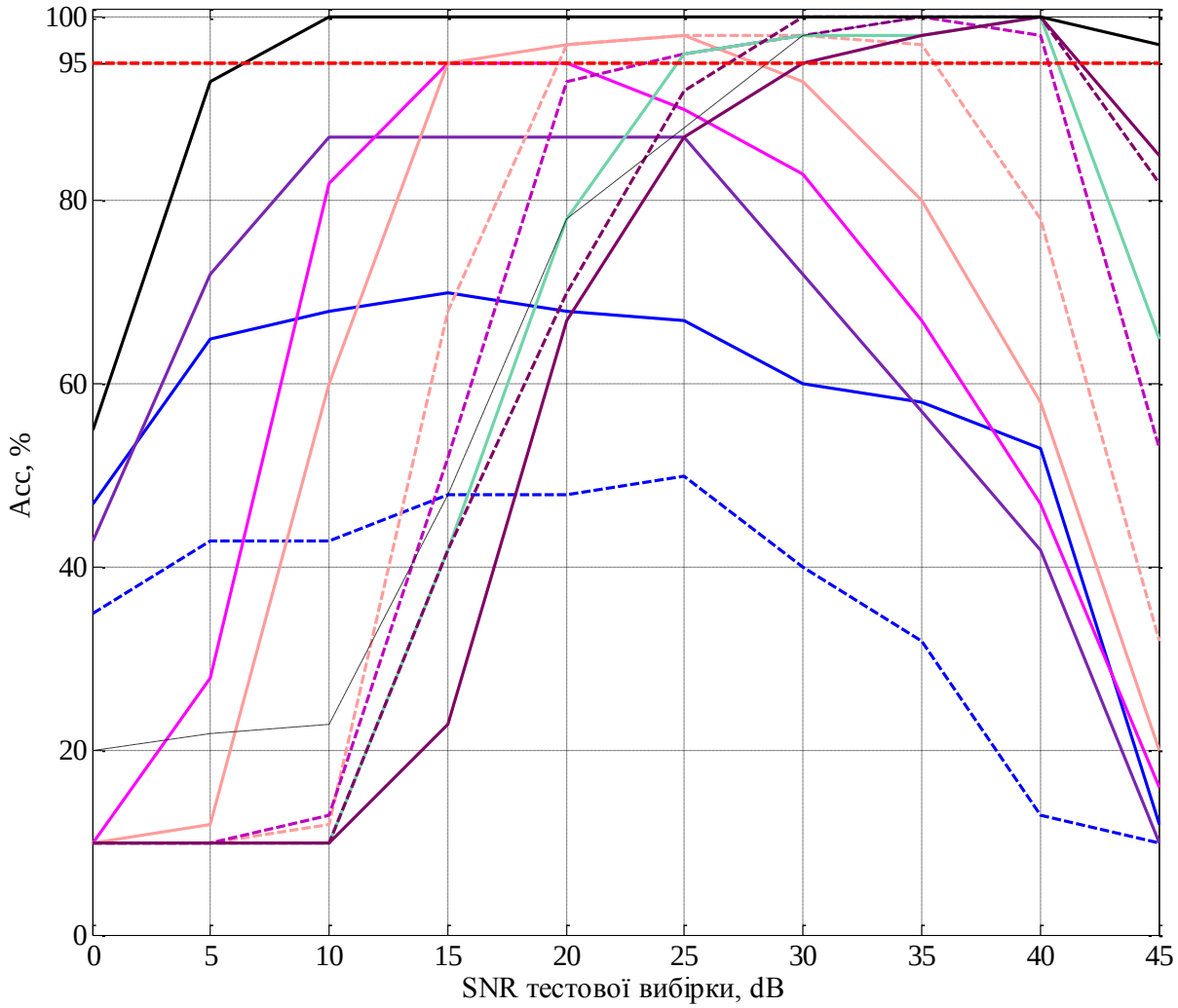


Рис. Б.1. Графік залежності точності розпізнавання від відношення сигнал-шум тестової вибірки для шуму **у фойє залізничного вокзалу**. SNR навчальних вибірок:

0 dB — 5 dB — 10 dB — 15 dB — 20 dB — 25 dB — 30 dB — 35 dB — 40 dB — 45 dB — Універсальна вибірка — "Чисті" сигнали

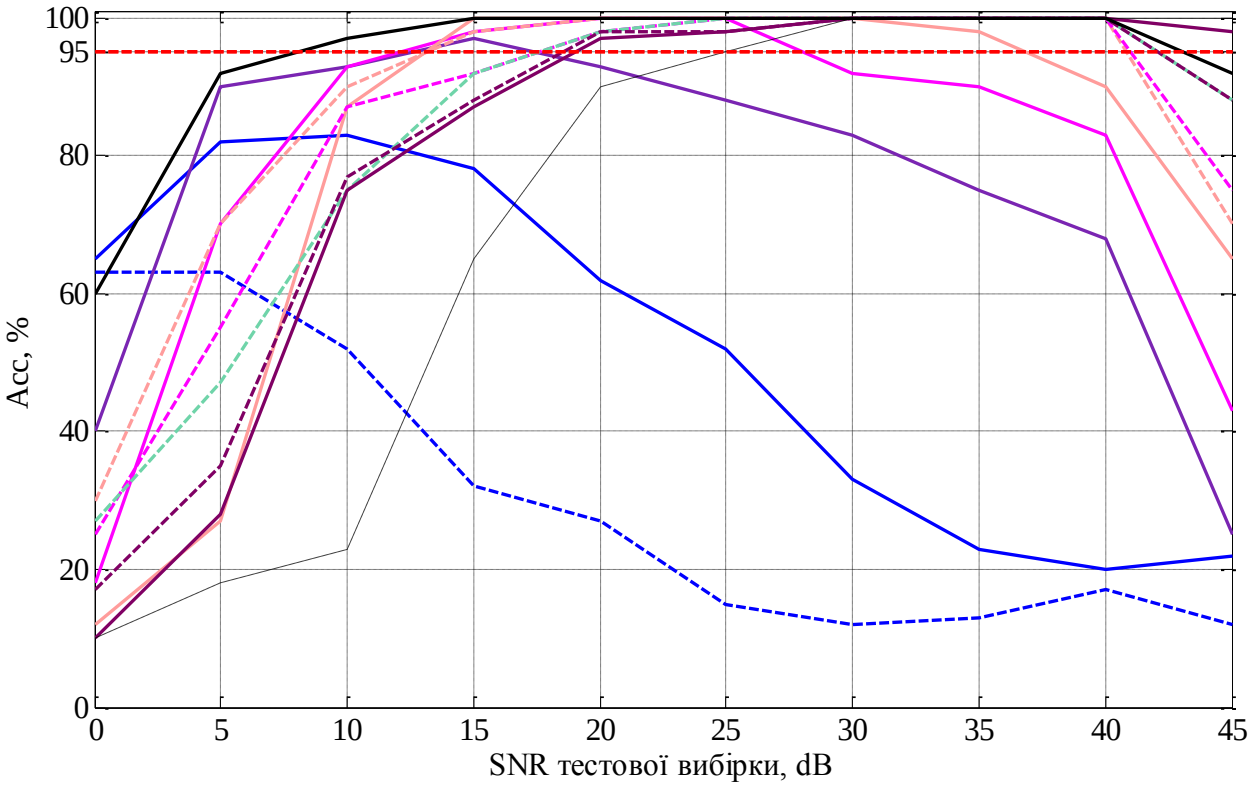


Рис. Б.2. Графік залежності точності розпізнавання від відношення сигнал-шум тестової вибірки для шуму **на тролейбусній зупинці**.

