Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет _ інформаційних радіотехнологій та технічного захисту інформації_
(повна назва)

Кафедра _ медіаінженерії та інформаційних радіоелектронних систем_____
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти _____другий (магістерський)_____

_____Методи придушення нестаціонарних шумів в звукозаписі.___ _____
_____
(тема)

Виконав:
студент _2_ курсу, групи ___МІм-22-1____
_____Захаров В.П._ _____
(прізвище, ініціали)

Спеціальність 172 Телекомунікації та_____
_____радіотехніка_____
(код і повна назва спеціальності)
Тип програми ___освітньо-професійна____
Освітня програма ____Медіаінженерія____
(повна назва освітньої програми)

Керівник ____доц. Шаповалов С.В.____
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ ___Володимир КАРТАШОВ__
(підпис)

2023 р.

Харківський національний університет радіоелектроніки

Факультет інформаційних радіотехнологій та технічного захисту інформації
Кафедра медіаінженерії та інформаційних радіоелектронних систем

Рівень вищої освіти другий (магістерський)

Спеціальність 172 Телекомунікації та радіотехніка
(код і повна назва)

Тип програми освітньо-професійна

Освітня програма Медіаінженерія
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«____»_____ 20 ___ р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

студенту Захарову Владиславу Павловичу
(прізвище, ім'я, по батькові)

1. Тема роботи Методи придушення нестаціонарних шумів в звукозаписі.

затверджена наказом по університету від "_20_"_____10_____ 2023 р. № ___1224 Ст____

2. Термін подання студентом роботи до екзаменаційної комісії _____10.01.2024 р._____

3. Вихідні дані до роботи Тип сигналу – мовний. Типи шуму – нестаціонарні змінні, переривчасті, імпульсні. Відношення сигнал-шум 20…24дБ. Провести аналіз шляхів зашумлення звукозаписів. Навести класифікацію шумів за статистичними, спектральними, часовими характеристиками. Оцінити складність задач подавлення шумів різного типу. Провести аналіз наукових робіт з дослідження аналітичних і нейромережевих методів подавлення шуму у звукозаписах. Проведести тестування програмних методів шумоочищення, підвищення якості та розбірливості мови: метода спектрального віднімання Noise Reduction та нейромережевого метода Adobe Podcast.

4. Перелік питань, що потрібно опрацювати в роботі
__Вступ
__1 Аналіз шляхів зашумлення звукозаписів та класифікація методів шумоочистки.
__2 Теоретичний аналіз цифрових методів шумоочищення, підвищення якості та розбірливості мови.
__3 Тестування програмних методів шумоочищення, підвищення якості та розбірливості мови.
__Висновки
__Перелік посилань
__Додатки

5. Перелік графічного матеріалу із зазначенням обов’язкових креслеників, схем, плакатів, комп'ютерних ілюстрацій

1. Постановка задачі (1 аркуш А4).
2. Класифікація джерел шуму (1 аркуш А4).
3. Піраміда складності задач шумоочищення (1 аркуш А4).
4. Методи подавлення шуму (1 аркуш А4).
5. Адаптивні компенсатори перешкод (1 аркуш А4).
6. Авторегресивна фільтрація (1 аркуш А4).
7. Прихована Марківська модель (1 аркуш А4).
8. Метод спектрального віднімання (1 аркуш А4).
9. Нейромережеве шумоподавлення (1 аркуш А4).
10. Підготовка тестових аудіозаписів (1 аркуш А4).
11. Мікшування аудіодоріжок (1 аркуш А4).
12. Шумоподавлювач Noise Reduction (2 аркуша А4).
13. Нейронна мережа Adobe Podcast (2 аркуша А4).
14. Спектр обробленого сигналу (1 аркуш А4).
15. Висновки (1 аркуш А4).

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1	Аналітичний огляд літератури	01.09.23–27.09.23	
2	Теоретичний аналіз методів шумоподавлення	28.09.23–11.10.23	
3	Підготовка аудіофрагментів	12.10.23–10.11.23	
4	Експериментальна частина	11.11.23–03.12.23	
5	Обробка результатів	04.12.23–17.12.23	
6	Графічна частина роботи	18.12.23–17.12.23	
7	Перевірка керівником	18.12.23–30.12.23	
8	Перевірка на академічний плагіат	02.01.24–05.01.24	
9	Перевірка завідувачем кафедри, рецензування	06.01.24–09.01.24	

Дата видачі завдання 20.10.2023 р.

Студент _____ Владислав ЗАХАРОВ
(підпис)

Керівник роботи _____ Сергій ШАПОВАЛОВ
(підпис)

РЕФЕРАТ

Пояснювальна записка до кваліфікаційної роботи: 69 сторінок, 35 рисунків, 1 таблиця, 47 джерел.

АКУСТИЧНИЙ ШУМ, МОВНИЙ СИГНАЛ, НЕЙРОННА МЕРЕЖА, РЕВЕРБЕРАЦІЯ, СПЕКТР, СПЕКТРОГРАМА, ХВИЛЬОФОРМА

Мета роботи – огляд, аналіз та тестування нових адаптивних методів та алгоритмів шумозаглушення для мовлення, визначення на основі цього аналізу та тестування найбільш ефективних алгоритмів детектування та придушення шуму.

В роботі проведено аналіз шляхів зашумлення звукозаписів. Наведено класифікацію шумів за статистичними, спектральними, часовими характеристиками. Складено піраміду складності задач подавлення шумів різного типу. Наведено загальну класифікацію методів подавлення шуму. Проведено теоретичний аналіз робіт з дослідження аналітичних і нейромережевих методів подавлення шуму у звукозаписах. Проведено тестування програмних методів шумоочищення, підвищення якості та розбірливості мови: метода спектрального віднімання Noise Reduction та нейромережевого метода Adobe Podcast.

ABSTRACT

Explanatory note to the qualification work: 69 pages, 35 figures, 1 table, 47 sources.

ACOUSTIC NOISE, SPEECH SIGNAL, NEURAL NETWORK, REVERBERATION, SPECTRUM, SPECTROGRAM, WAVEFORM

The purpose of the work is to review, analyze and test new adaptive methods and algorithms for noise suppression for speech, to determine based on this analysis and testing the most effective noise detection and suppression algorithms.

The paper analyzes ways of making sound recordings noisy. The classification of noises according to statistical, spectral, and temporal characteristics is presented. A pyramid of the complexity of noise suppression tasks of various types has been compiled. A general classification of noise suppression methods is provided. A theoretical analysis of works on the research of analytical and neural network methods of noise suppression in sound recordings was carried out. Testing of software methods of noise reduction, improving the quality and intelligibility of speech was carried out: the Noise Reduction spectral subtraction method and the Adobe Podcast neural network method.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

Adobe Podcast – веб-інструмент для запису та редагування звуку на основі штучного інтелекту;

Audacity — безкоштовний багатоплатформенний аудіоредактор звукових файлів, орієнтований на роботу з кількома дорожками;

Conv-TasNet – модифікація алгоритму TasNet, яка використовує у якості вагової функції згорткові шари з розширенням (dilation);

DEMUCS – Music source separation – нейронний алгоритм розділення музичних джерел;

Denoise – лінійна система шумоподавлення;

EM – Expectation Maximization – алгоритм максимізації математичного очікування;

HiFi-GAN – генеративна нейронна мережа для ефективної та високоякісної мови;

MSE – Mean Square Error – середньоквадратична помилка;

MMSE – мінімізація середньоквадратичної помилки;

PSNR – peak signal-to-noise ratio – пікове відношення сигнал-шум;

SNR – signal-to-noise ratio – відношення сигнал-шум;

TasNet – Time-Domain Audio Separation Network for Real-Time – нейронний метод одноканального розділення мови в часовій області;

TIMIT Acoustic-Phonetic Continuous Speech Corpus – стандартний набір даних, який використовується для оцінки систем автоматичного розпізнавання мовлення;

МІРЕС – медіаінженерія та радіоелектронні системи;

ПВКК – пристрій управління ваговими коефіцієнтами;

ЦОС – цифрова обробка сигналів.

ВСТУП

Широке використання телекомунікаційних та інформаційних мереж стимулює розробку алгоритмів цифрової обробки сигналів, зокрема алгоритмів цифрової обробки звуку. Це пов'язано з тим, що одним з основних видів даних, що передаються по мобільних мережах і мережах Інтернет, є мова.

На сьогоднішній день активно розвиваються наступні напрямки цифрової обробки звуку голосу:

– транскрибування – переклад мови в текст;

– діаризація – поділ вихідного аудіопотоку на окремі аудіопотоки по дикторах і розпізнавання дикторів;

– визначення мови мовлення;

– визначення емоційного забарвлення промови.

Для спрощення роботи даних високорівневих алгоритмів аналізу та розпізнавання мови, а також для поліпшення якості зв'язку між кінцевими абонентами голосового зв'язку вихідний звук піддається передобробці – шумоочищенню.

В даний час існує кілька класів методів шумоочищення та шумозаглушення, проте усі вони мають недоліки. Продуктивних і добре описаних методів шумоочищення голосу для систем реального часу (СРЧ), що накладають на використовувані в них алгоритми ряд суттєвих обмежень, недостатньо. Так, застосування смугових фільтрів і гейтів [1], що задовольняють вимогам СРЧ щодо часу відклику алгоритму і передбачуваності його роботи, виявляється марним при зміні у часі виду та напрямку шуму, а найбільш ефективні сучасні нейромережеві підходи [2, 3] або вносять у роботу СРЧ велику затримку, або потребують специфічного дорогого обладнання та (або) допоміжних даних.

Компромісним рішенням, що забезпечує прийнятні швидкість і якість шумоочищення, є розробка алгоритмів фільтрації, які передбачають, що

сигнали шуму та голосу некорельовані. Зазвичай такі алгоритми складаються із двох частин:

– детектора шуму;

– його безпосереднього подавлювача.

Мета даної роботи – огляд, аналіз та тестування нових адаптивних методів та алгоритмів шумозаглушення для мовлення, визначення на основі цього аналізу та тестування найбільш ефективних алгоритмів детектування та придушення шуму.

1 АНАЛІЗ ШЛЯХІВ ЗАШУМЛЕННЯ ЗВУКОЗАПИСІВ ТА КЛАСИФІКАЦІЯ МЕТОДІВ ШУМООЧИСТКИ

1.1 Поняття зашумлення і шумоочистки

Дамо формальні визначення понять "мова", "шум" і "процес шумоочищення" в термінах цифрової обробки сигналів (ЦОС).

Мова – корисний сигнал $s_1(t)$, що надходить на вхід лінійної системи шумоподавлення Denoise.

Шум – адитивний сигнал-перешкода $s_2(t)$, що надходить на вхід лінійної системи шумоподавлення Denoise.

Процес шумоочищення – лінійна система шумоподавлення Denoise, яка виконує наступне перетворення:

$$s_1(t) + s_2(t) \to s_1(t). \tag{1.1}$$

Як показує практика, створення систем, що точно задовольняють формулі (1.1), є нездійсненним завданням, тому часто завдання шумоочищення спрощують.

Детальний опис процесу шумоподавлення представимо наступною формулою:

$$s_1(t) + s_2(t) \to s_3(t) \cong s_1(t), \tag{1.2}$$

де $s_3(t)$ – сигнал, що є наближенням сигналу $s_1(t)$.

Отже, шумоочищення – це процес усунення шумів з корисного сигналу з метою підвищення його суб'єктивної якості. Методи шумоподавлення концептуально дуже схожі незалежно від сигналу, що обробляється, проте

попереднє знання характеристик сигналу, що передається, може значно вплинути на реалізацію цих методів залежно від типу сигналу.

Формалізований алгоритм шумоочищення показано на рис.1.1.

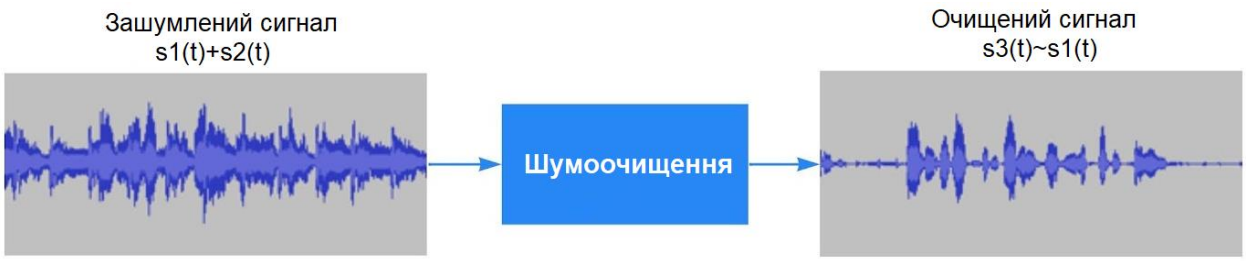


Рисунок 1.1 – Формалізований алгоритм шумоочищення

Зазначимо, що використання алгоритму шумоподавлення в СРЧ накладає на нього наступне суттєве обмеження: алгоритм повинен обробляти звукові дані блочно-послідовно, вносячи в них затримку, що залишається комфортною для слухачів. Дослідження показують, що максимально допустима затримка складає менше 400 мс [4].

1.2 Класифікація шумів в звукозаписі

Шум – хаотичні коливання різної фізичної природи, що відрізняються складністю часової та спектральної структури [5, 6]. Спочатку слово шум відносилося виключно до звукових коливань, проте в сучасній науці воно було поширене і на інші види коливань (радіо, електроніка).