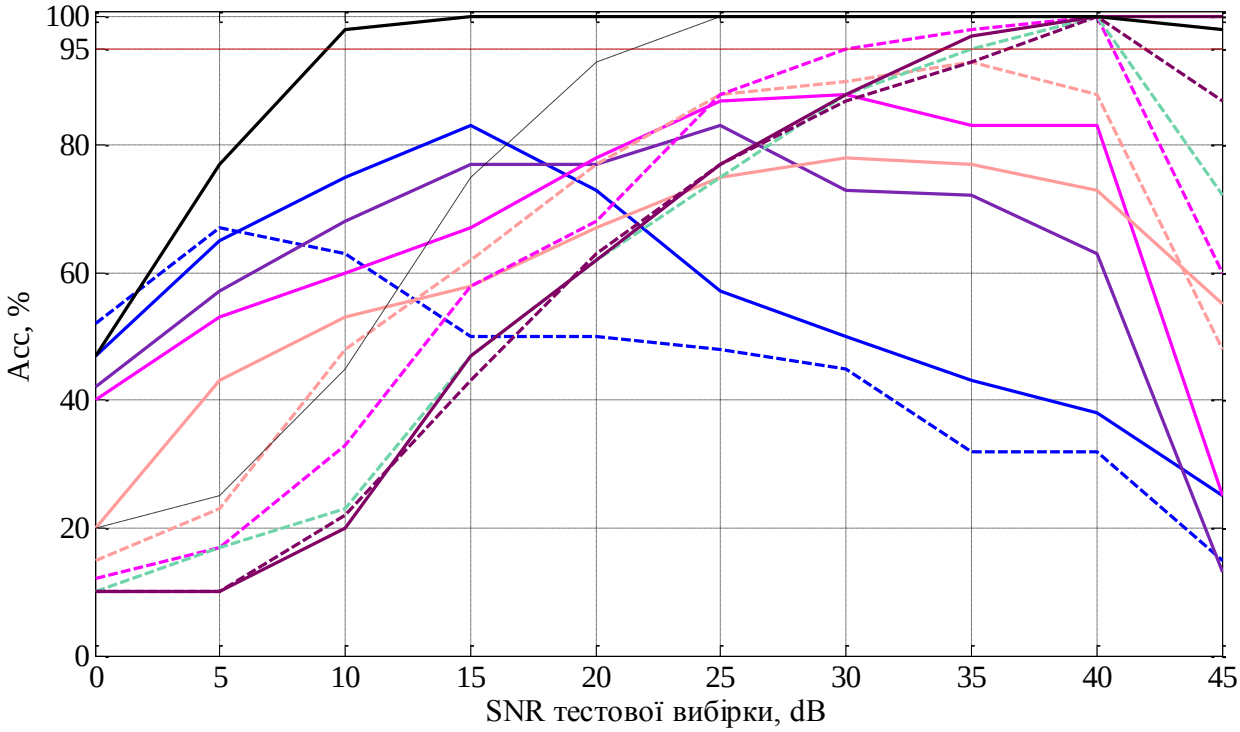


Рис. Б.3. Графік залежності точності розпізнавання від відношення сигнал-шум тестової вибірки для шуму **пральної машини.** SNR навчальних вибірок:

0 dB, 5 dB, 10 dB, 15 dB, 20 dB, 25 dB, 30 dB, 35 dB, 40 dB, 45 dB, Універсальна вибірка, "Чисті" сигнали

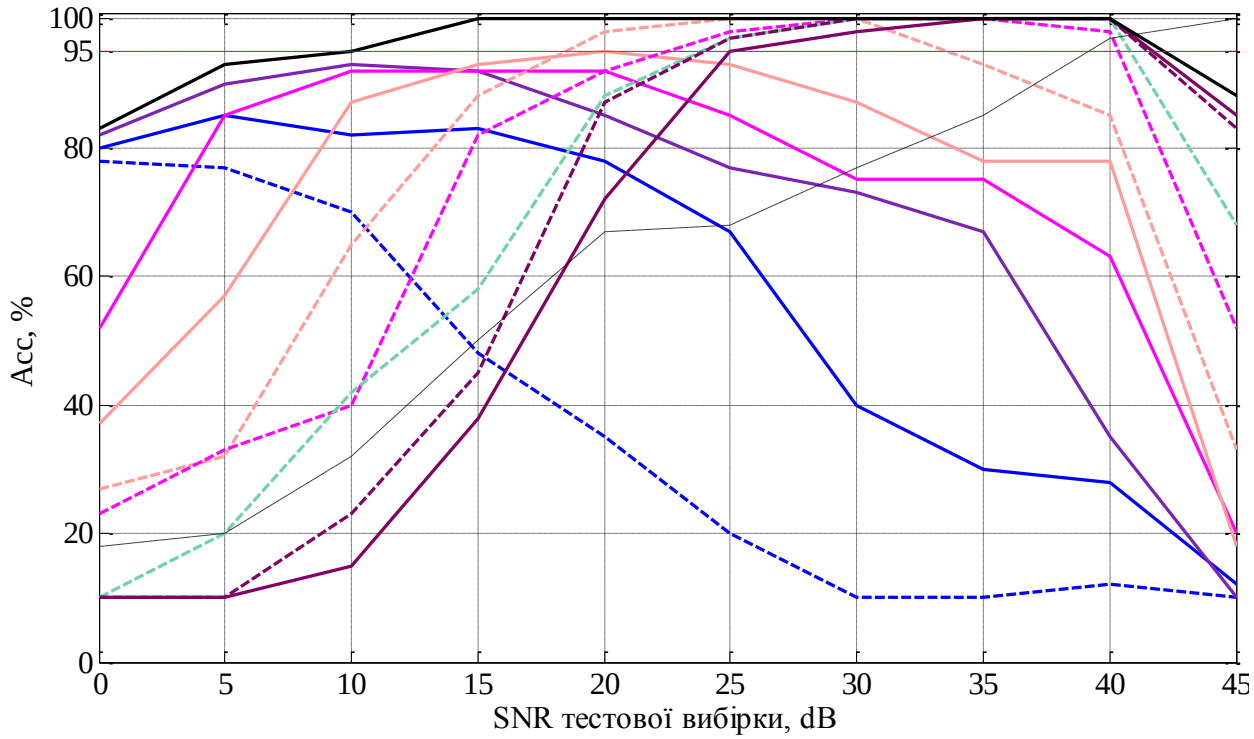


Рис. Б.4. Графік залежності точності розпізнавання від відношення сигнал-шум тестової вибірки для шуму **у підземному переході між вокзалами.**