Шум, незалежно від фізичної природи, відрізняється від періодичних коливань випадковою зміною миттєвих значень величин, що характеризують даний процес. Часто шум є сумішшю випадкових і періодичних коливань.

Для опису шуму застосовують різні математичні моделі відповідно до їх часової, спектральної та просторової структури. Для кількісної оцінки шуму користуються усередненими параметрами, що визначаються на підставі статистичних законів, що враховують структуру шуму в джерелі та властивості середовища, в якому поширюється шум.

Шуми поділяються на статистично стаціонарні та нестаціонарні [5, 6].

Стаціонарний шум характеризується сталістю середніх параметрів – інтенсивності (потужності), розподілу інтенсивності за спектром (спектральна щільність), автокореляційної функції (середнє за часом від миттєвих значень двох шумів, зрушених тимчасово затримки).

На практиці часто спостерігається шум, що виникає в результаті дії багатьох окремих незалежних джерел (наприклад, шум натовпу людей, моря, виробничих верстатів, шум вихрового повітряного потоку, шум на виході радіо і ін.), є квазістаціонарним.

Шум, що триває короткі проміжки часу (менше, ніж час усереднення у вимірниках), називається нестаціонарним. До нестаціонарних шумів відносять, наприклад [5, 6]:

– вуличний шум транспорту, що проходить,

– окремі стуки у виробничих або побутових умовах,

– рідкісні імпульсні перешкоди в радіотехніці і т.п.

Приклади часових реалізацій стаціонарного і нестаціонарного шуму показані на рис.1.2.

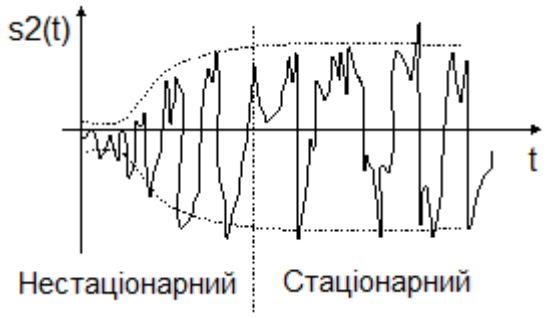


Рисунок 1.2 – Приклади часових реалізацій стаціонарного і нестаціонарного шуму

Нестаціонарний шум ділиться на переривчастий, змінний і імпульсний.

Переривчастий шум характеризується різким падінням рівня звуку рівня фонового шуму, причому тривалість інтервалів, протягом яких рівень

залишається постійним і перевищує рівень фонового шуму, становить 1 с і більше.

Змінний шум має рівень звуку, що безперервно змінюється в часі.

Імпульсний шум – це шумовий сигнал у вигляді окремих імпульсів тривалістю від 1 до 200 мс, або імпульсів, що йдуть один за одним в інтервалі більше 10 мс (але менше 1 с) і сприймається людським вухом як наступні один за одним удари.

Приклади часових реалізацій постійного, переривчастого, змінного і імпульсного шумів показані на рис. 1.3.

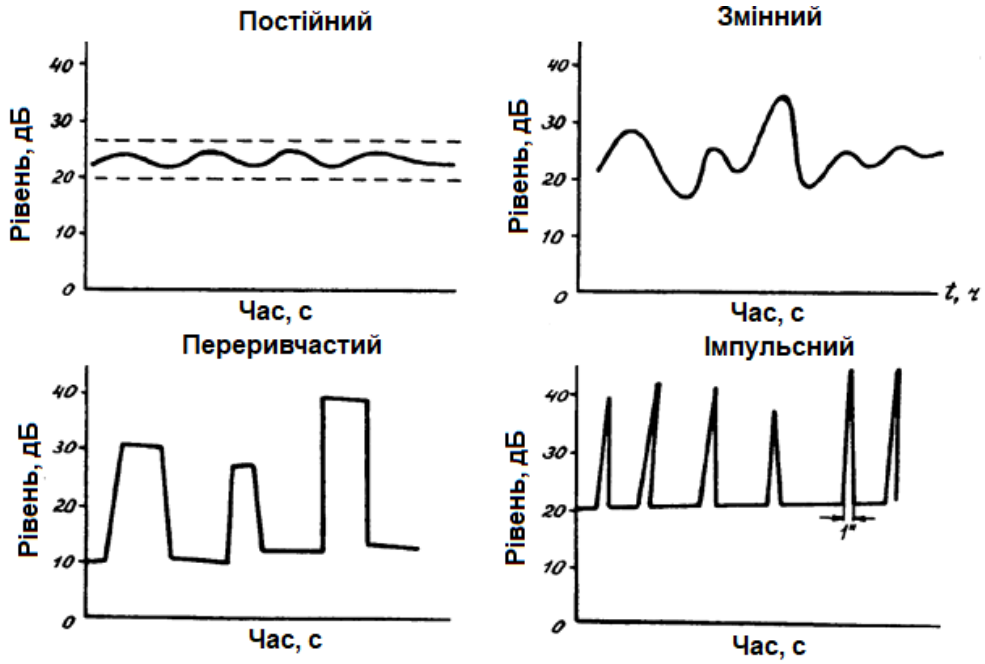


Рисунок 1.3 – Приклади часових реалізацій постійного, переривчастого, змінного і імпульсного шумів

За частотною характеристикою шуми поділяються на:

– низькочастотні (<300 Гц);

– середньочастотні (300-800 Гц);

– високочастотні (>800 Гц).

За характером спектра шуми поділяють на:

– широкосмуговий шум з безперервним спектром шириною понад 1 октаву;

– тональний шум, у діапазоні якого є виражені тони. Вираженим тон вважається тоді, якщо одна з третьоктавних смуг частот перевищує інші щонайменше, ніж 10 дБ.

Всі пристрої запису, як аналогові, так і цифрові, мають властивості, які роблять їх сприйнятливими до шуму. Шум може бути [5, 6]:

– випадковим і не когерентним, тобто не пов'язаний із самим сигналом,

– когерентним, що вноситься пристроями запису та алгоритмами обробки.

Будь-які аналогові схеми посилення та перетворення сигналів є джерелами шуму [5, 6]:

– по-перше, це тепловий шум, який викликаний тепловими процесами, що впливають на напрямок руху електронів;

– по-друге, це дробовий шум, причиною якого є дискретність носіїв електричного заряду – електронів, іонів.

Дослідження шуму переслідує різноманітні цілі – вивчення джерел шуму зменшення їх шкідливого на людини і різні системи; пошук способів та засобів найкращого (оптимального) прийому, виявлення та вимірювання параметрів різних сигналів у присутності шуму; підвищення точності вимірювань в аналогових та цифрових пристроях обробки інформації та ін.

Для вимірювання характеристик шуму застосовуються шумоміри, частотні аналізатори, корелометри та ін.

Для оцінки рівня мінливих шумів використовується так званий еквівалентний рівень звуку. Еквівалентний рівень звуку даного непостійного шуму чисельно дорівнює рівню звуку постійного, широкосмугового, неімпульсного шуму, що має такий самий вплив на людину, як і постійний шум. При вимірюваннях за допомогою шумоміра еквівалентний рівень шуму визначають за формулою [7]:

$$L_e = 10 \lg \frac{1}{T} \sum_{i=1}^{m} t_i 10^{L_i/10}, \qquad (1.3)$$

де T – час усереднення,

m – число вимірювань,

L_i – результат окремого вимірювання,

t_i – інтервал часу між вимірюваннями.

Зазвичай інтервал між вимірюваннями t_i=2…3 с, а час усереднення вибирають залежно від шуму.

1.3 Класифікація складності задач шумоочищення

Системи шумоочищення – системи обробки сигналу, реалізовані у вигляді електронних схем чи програмних алгоритмів, призначені збільшення відношення сигнал/шум.

Більшість систем шумоочищення ділиться на два типи:

– фільтрування, коли система обробляє сигнал під час прийому (відтворення) чи запису (передачі) намагаючись очистити корисний сигнал від шуму;

– системи, що модифікують сигнал для передачі по шумних каналах (або для запису сигналу на носій), з подальшим зворотним перетворенням на приймальній стороні (при відтворенні).

Як показано в попередньому підрозділі, існує безліч різних класифікацій шумів, наприклад, за характером спектра або частотою хвиль. Однак, коли ми хочемо позбавитися шумів у записі мови, варто в першу чергу враховувати категоризацію шумів за часовими характеристиками. Наведемо наочно таку класифікацію на рис. 1.4.

Як можна помітити, часові характеристики шуму тісно пов'язані зі способом утворення шуму: стаціонарний шум або змінний шуми, як правило, утворені якимись постійними процесами (природними або штучними), тоді як переривчастий і імпульсний – різкими одноразовими процесами. Переривчастий шум для простоти можна сприймати як імпульсний шум, що повторюється з деякою періодичністю.

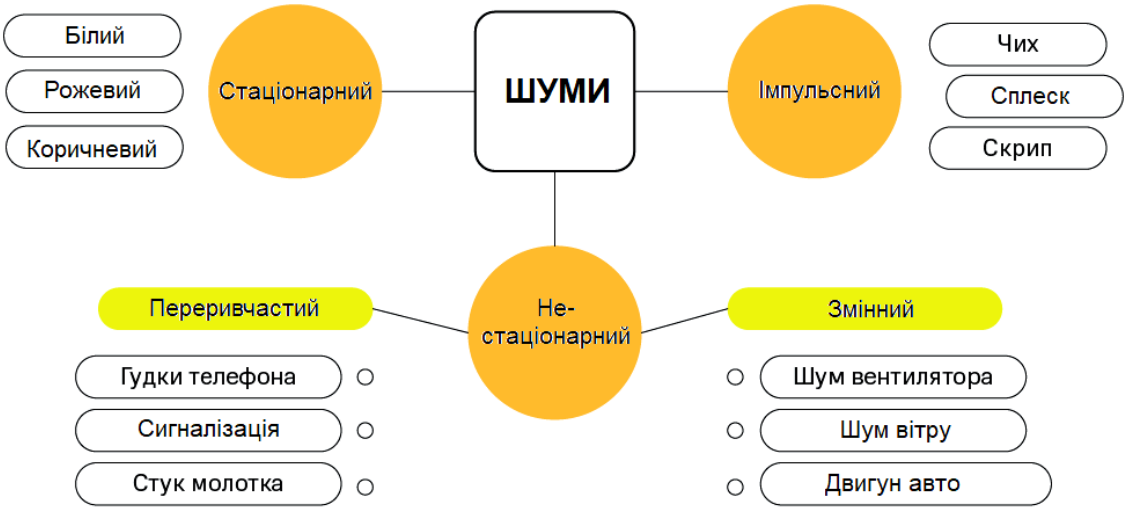


Рисунок 1.4 – Категоризація шумів за часовими характеристиками

Категорії шуму наведені у тому, щоб розмежувати шуми за складністю їх придушення.

Складність завдання шумоподавлення у непередбачуваності шумів, які можуть виникнути звуковому сигналі. Ми можемо з відносною легкістю прибрати шум із сигналу, якщо нам наперед відомо, який вид шуму знаходиться в цьому сигналі і де. Крім того, нам досить легко позбавлятися стаціонарного шуму, тому що ми легко можемо визначити поріг гучності в спектрі, так як білий шум буде рівномірно розподілений по всьому сигналу, і у фрагментах тиші ми будемо чітко спостерігати амплітуди шуму. Можна побудувати наступну піраміду складності задач, наведену на рис. 1.5.

Рисунок 1.5 – Піраміда складності задач шумоочищення

Якщо задачі нагорі піраміди можна вирішити обчислювальними методами, то задачі в нижній частині піраміди можна вирішити лише методами машинного навчання. Якщо обчислювальні методи вирішують задачі позбавлення сигналу певного шуму, то нейромережеві методи навчаються вирішувати задачі виділення лише релевантної мовної інформації з усього аудіопотоку. Схематично роботу методів вирішення задач шумоочищення нагорі і знизу піраміди складності показано на рис.1.6 [9].

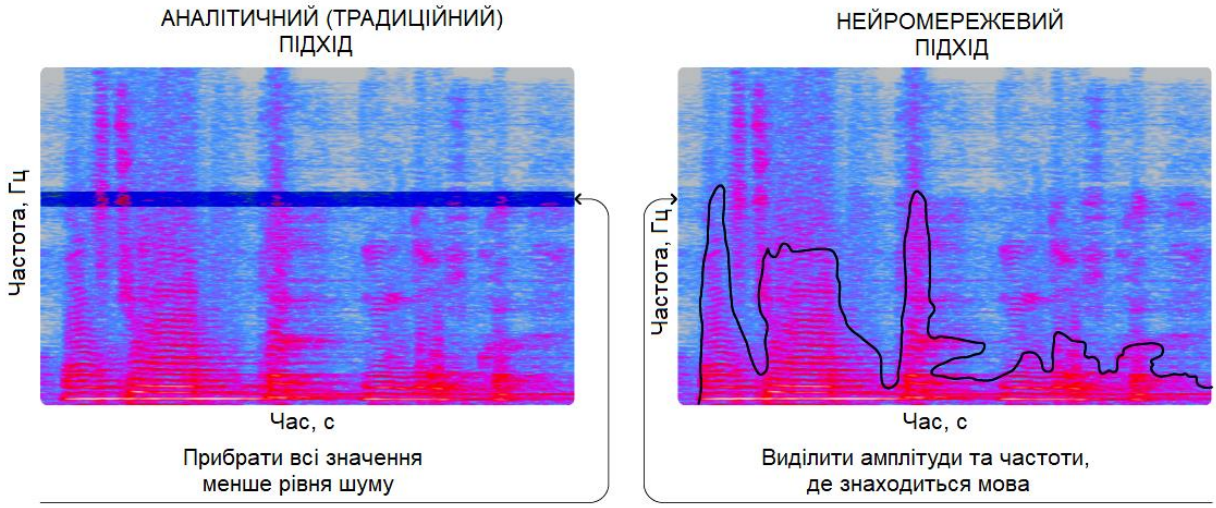


Рисунок 1.6 – Принцип роботи методів вирішення задач шумоочищення нагорі (а) і знизу (б) піраміди складності

Тепер докладно розберемо як традиційні [8], так і передові [9-13] методи шумоподавлення в аудіо.