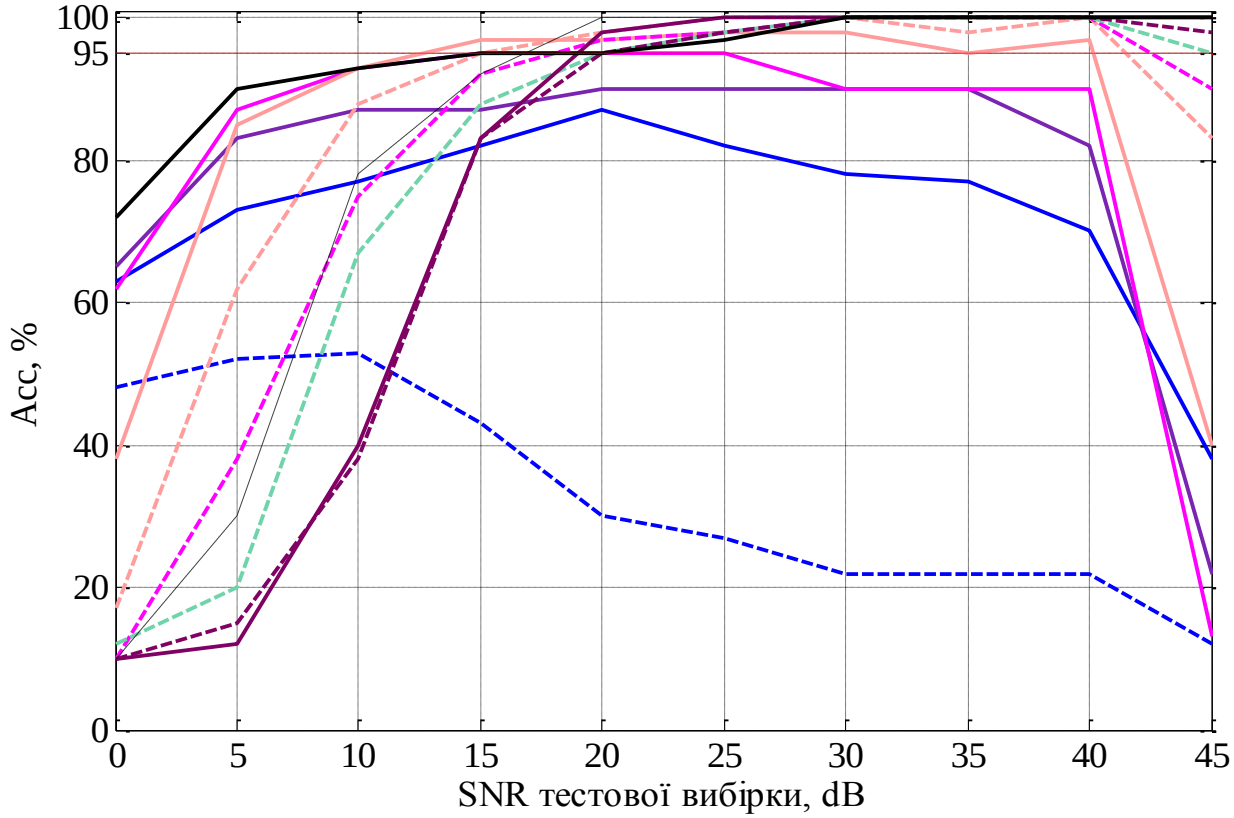


Рис. Б.5. Графік залежності точності розпізнавання від відношення сигнал-шум тестової вибірки для шуму **біля входу у вокзал.** SNR навчальних вибірок:

0 dB — 5 dB — 10 dB — 15 dB — 20 dB — 25 dB — 30 dB — 35 dB — 40 dB — 45 dB — Універсальна вибірка — "Чисті" сигнали

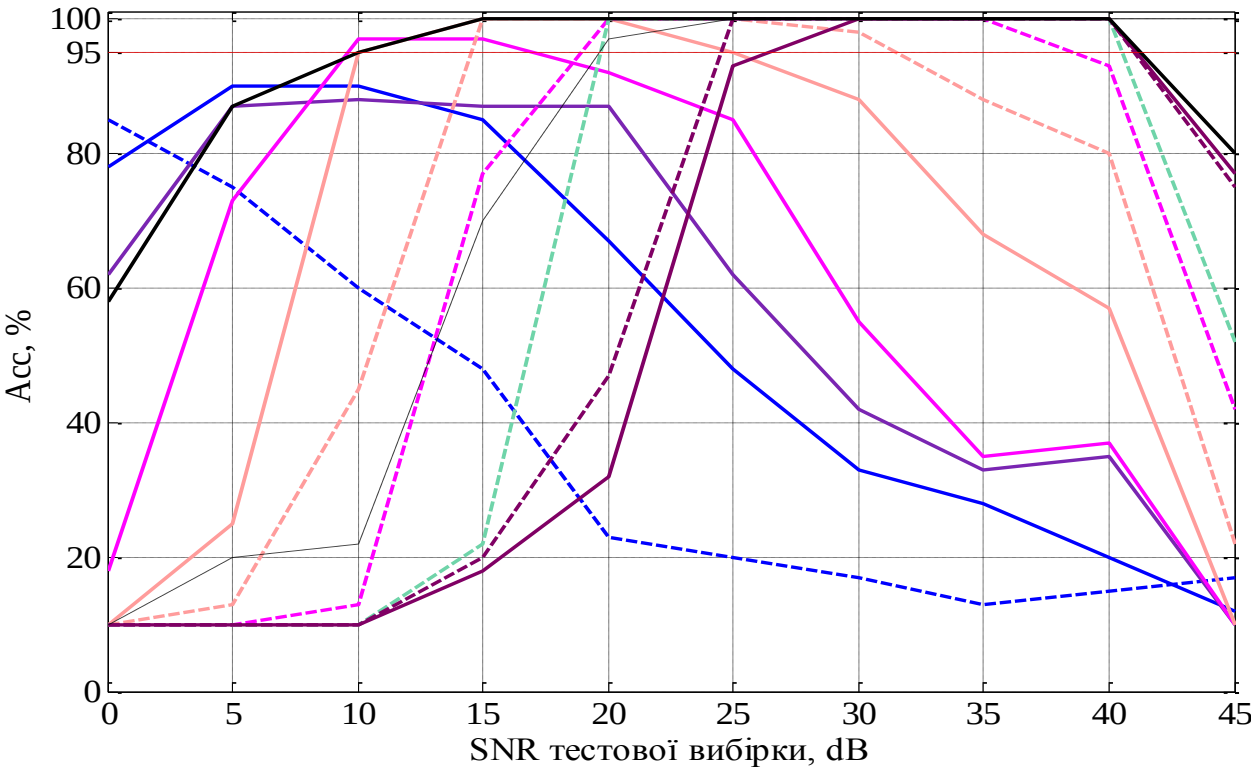


Рис. Б.6. Графік залежності точності розпізнавання від відношення сигнал-шум тестової вибірки для шуму **мікрохвильової печі.**

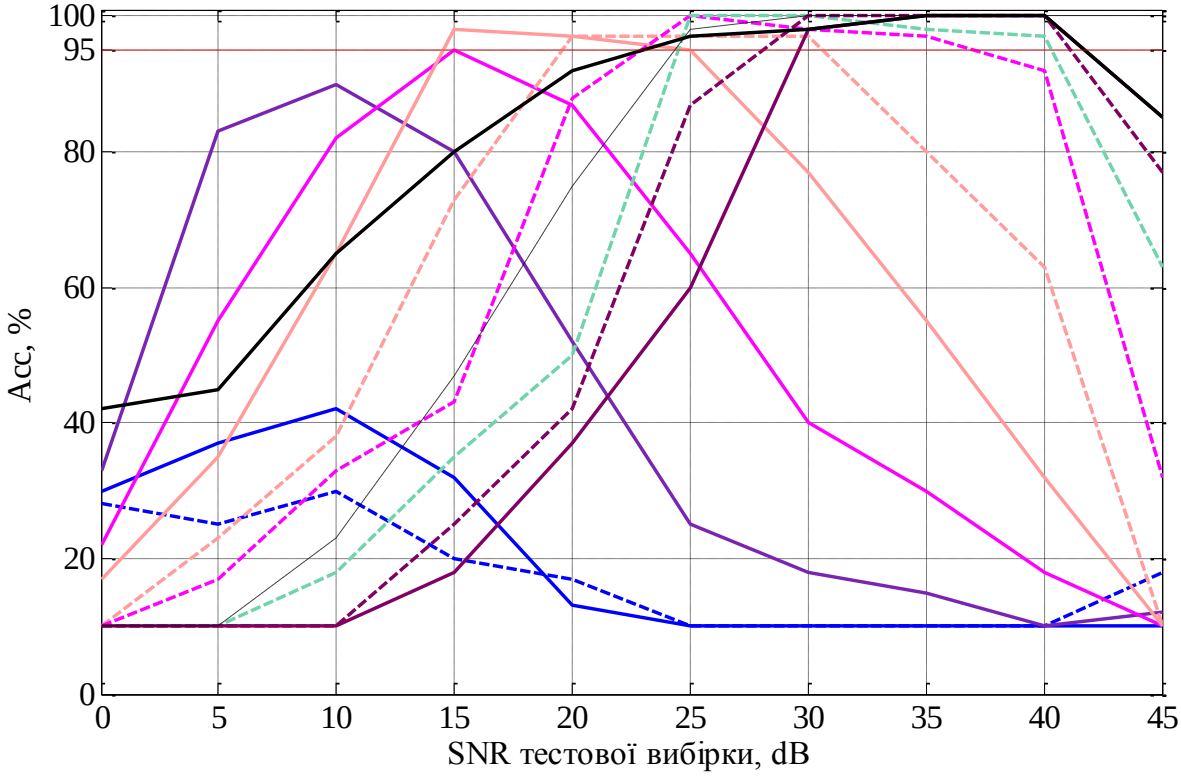


Рис. Б.7. Графік залежності точності розпізнавання від відношення сигнал-шум тестової вибірки для шуму **поїзда метро під час розгону.** SNR навчальних вибірок:

0 dB — 5 dB — 10 dB — 15 dB — 20 dB — 25 dB — 30 dB
35 dB — 40 dB — 45 dB — Універсальна вибірка — — "Чисті" сигнали

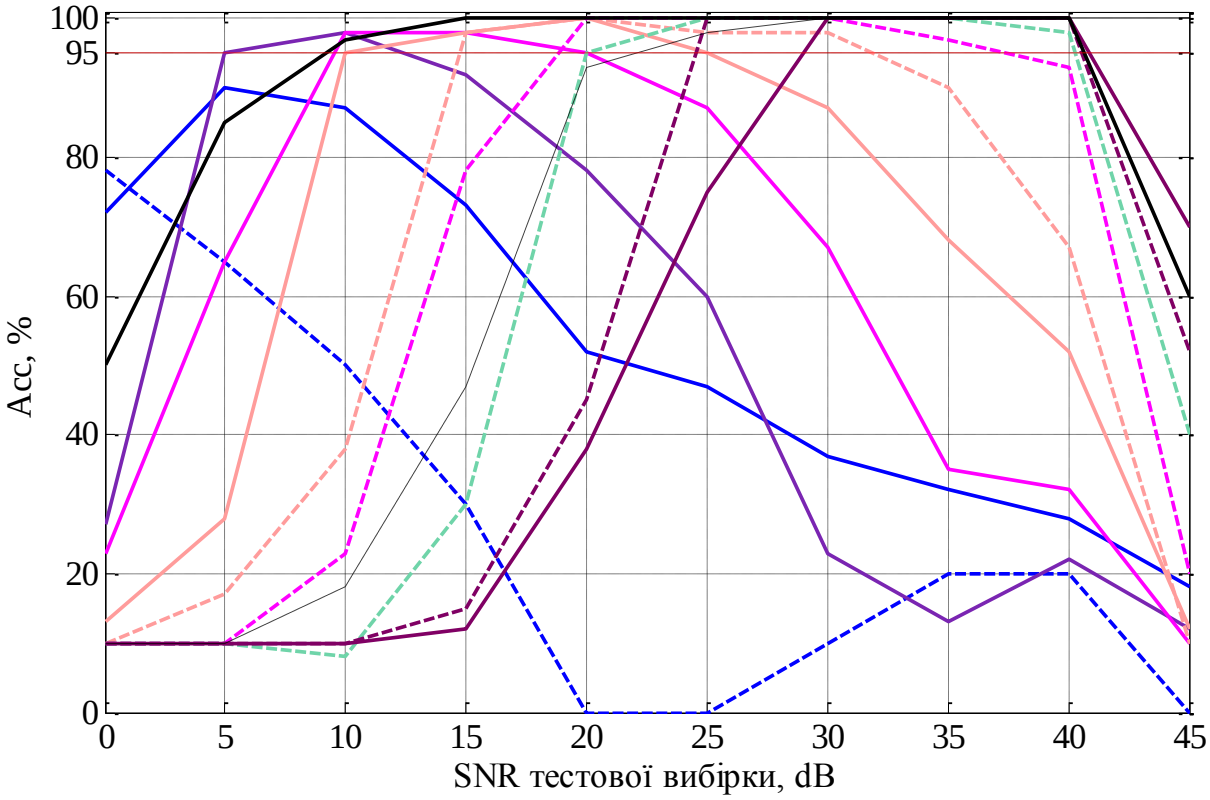


Рис. Б.8. Графік залежності точності розпізнавання від відношення сигнал-шум тестової вибірки для шуму **комп'ютера.**

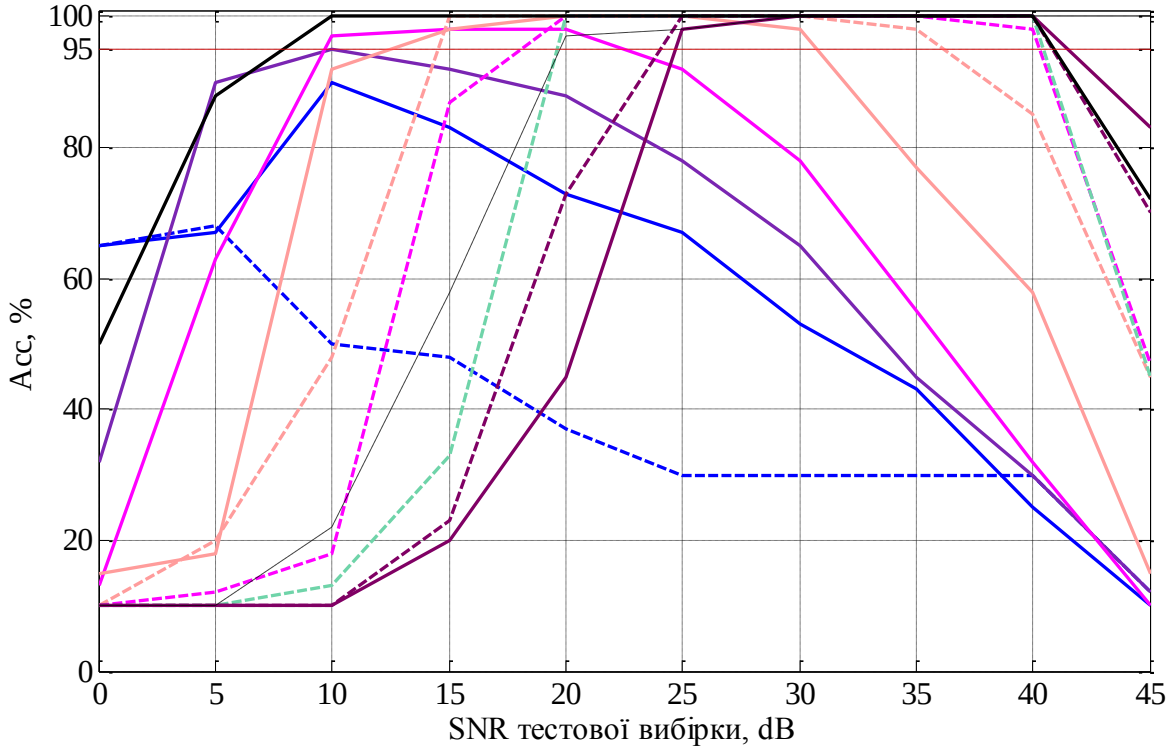


Рис. Б.9. Графік залежності точності розпізнавання від відношення сигнал-шум тестової вибірки для шуму **вантажівок.** SNR навчальних вибірок:

0 dB 5 dB 10 dB 15 dB 20 dB 25 dB 30 dB
35 dB 40 dB 45 dB Універсальна вибірка "Чисті" сигнали