1.4 Класифікація методів та систем шумоочищення

1.4.1 Традиційні методи шумоочищення

Найпростіші традиційні методи шумоподавлення використовуються в умовах, коли ми програмно не знаємо, який характер шуму та мови. Така відсутність апріорної інформації також спостерігається, коли ми хочемо

позбавлятися шуму на льоту. При такому шумоочищенні використовуються звичайні або спектральні пороги – заглушуються будь-які відгуки, якщо вони не перевищують певного порогу гучності.

В основі інших традиційних методів лежить моделювання розподілу чистої мови чи шуму. Робиться це за допомогою знаходження спектральної густини потужності (гучності) сигналу.

Спектральна щільність – базується на перетворенні Фур'є уявлення залежних від часу сигналів (як детермінованих, так і випадкових процесів) у вигляді спектрів.

Якщо процес $s(t)$ має кінцеву енергію і квадратично інтегруємий (а це нестаціонарний процес), то для однієї реалізації процесу можна визначити перетворення Фур'є як випадкову комплексну функцію частоти f [6]:

$$S(f) = \int_{-\infty}^{\infty} s(t) e^{-j2\pi ft} dt . \qquad (1.4)$$

Однак вона виявляється майже марною для опису ансамблю. Виходом із цієї ситуації є відкидання деяких параметрів спектра, а саме спектру фаз, та побудова функції, що характеризує розподіл енергії процесу по осі частот.

Функція $P(f) = |S(f)|^2$ характеризує, таким чином, розподіл енергії реалізації по осі частот і називається спектральною щільністю реалізації. Усереднивши цю функцію за всіма реалізаціями можна отримати спектральну щільність процесу.

Отже, щільність потужності сигналу – варіант опису розподілу значень сигналу різні моменти часу. Спектральна щільність потужності сигналу, своєю чергою, – функція, що визначає розподіл потужності сигналу залежно від частоти, саме – можливу потужність у різні одиниці частоти. У такому разі, маючи спектральну щільність потужності шуму, можна використовувати метод спектрального віднімання (spectral subtraction).

На рис.1.7 показані приклади спектральної щільності потужності шуму (а) і спектральної щільності потужності мови (б) [7].

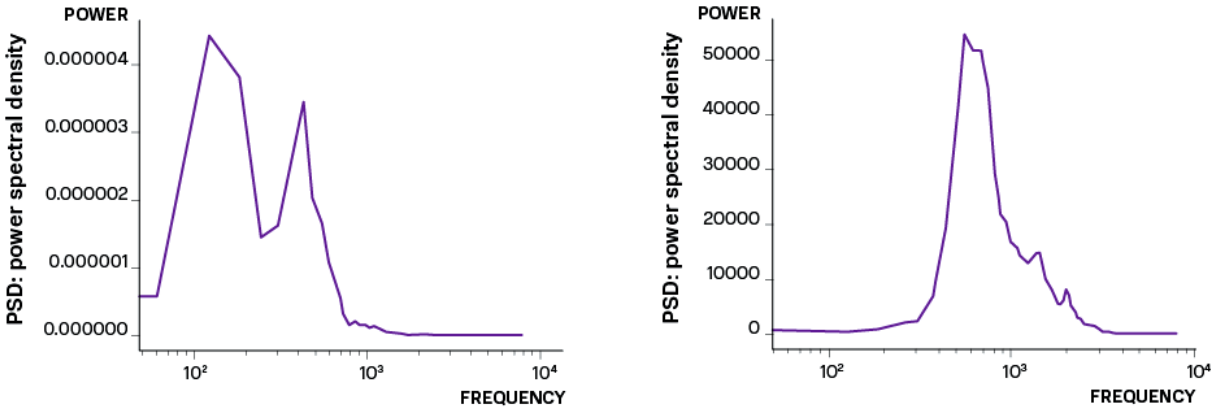


Рисунок 1.7 – Приклади спектральної щільності потужності шуму (а) і мови (б)

Вінерівське оцінювання (Wiener filter) використовується як один із традиційних навчальних способів шумоподавлення, частково схожий на метод спектрального віднімання. Цей підхід заснований на оптимальному підборі такого фільтра, який мінімізував би різницю між чистим сигналом і покращеним сигналом. Подібно до деяких алгоритмів машинного навчання, при обчисленні вінерівського фільтра мінімізується метрика Mean Square Error (MSE) [8].

$$H(f) = \frac{P_s(f)}{P_y(f)} = \frac{P_y(f) - P_d(f)}{P_y(f)}, \tag{1.5}$$

де $P_s(f)$ – спектр чистого сигналу,

$P_y(f)$ – спектр зашумленого сигналу,

$P_d(f)$ – спектр шуму.

Таким чином, оптимальний вінерівський фільтр можна знайти у випадках, коли нам відома «чиста версія» зашумленого сигналу, або якщо

нам відомий конкретний шум, який зустрічається в аудіозаписах і який ми хочемо прибрати.

Найчастіше після операцій з фільтрації шуму застосовується згладжування, щоб позбавитися артефактів сигналу – "музичного" шуму – після чищення. Для згладжування застосовуються різні фільтри, наприклад, гаусовий фільтр (або розмиття по Гаусу) [9].

1.4.2 Нейромережеві методи шумоподавлення

Майже всі сучасні нейромережеві алгоритми використовуються як для розмежування спікерів або інструментів, так і для шумоподавлення. При шумоподавленні важливо зазначити, що шум і чиста мова – два незалежні процеси, які виникають одночасно у часі, як два окремі інструменти в музичній композиції.

Залежно від способу вирішення задачі шумоподавлення, розмежування спікерів або покращення сигналу алгоритми машинного навчання можна розділити на дві категорії:

– на основі масок;

– генеративні.

Описання і приклади даних категорій нейромережевих алгоритмів наведені у табл. 1.1.

Таблиця 1.1 – Описання і приклади категорій нейромережевих алгоритмів обробки звуку

	На основі масок	Генеративні
Опис	Передбачають маски для кожного спікера/інструмента чи чистого сигналу. Ці маски накладаються на оригінальний сигнал.	Передбачають новий сигнал для кожного спікера/інструмента чи чистий сигнал.
Приклади	Conv-TasNet	DEMUCS Wave-U-Net HiFiGAN

До описаних вище нейромережевих підходів використовувалися нейромережеві методи накладання масок на спектрограму разом із прямим і зворотним перетвореннями Фур'є. Приклад методу накладання масок на спектрограми показано на рис. 1.8 [10].