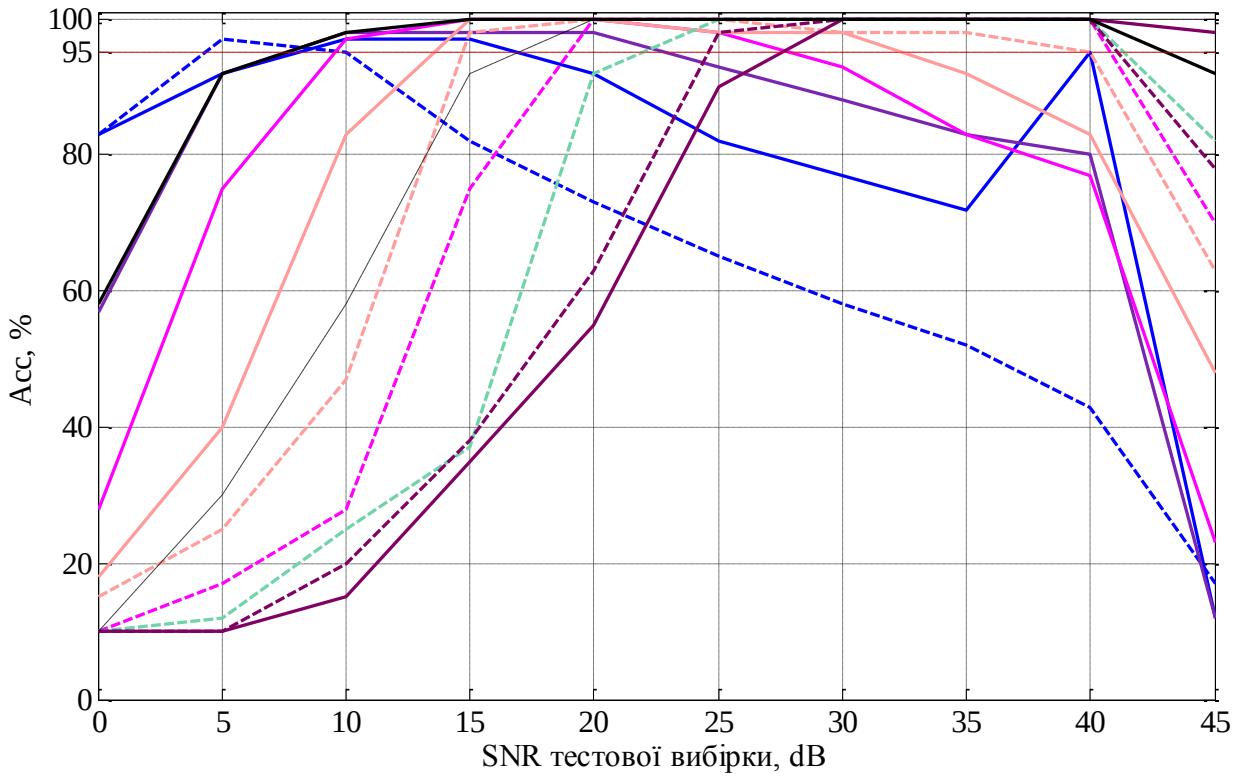


Рис. Б.10. Графік залежності точності розпізнавання від відношення сигнал-шум тестової вибірки для шуму **в тролейбусі.**

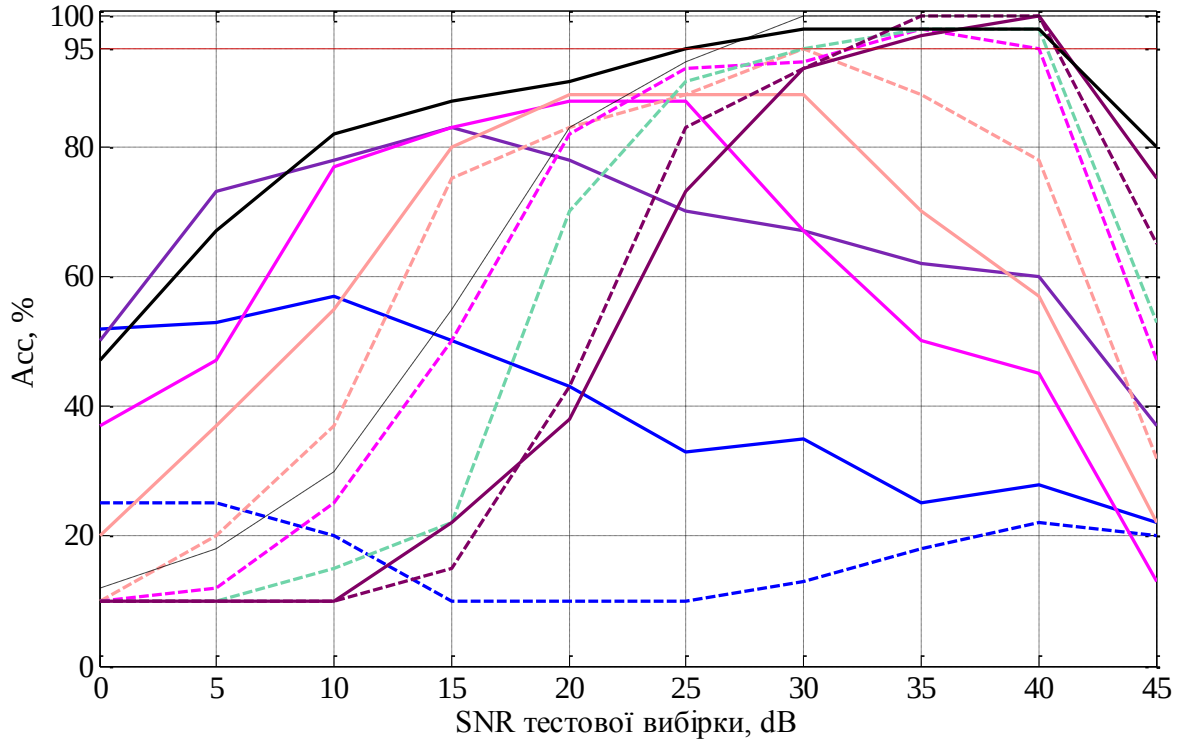


Рис. Б.11. Графік залежності точності розпізнавання від відношення сигнал-шум тестової вибірки для шуму **аудиторії, в якій присутні 13 чеоловік.** SNR навчальних вибірок:

0 dB 5 dB 10 dB 15 dB 20 dB 25 dB 30 dB 35 dB 40 dB 45 dB Універсальна вибірка "Чисті" сигнали

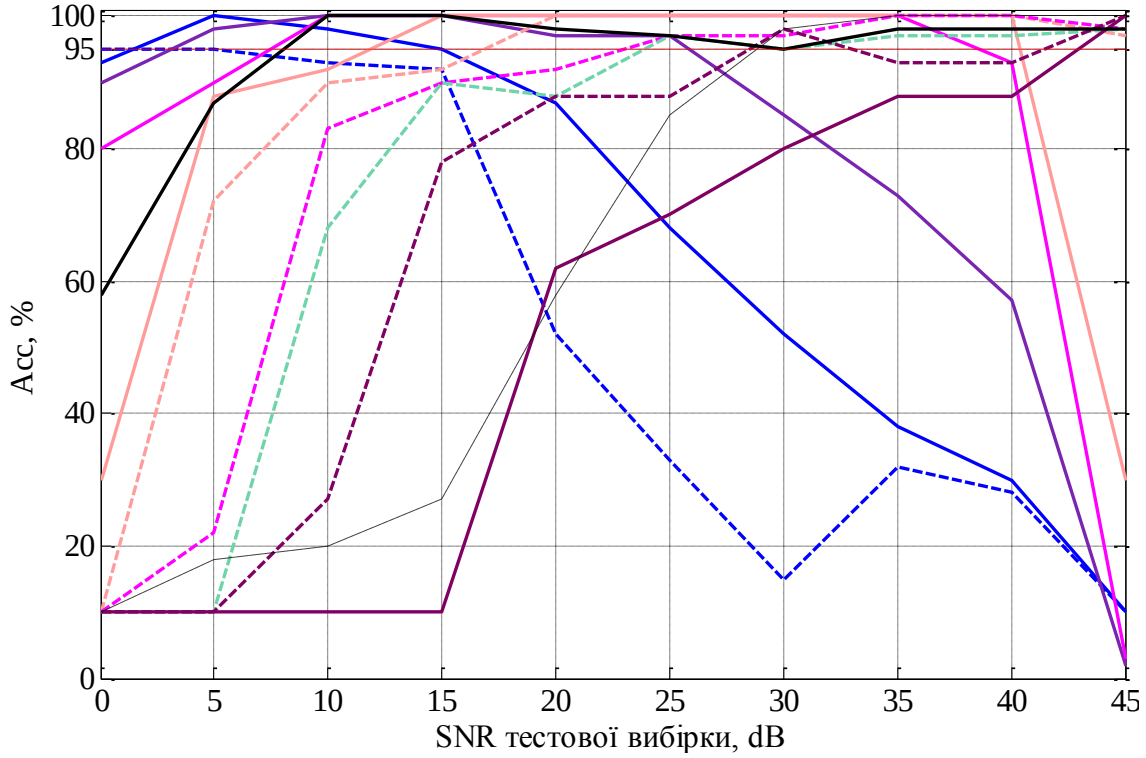


Рис. Б.12. Графік залежності точності розпізнавання від відношення сигнал-шум тестової вибірки для шуму **кавомолки.**