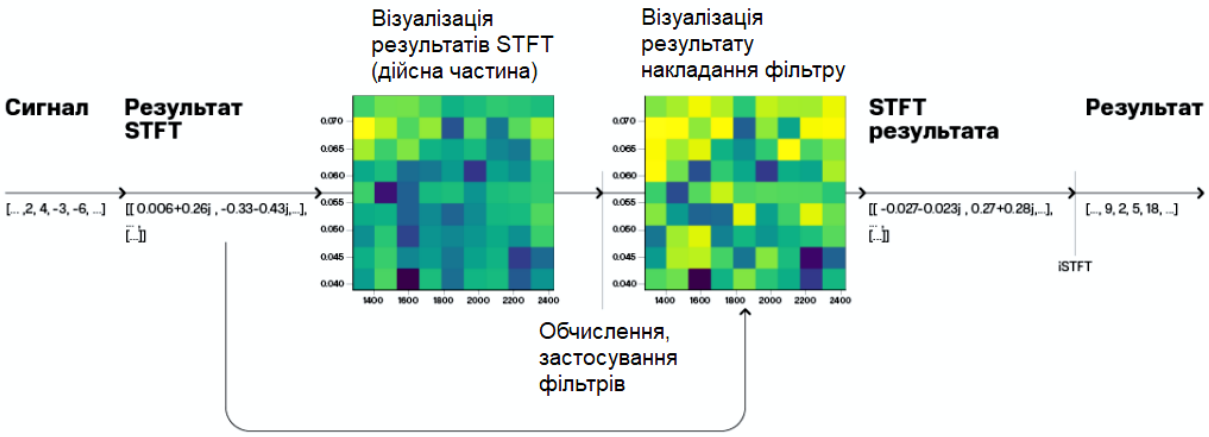


Рисунок 1.8 – Приклад методу накладання масок на спектрограми

Однак підходи, які ґрунтуються на маскування спектрограм, мають деякі недоліки. Наприклад, фаза хвилі в чистому сигналі може відрізнятись від фази хвилі в зашумленому сигналі. Тому навіть при обчисленні ідеальної маски для спектрограми відновлена з брудного сигналу фаза може вносити якісь елементи шуму і псувати підсумкову якість шумоподавлення.

1.5 Висновки по розділу 1

В результаті аналізу шляхів зашумлення звукозаписів було зроблено наступне.

1. Вияснено, що головними причинами зашумлення є акустичні або електронні шуми.

2. Наведено класифікацію шумів за статистичними, спектральними, часовими характеристиками. По цим ознакам існує безліч різних варіантів шумів. Однак, коли ми хочемо позбавитися шумів у записі мови, варто в

першу чергу враховувати категоризацію шумів за часовими характеристиками.

3. Проведено класифікацію шумі шумів за часовими характеристиками. Виділено:

– стаціонарні шуми,

– змінні у часі шуми,

– переривчасті шуми,

– імпульсні шуми.

4. Складено піраміду складності задач подавлення зазначених шумів. Серед нестаціонарних найскладнішим є випадок імпульсного шуму, оскільки він має найвищу ентропію.

5. Наведено загальну класифікацію методів подавлення шуму. Умовно методи розділено на аналітичні (традиційні), і передові – нейромережеві методи.

6. Зазначено загальні принципи, можливості, переваги і недоліки аналітичних (традиційних) і нейромережевих (передових) методів.

В даний час в програмах з обробки звуку або у веб-додатках реалізовано багато методів подавлення нестаціонарного шуму.

Об’єкт дослідження – процес очищення звукозапису мови від нестаціонарних шумів.

Предмет дослідження – методи очищення звукозапису мови від нестаціонарних шумів з метою наближення результату очищення до оригіналу як за об’єктивними характеристиками, так і за сприйняттям на слух.

Мета кваліфікаційної роботи – огляд, аналіз та тестування нових адаптивних методів та алгоритмів шумозаглушення для мовлення, визначення на основі цього аналізу та тестування найбільш ефективних алгоритмів детектування та придушення шуму в певних умовах.

Кваліфікаційна робота виконується на кафедрі МІРЕС ХНУРЕ. На кафедрі проводяться дослідження в таких наукових областях, як виявлення та

розпізнавання БПЛА за результатами акустичного спостереження [14-17], створення систем зондування атмосфери за допомогою акустичних хвиль [18-21]. Цілий ряд студентських доповідей [22-25] і атестаційних робіт магістрів минулих років [26-28] присвячені дослідженню систем звукозапису. Отже, дослідження в даній роботі пов’язані і ґрунтуються на традиційному напрямку робіт колективу і студентів кафедри МІРЕС.

2 ТЕОРЕТИЧНИЙ АНАЛІЗ ЦИФРОВИХ МЕТОДІВ ШУМООЧИЩЕННЯ, ПІДВИЩЕННЯ ЯКОСТІ ТА РОЗБІРЛИВОСТІ МОВИ

2.1 Постановка задачі

Мовні сигнали, з якими доводиться мати справу на практиці, завжди тією чи іншою мірою зашумлені. У тих випадках, коли шум має значну інтенсивність, його наявність може суттєво спотворити результати обробки, аналізу чи розпізнавання мови. У низці інших випадків, наприклад, при аналізі зашумлених записів у криміналістичних цілях чи відновленні аудіозаписів в архівах, завдання очищення сигналу від шуму носить самостійний характер і єдиною метою роботи. Тому вибір методів очищення сигналу від шуму є дуже актуальним напрямом досліджень.

На сьогодні розроблено дуже велику кількість різних методів цифрової обробки зашумлених мовних сигналів. Основним типом шумів для методів, розглянутих в даному аналізі, є адитивний шум.

З метою упорядкування аналізу методів очищення сигналу від шуму доцільно провести їхню класифікацію. Основною ознакою, за якою будуть класифікуватися алгоритми, є характер або тип тих закономірностей, які є основою виділення мовного сигналу з суміші із шумом. Як допоміжна ознака буде використовуватися класифікація на кшталт того математичного чи алгоритмічного апарату, який використаний для фільтрації. Подібна класифікація, звичайно, дуже умовна, оскільки багато з цих методів не можна беззастережно віднести до будь-якої однієї категорії. Як правило, одні й ті самі методи використовують одночасно різні принципи, і в цьому випадку можна говорити лише про переважний вплив будь-якої концепції.