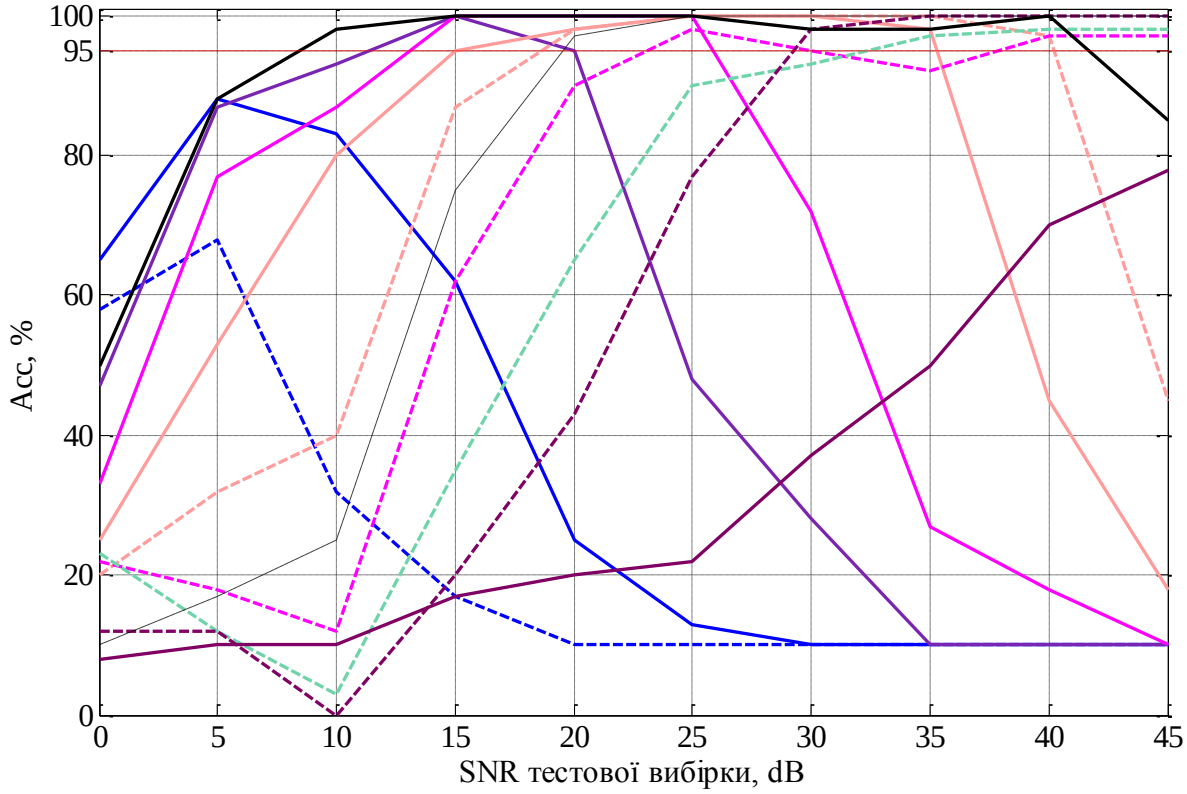


Рис. Б.13. Графік залежності точності розпізнавання від відношення сигнал-шум тестової вибірки для шуму **фойє метро.** SNR навчальних вибірок:

0 dB 5 dB 10 dB 15 dB 20 dB 25 dB 30 dB
35 dB 40 dB 45 dB Універсальна вибірка "Чисті" сигнали

ДОДАТОК В

Результати дослідження точності розпізнавання системи АРМ при методах навчання з використанням зашумлених сигналів (табличне представлення)

Таблиця В.1. Точність розпізнавання Acc, %, шум у фойє метро

	Номер частини експерименту	1	2										3
	SNR тестової вибірки, дБ	SNR навчальної вибірки, дБ											
		45,«чистий» сигнал	0	5	10	15	20	25	30	35	40	45	Усі SNR
Acc, %	0	10	58	65	47	33	25	20	22	23	12	8	50
	5	17	68	88	87	77	53	32	18	12	12	10	88
	10	25	32	83	93	87	80	40	12	3	0	10	98
	15	75	17	62	100	100	95	87	62	35	20	17	100
	20	97	10	25	95	100	98	98	90	65	43	20	100
	25	100	10	13	48	100	100	100	98	90	77	22	100
	30	100	10	10	28	72	100	100	95	93	98	37	98
	35	100	10	10	10	27	98	100	92	97	100	50	98
	40	100	10	10	10	18	45	97	97	98	100	70	100
	45	100	10	10	10	10	18	45	97	98	100	78	85

Таблиця В.2. Точність розпізнавання Acc, %, шум кавомолки

	Номер частини експерименту	1	2										3
	SNR тестової вибірки, дБ	SNR навчальної вибірки, дБ											
		45, «чистий» сигнал	0	5	10	15	20	25	30	35	40	45	Усі SNR
Acc, %	0	10	95	93	90	80	30	10	10	10	10	10	58
	5	18	95	100	98	90	88	72	22	10	10	10	87
	10	20	93	98	100	100	92	90	83	68	27	10	100
	15	27	92	95	100	100	100	92	90	90	78	10	100
	20	58	52	87	97	100	100	100	92	88	88	62	98
	25	85	33	68	97	100	100	100	97	97	88	70	97
	30	98	15	52	85	100	100	100	97	95	98	80	95
	35	100	32	38	73	100	100	100	100	97	93	88	98
	40	100	28	30	57	93	100	100	100	97	93	88	98
	45	100	10	10	2	3	30	97	98	98	100	100	98

Таблиця В.3. Точність розпізнавання Acc, %, шум в аудиторії, в якій присутні 13 людей

	Номер частини експерименту	1	2										3
	SNR тестової вибірки, дБ	SNR навчальної вибірки, дБ											
		45, «чистий» сигнал	0	5	10	15	20	25	30	35	40	45	Усі SNR
Acc, %	0	12	25	52	50	37	20	10	10	10	10	10	47
	5	18	25	53	73	47	37	20	12	10	10	10	67
	10	30	20	57	78	77	55	37	25	15	10	10	82
	15	55	10	50	83	83	80	75	50	22	15	22	87
	20	83	10	43	78	87	88	83	82	70	43	38	90
	25	93	10	33	70	87	88	88	92	90	83	73	95
	30	100	13	35	67	67	88	95	93	95	92	92	98
	35	100	18	25	62	50	70	88	98	98	100	97	98
	40	100	22	28	60	45	57	78	95	98	100	100	98
	45	100	20	22	37	13	22	32	47	53	65	75	80

Таблиця В.4. Точність розпізнавання Acc, %, шум в тролейбусі

	Номер частини експерименту	1	2										3
	SNR тестової вибірки, дБ	SNR навчальної вибірки, дБ											
		45, «чистий» сигнал	0	5	10	15	20	25	30	35	40	45	Усі SNR
Acc, %	0	10	83	83	57	28	18	15	10	10	10	10	58
	5	30	97	92	92	75	40	25	17	12	10	10	92
	10	58	95	97	98	97	83	47	28	25	20	15	98
	15	92	82	97	98	100	100	98	75	37	38	35	100
	20	100	73	92	98	100	100	100	100	92	63	55	100
	25	100	65	82	93	98	98	100	100	100	98	90	100
	30	100	58	77	88	93	98	98	100	100	100	100	100
	35	100	52	72	83	83	92	98	100	100	100	100	100
	40	100	43	95	80	77	83	95	100	100	100	100	100
	45	100	17	12	12	23	48	63	70	82	78	98	92

Таблиця В.5. Точність розпізнавання Acc, %, шум вантажівки

	Номер частини експерименту	1	2										3
	SNR тестової вибірки, дБ	SNR навчальної вибірки, дБ											
		45, «чистий» сигнал	0	5	10	15	20	25	30	35	40	45	Усі SNR
Acc, %	0	10	65	65	32	13	15	10	10	10	10	10	50
	5	10	68	67	90	63	18	20	12	10	10	10	88
	10	22	50	90	95	97	92	48	18	13	10	10	100
	15	58	48	83	92	98	98	100	87	33	23	20	100
	20	97	37	73	88	98	100	100	100	100	73	45	100
	25	98	30	67	78	92	100	100	100	100	100	98	100
	30	100	30	53	65	78	98	100	100	100	100	100	100
	35	100	30	43	45	55	77	98	100	100	100	100	100
	40	100	30	25	30	32	58	85	98	100	100	100	100
	45	100	12	10	12	10	15	45	47	45	70	83	72

Таблиця В.6. Точність розпізнавання Acc, %, шум комп'ютера

	Номер частини експерименту	1	2										3
	SNR тестової вибірки, дБ	SNR навчальної вибірки, дБ											
		45, «чистий» сигнал	0	5	10	15	20	25	30	35	40	45	Усі SNR
Acc, %	0	10	78	72	27	23	13	10	10	10	10	10	50
	5	10	65	90	95	65	28	17	10	10	10	10	85
	10	18	50	87	98	98	95	38	23	8	10	10	97
	15	47	30	73	92	98	98	98	78	30	15	12	100
	20	93	0(-12)	52	78	95	100	100	100	95	45	38	100
	25	98	0(-38)	47	60	87	95	98	100	100	100	75	100
	30	100	0(-23)	37	23	67	87	98	100	100	100	100	100
	35	100	20	32	13	35	68	90	97	100	100	100	100
	40	100	20	28	22	32	52	67	93	98	100	100	100
	45	100	10	18	12	10	12	10	20	40	52	70	60

Таблиця В.7. Точність розпізнавання Acc, %, шум поїзда метро під час розгону

	Номер частини експерименту	1	2										3
	SNR тестової вибірки, дБ	SNR навчальної вибірки, дБ											
		45, «чистий» сигнал	0	5	10	15	20	25	30	35	40	45	Усі SNR
Acc, %	0	10	28	30	33	22	17	10	10	10	10	10	42
	5	10	25	37	83	55	35	23	17	10	10	10	45
	10	23	30	42	90	82	65	38	33	18	10	10	65
	15	47	20	32	80	95	98	73	43	35	25	18	80
	20	75	17	13	52	87	97	97	88	50	42	37	92
	25	98	10	10	25	65	95	97	100	100	87	60	97
	30	100	10	10	18	40	77	97	98	100	100	98	98
	35	100	10	10	15	30	55	80	97	98	100	100	100
	40	100	10	10	10	18	32	63	92	97	100	100	100
	45	100	18	10	12	10	10	10	32	63	77	85	85

Таблиця В.8. Точність розпізнавання Acc, %, шум мікрохвильової печі

	Номер частини експерименту	1	2										3
	SNR тестової вибірки, дБ	SNR навчальної вибірки, дБ											
		45, «чистий» сигнал	0	5	10	15	20	25	30	35	40	45	Усі SNR
Acc, %	0	10	85	78	62	18	10	10	10	10	10	10	58
	5	20	75	90	87	73	25	13	10	10	10	10	87
	10	22	60	90	88	97	95	45	13	10	10	10	95
	15	70	48	85	87	97	100	100	77	22	20	18	100
	20	97	23	67	87	92	100	100	100	100	47	32	100
	25	100	20	48	62	85	95	100	100	100	100	93	100
	30	100	17	33	42	55	88	98	100	100	100	100	100
	35	100	13	28	33	35	68	88	100	100	100	100	100
	40	100	15	20	35	37	57	80	93	100	100	100	100
	45	100	17	12	10	10	10	22	42	52	75	77	80

Таблиця В.9. Точність розпізнавання Acc, %, шум біля входу у вокзал

	Номер частини експерименту	1	2										3
	SNR тестової вибірки, дБ	SNR навчальної вибірки, дБ											
		45, «чистий» сигнал	0	5	10	15	20	25	30	35	40	45	Усі SNR
Acc, %	0	10	48	63	65	62	38	17	10	12	10	10	72
	5	30	52	73	83	87	85	62	38	20	15	12	90
	10	78	53	77	87	93	93	88	75	67	38	40	93
	15	92	43	82	87	95	97	95	92	88	83	83	95
	20	100	30	87	90	95	97	98	97	95	95	98	95
	25	100	27	82	90	95	98	100	98	98	98	100	97
	30	100	22	78	90	90	98	100	100	100	100	100	100
	35	100	22	77	90	90	95	98	100	100	100	100	100
	40	100	22	70	82	90	97	100	100	100	100	100	100
	45	100	12	38	22	13	40	83	90	95	98	100	100

Таблиця В.10. Точність розпізнавання Acc, %, шум у підземному переході між вокзалами

	Номер частини експерименту	1	2										3
	SNR тестової вибірки, дБ	SNR навчальної вибірки, дБ											
		45, «чистий» сигнал	0	5	10	15	20	25	30	35	40	45	Усі SNR
Acc, %	0	18	78	80	82	52	37	27	23	10	10	10	83
	5	20	77	85	90	85	57	32	33	20	10	10	93
	10	32	70	82	93	92	87	65	40	42	23	15	95
	15	50	48	83	92	92	93	88	82	58	45	38	100
	20	67	35	78	85	92	95	98	92	88	87	72	100
	25	68	20	67	77	85	93	100	98	97	97	95	100
	30	77	10	40	73	75	87	100	100	100	100	98	100
	35	85	10	30	67	75	78	93	100	100	100	100	100
	40	97	12	28	35	63	78	85	98	100	100	100	100
	45	100	10	12	10	20	18	33	52	68	83	85	88

Таблиця В.11. Точність розпізнавання Acc, %, шум пральної машини

	Номер частини експерименту	1	2										3
	SNR тестової вибірки, дБ	SNR навчальної вибірки, дБ											
		45, «чистий» сигнал	0	5	10	15	20	25	30	35	40	45	Усі SNR
Acc, %	0	20	52	47	42	40	20	15	12	10	10	10	47
	5	25	67	65	57	53	43	23	17	17	10	10	77
	10	45	63	75	68	60	53	48	33	23	22	20	98
	15	75	50	83	77	67	58	62	58	47	43	47	100
	20	93	50	73	77	78	67	77	68	62	63	62	100
	25	100	48	57	83	87	75	88	88	75	77	77	100
	30	100	45	50	73	88	78	90	95	88	87	88	100
	35	100	32	43	72	83	77	93	98	95	93	97	100
	40	100	32	38	63	83	73	88	100	100	100	100	100
	45	100	15	25	13	25	55	48	60	72	87	100	98