З урахуванням зробленого зауваження можна виділити такі групи методів цифрової обробки зашумлених мовних сигналів:

– методи адаптивної компенсації перешкод;

– методи, що ґрунтуються на використанні математичних моделей

мовних сигналів у часовій області (наприклад, авторегресійна модель мовного сигналу та рекурентні алгоритми оцінки параметрів та мовного сигналу);

– методи, що ґрунтуються на використанні математичних моделей мовних сигналів у частотній області (оцінювання мінімальної середньоквадратичної помилки, марківські моделі сигналу та шуму);

– методи, що базуються на використанні спектральних характеристик шуму (віднімання амплітудних спектрів, вінерівська фільтрація);

– методи, засновані на використанні моделей штучних нейронів мереж;

– методи, що базуються на моделях сприйняття мови людиною.

Перейдемо до розгляду конкретних методів цифрової обробки зашумлених мовних сигналів.

2.2 Адаптивні компенсатори шумів

Цей клас методів цифрової обробки зашумлених сигналів заснований на використанні, крім власне зашумленого сигналу, що підлягає очищенню, а також одного або декількох опорних сигналів – сигналів, які корельовані з шумовим сигналом і некорельовані (або слабко корельовані) з корисним сигналом, що підлягає виділенню.

За допомогою опорних сигналів формується сигнал, що є оцінкою перешкоди. Цей сигнал потім віднімається із зашумленого сигналу і результат цієї операції розглядається як оцінка незашумленого сигналу. На рис. 2.1 представлено схему адаптивного компенсатора перешкод, який використовує один опорний сигнал [29].

На рис. 2.1 введені наступні позначення: $u(n)$ – дискретний відлік корисного сигналу на момент часу n, n=0,1,2…; $e(n)$ – шумовий сигнал, $e_1(n)$ – опорний сигнал, $\varepsilon(n)$ – сигнал помилки, $u_1(n)$ – вихідний сигнал компенсатора, ПКВК – пристрій керування ваговими коефіцієнтами.

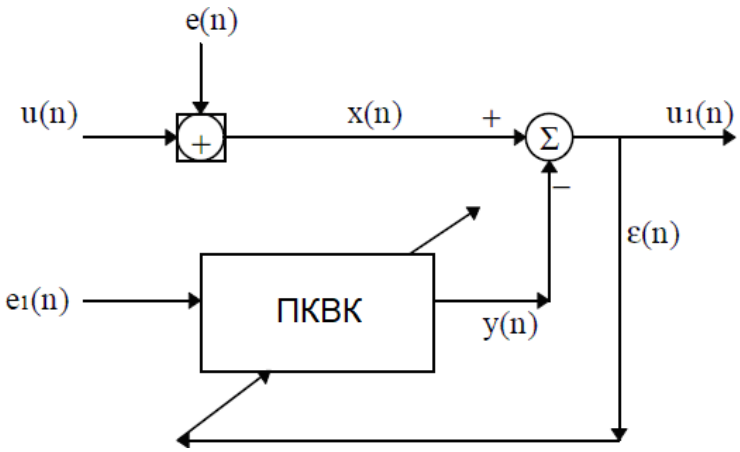


Рисунок 2.1 – Схема адаптивного компенсатора перешкод

Найбільш важливою частиною адаптивного компенсатора перешкод є пристрій управління ваговими коефіцієнтами – лінійний фільтр, через який пропускається опорний сигнал $e_1(n)$. Завдання адаптивної компенсації перешкоди $e(n)$ зводитись до підбору коефіцієнтів фільтра таким чином, щоб мінімізувати енергію сигналу на виході компенсатора $u_1(n)$. У цьому випадку буде максимізовано вихідне відношення сигнал/шум.

Практична схема реалізації адаптивного метода шумоподавлення показана на рис. 2.2.

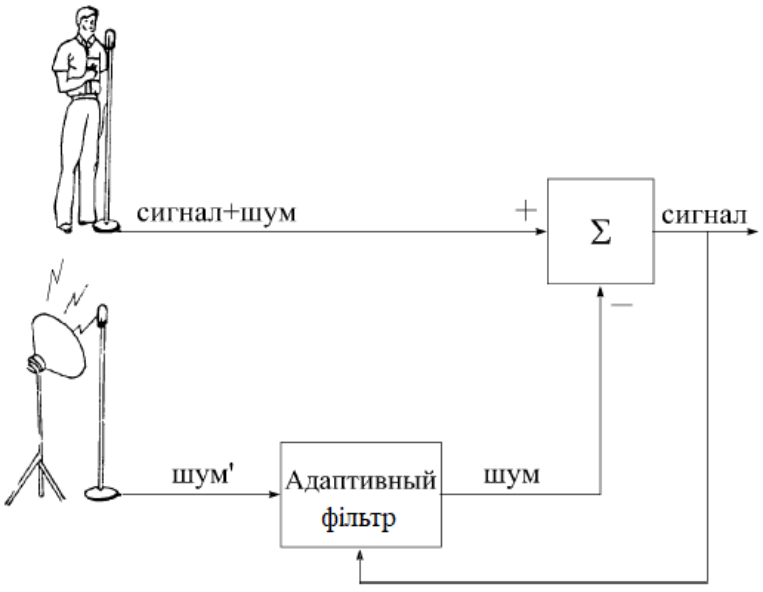


Рисунок 2.2 – Практична схема реалізації адаптивного метода шумоподавлення

Мінімізація енергії зазвичай здійснюється на основі градієнтних методів пошуку екстремуму функцій багатьох змінних [30].

Якщо адаптивний фільтр розглядати як трансверсальний, його вихідний сигнал визначається виразом:

$$y[n] = \sum_{k=1}^{N} W_k[n-1] \cdot x[n-k+1] = W_N^T(n-1) \cdot X_N(k), \qquad (2.1)$$

де $X_N(m) = \{x[n], x[n-1], ..., x[k-n+1], ..., x[k-N+1]\}$ – вектор відліків вхідного сигналу;

$W_N(n-1) = \{w_1[n-1], w_2[n-1], ..., w_k[n-1], ..., w_N[n-1]\}$ – вектор вагових коефіцієнтів.

Оптимальне оцінювання ґрунтується на мінімізації цільової функції середньоквадратичної помилки

$$\xi = e_2[n]. \qquad (2.2)$$

Значення вектора вагових коефіцієнтів, що відповідає мінімуму ξ, знаходять відповідно до рівняння Вінера-Хопфа:

$$W_N = R_N^{-1} \cdot r_N, \qquad (2.3)$$

де R_N – автокореляційна матриця вхідного сигналу, кожен елемент якої є добутком двох її відліків;

r_N – кореляційний вектор зразкового та вхідного сигналів, кожен елемент якого є добутком поточного відліку сигналу d[n] і одного з елементів X_N.

Відомо, що адаптивні компенсатори перешкод дозволяють значно покращити якість зашумлених сигналів – на кілька десятків децибелів [31],

але вимога наявності опорного сигналу суттєво звужує їх галузь застосування. У багатьох додатках цифрової обробки мовних сигналів (наприклад, під час реставрації архівних записів або в криміналістиці), опорного сигналу, принаймні, у явному вигляді немає. Тому для застосування методів адаптивної компенсації перешкод опорний сигнал у таких випадках доводиться отримувати на основі непрямих міркувань, пов'язаних з особливостями мовного сигналу, а сам адаптивний компенсатор у цьому випадку буде бути однією з складових частин складнішого алгоритму виділення мовного сигналу.