Таблиця В.12. Точність розпізнавання Acc, %, шум на тролейбусній зупинці

	Номер частини експерименту	1	2										3
	SNR тестової вибірки, дБ	SNR навчальної вибірки, дБ											
		45, «чистий» сигнал	0	5	10	15	20	25	30	35	40	45	Усі SNR
Acc, %	0	10	63	65	40	18	12	30	25	27	17	10	60
	5	18	63	82	90	70	27	70	55	47	35	28	92
	10	23	52	83	93	93	87	90	87	75	77	75	97
	15	65	32	78	97	98	100	98	92	92	88	87	100
	20	90	27	62	93	100	100	100	98	98	98	97	100
	25	95	15	52	88	100	100	100	100	100	98	98	100
	30	100	12	33	83	92	100	100	100	100	100	100	100
	35	100	13	23	75	90	98	100	100	100	100	100	100
	40	100	17	20	68	83	90	100	100	100	100	100	100
	45	100	12	22	25	43	65	70	75	88	88	98	92

Таблиця В.13. Точність розпізнавання Acc, %, шум у фойє залізничного вокзалу

	Номер частини експерименту	1	2										3
	SNR тестової вибірки, дБ	SNR навчальної вибірки, дБ											
		45, «чистий» сигнал	0	5	10	15	20	25	30	35	40	45	Усі SNR
Acc, %	0	20	35	47	43	10	10	10	10	10	10	10	55
	5	22	43	65	72	28	12	10	10	10	10	10	93
	10	23	43	68	87	82	60	12	13	10	10	10	100
	15	48	48	70	87	95	95	68	52	42	42	23	100
	20	78	48	68	87	95	97	97	93	78	70	67	100
	25	88	50	67	87	90	98	98	96	96	92	87	100
	30	98	40	60	72	83	93	98	98	98	100	95	100
	35	100	32	58	57	67	80	97	100	98	100	98	100
	40	100	13	53	42	47	58	78	98	100	100	100	100
	45	100	10	12	10	16	20	32	53	65	82	85	97

Таблиця В.14. Точність розпізнавання Acc, %, шум вулиці, вкладеної бруківкою

	Номер частини експерименту	1	2										3
	SNR тестової вибірки, дБ	SNR навчальної вибірки, дБ											
		45, «чистий» сигнал	0	5	10	15	20	25	30	35	40	45	Усі SNR
Acc, %	0	10	65	78	48	47	40	15	10	10	10	10	62
	5	12	63	90	90	85	62	50	28	15	13	12	95
	10	22	63	93	100	100	92	75	57	38	35	25	100
	15	47	47	78	95	100	100	98	90	63	65	52	100
	20	78	30	65	88	100	100	100	100	95	88	80	100
	25	88	15	58	88	97	98	100	100	100	98	97	100
	30	98	15	50	77	90	100	98	100	100	100	100	100
	35	100	13	30	72	87	95	100	100	100	100	100	100
	40	100	12	18	68	85	88	95	100	100	100	100	100
	45	100	10	12	32	50	72	85	85	92	95	97	88

Додаток Г

Перелік імпульсних характеристик приміщень та часу реверберації Т20

Таблиця Г.1. Перелік ІХ приміщень та часу реверберації Т20

Приміщення	Відстань «гучномовець-мікрофон», м	№ запису	Т20, с
1 Кімната для нарад	1.45	1, 2, 3, 4	0.3
	1.7	1, 2, 3, 4	0.3
	2.25	1, 2, 3, 4	0.3
	2.8	1, 2, 3, 4	0.3
2 Офісне приміщення	1	1,2	0.4
		3.4	0.5
	2	1, 2, 3, 4	0.6
	3	1, 2, 3, 4	0.6
3 Сходовий майданчик	1	1, 2	0.8
	2	1, 2	0.9
	3	1, 2	1
4 Лекційна аудиторія	2.25	1, 2	0.7
		3, 4	0.8
	4	1, 2, 3, 4	0.8
	5.56	1, 2	0.8
		3, 4	0.8
	7.1	1, 2, 3, 4	0.9
	8.68	1, 2, 3, 4	0.9
	10.2	1, 2, 3, 4	0.9

ДОДАТОК Ґ

Результати дослідження точності розпізнавання системи АРМ в умовах дії реверберaційної завади

Таблиця Ґ.1 – Результати експериментального дослідження робастності системи АРМ до дії реверберaційної завади при навчанні за методами Clean training. Reberb-matched training. All-training

T20 тестувальної вибірки, с / T20, с / метод навчальної вибірки	0	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Clean training	100	88.3	78.3	71.7	71.7	67.5	67.5	58.3	54.2
0.3	94.2	98.8	99.2	96.7	95.8	89.6	91.3	91.3	87.1
0.4	88.3	95.8	94.6	94.6	91.3	90	87.5	90.8	75
0.5	96.7	97.1	96.7	96.7	96.7	90	90	90.4	85.4
0.6	67.5	95	95.8	95.4	95.4	87.9	86.7	91.6	70.8
0.7	68.3	95	89.6	87.9	82.1	88.3	87.9	87.5	67.9
0.8	80	96.3	92.9	89.5	87.9	85.8	89.6	90	69.2
0.9	54.1	88.3	87.9	87.1	81.3	81.7	82.5	87.1	63.3
1	69.2	92.1	84.2	90	87.5	90	9038	89.6	89.5
All-training	91.7	93.3	89.2	87.5	90.8	91.7	89.2	87.5	83.3

Таблиця Ґ.2 – експериментального дослідження робастності системи АРМ до дії реверберaційної завади при навчанні за методом Room-training (офісне приміщення)

Офісне приміщення	Відстань «гучномовець - мікрофон» тестувальної вибірки, м		
Відстань «гучномовець -- мікрофон» навчальної вибірки, м / метод навчання	1.45	2.25	2.8
Clean training	75	72.9	69.2
1	95.1	92.8	90.1
2	95.9	94.5	91
3	91.6	92.3	91.1
Room-training	93.3	93.3	89.6
All-training	90.4	90.8	89

Таблиця Ґ.3 – експериментального дослідження робастності системи АРМ до дії реверберaційної завади при навчанні за методом Room-training (кімната для нарад)

Кімната для нарад	Відстань «гучномовець - мікрофон» тестувальної вибірки, м				
Відстань «гучномовець -- мікрофон» навчальної вибірки, м / метод навчання	1.45	1.7	1.9	2.25	2.8
Clean training	91.7	92.5	90.1	91.3	88.7
1.45	97.3	97.8	98.4	97.8	95.2
1.7	98.5	98.4	99	99.4	98.2
1.9	97	97.9	98.3	98	95.6
2.25	97.5	98.1	99	98.5	97.4
2.8	93.4	95.9	96.7	97.6	96.6
Room-training	98.3	98.3	98.3	98.3	98.3
All-training	92.1	92.5	92.5	95	92.5

Таблиця Ґ.4 – експериментального дослідження робастності системи АРМ до дії реверберaційної завади при навчанні за методом Room-training (сходовий майданчки)

Сходовий майданчик	Відстань «гучномовець - мікрофон» тестувальної вибірки, м		
Відстань «гучномовець -- мікрофон» навчальної вибірки, м / метод навчання	1	2	3
Clean training	67.5	59.2	52.5
1	89.5	90	88.3
2	88.3	89.2	88.8
3	83.7	88.3	88.3
Room-training	87.5	88.3	87.5
All-training	87.5	86.7	81.7

Таблиця Ґ.5 – Результати експериментального дослідження робастності системи АРМ до дії реверберaційної завади при навчанні за методом Room-training (лекційна аудиторія)

Лекційна аудиторія	Відстань «гучномовець - мікрофон» тестувальної вибірки, м					
Відстань «гучномовець - мікрофон» навчальної вибірки, м / метод навчання	2.25	4	5.56	7.1	8.68	10.2
Clean training	67.5	67.5	60	56.7	58.3	56.7
2.25	92.5	90.8	92.9	90	92	91.7
4	93.3	91.7	92.5	91.7	91.7	90.4
5.56	95.4	91.7	91.7	93.7	91.3	92.9
7.1	95	90.1	90	93.8	90.1	92.5
8.68	93.3	90	91.7	91.7	88.3	89.5
10.2	84.6	85	87.9	86.3	86.3	88.3
Room-training	90.8	91.7	88.3	90	88.3	88.3
All-training	86.7	87.5	86.7	86.7	85.8	86.7