2.3 Методи на основі статистичних моделей мовних сигналів у часовій області

Завдання виділення мовного сигналу із суміші з шумом у разі використання достатньо адекватної моделі зводиться до оцінки якимось чином параметрів цієї моделі і наступним синтезом або фільтрацією мовного сигналу фільтром, побудованим на основі чи за допомогою оцінених параметрів.

Одними з найбільш перспективних методів у цьому класі є методи статистичної фільтрації у часовій області, які розвивалися в роботах [32]. Фільтрування мовного сигналу, моделюється авторегресією, здійснюється при цьому методами теорії оптимального оцінювання, наприклад, за допомогою побудови оптимального лінійного фільтра (фільтра Калмана [33]).

Припустимо, що деяка лінійна система зі змінними параметрами збуджується шумовим сигналом w(k), де k – індекс, відповідний дискретному часу. Співвідношення між вихідним сигналом системи x(k) (вектором стану) та сигналом збудження w(k) в момент часу k =1 матиме вигляд

$$x(k) = F(k+1, k)\, x(k) + G(k)w(k). \qquad (2.4)$$

У виразі (2.4) передбачається, що сигнали x і w – векторні величини, компоненти є випадковими величинами. Матриці F(k+1,k) та G(k) характеризують стан системи у відповідні моменти часу.

Допустимо далі, що вектор стану невідомий і потрібно зробити його оцінку за спостережуваними (до моменту часу k включно) величин z(k) (спостережень), які пов'язані з вектором стану x(k) співвідношенням:

$$z(k) = H(k)x(k)+v(k), \qquad (2.5)$$

де v(k) – шум, який потрібно відфільтрувати.

Якщо задані матриці F(i,i+1), G(i), H(i), 0<=i<=k, визначено статистичні властивості шумів w,v і вказані відповідні початкові умови у нульовий момент часу: x(0), то оптимальна, за критерієм мінімуму дисперсії помилки, лінійна оцінка вектора стану x(i) за спостереженнями, z(1), z(2),…, z(i), для процесу, що описується співвідношеннями (3.1), (3.2), дається в рекурентному вигляді алгоритмом фільтрації Калмана [33]:

$$x(i) = F(i+1,i)*x(i-1)+K(i)[z(i)-H(i)f(i,i-1)x(i-1)]. \qquad (2.6)$$

Однією з найпоширеніших моделей мовних сигналів є модель авторегресії, або її еквіваленти [32].

Відповідно до цієї моделі мовний сигнал {x(n)}, n=..-1,0,1,... описується рівнянням авторегресії:

$$s(n) = \Sigma_{k=1}\, pa(k)s(n-k) + be(n-1). \qquad (2.7)$$

де e(n) – послідовність некорельованих випадкових величин, таких, що E(e(n)), E(E2(n))=1, n=…-1,0,1,…;

a(k) – параметри моделі,

b – постійний коефіцієнт. Розмір p називається порядком моделі.

Відповідно до гіпотези про локальну сталість параметрів, параметри моделі авторегресії зазвичай вважаються постійними протягом малих проміжків часу (10-20 мс), або якимось чином задається їх закон зміни.

Модель (2.7) є лінійною моделлю, передатна функція якої містить лише полюси. Для опису деяких звуків мови, наприклад, назальних, "м". "н", найбільш адекватною є лінійна модель авторегресії зі ковзним середнім:

$$s(n) = \sum_{k=1}^{p} a(k)s(n-k) + \sum_{m=0}^{q} b(m)e(n-m) \qquad (2.8)$$

Передатна функція цієї моделі містить як полюси, так і нулі. Тому для покращення якості мовних сигналів часто вигідніше використовувати саме цю модель. Порівняння кількох методів моделювання зашумленого мовного сигналу на основі моделі (2.8) показало. що виграш щодо сигнал/шум у цьому випадку приблизно на 5 дБ перевищує аналогічний виграш при використанні тих самих методів фільтрації, але для авторегресійної моделі мовного сигналу.

Обчислювально ефективна (але з менш вдалим результатом обробки) реалізація алгоритму фільтрації мовного сигналу, що моделюється авторегресійною моделлю з параметрами, пов'язаними в марківське коло, запропонована в [34]. Спільна оцінка сигналу та параметрів марківського кола обчислюються рекурентним способом за допомогою алгоритму максимізації математичного очікування (expectation maximization - EM), причому для обчислення умовного очікування (expectation step) сигналу щодо спостережень використаний фільтр Калмана-Бьюсі.

Експериментальні випробування на мовному сигналі в суміші з некорельованим адитивним білим шумом з відношенням сигнал/шум 0, 10 і 20 дБ показали збільшення відношення сигнал/шум у середньому на 4 дБ.

Авторегресійна модель мовного сигналу, як показує практика, не має такого вираженого дефекту як музичні тони, проте, артефакти обробки також мають місце.

2.4 Методи на основі обробки мовного сигналу з використанням апарату прихованих марківських моделей

Іншим класом методів обробки зашумлених мовних сигналів заснованих на використанні статистичних моделей мовного сигналу є методи, у яких мовний сигнал моделюється прихованим марківським колом. Тобто, для моделювання мовного сигналу використаний найбільш ефективний для розпізнавання мови підхід.

Відомо, що традиційні методи фільтрації (віднімання спектрів або фільтр Вінера) не використовують фонетичну інформацію, переносиму мовним сигналом. Дослідження [35] показали, що знання та застосування в процесі обробки фонетичної структур сигналу призводить до покращення якості фільтрації. Тому цілком природним є застосування в процесі очищення мовного сигналу від шумів його статистичної моделі у вигляді прихованого марківського кола, який пов'язаний фонетичною структурою сигналу.

Ідея реалізації такого підходу полягає у тому. що спочатку, по записам незашумленого мовного сигналу будуються статистичні моделі одиниць мовного потоку (фонів чи ширших класів звуків). Після того як статистична модель для множини станів сигналу побудована, за нею можна розрахувати оптимальний фільтр Вінера.

При обробці зашумленого мовного сигналу спочатку оцінюється (за фільтрованим на попередньому кроці сигналом) поточний стан харківської моделі, відповідно до якої вибирається оптимальний фільтр, який потім використовується для фільтрації сигналу та отримання чергової оцінки.

Алгоритм фільтрації виглядає так, як показано на рис.2.3 [36]. У дослідженнях [36] спочатку, використовуючи стандартну базу даних (марківські моделі формувалися на TIMIT) будувалися моделі станів для незашумленого мовного сигналу. Для кожного стану моделі і зкожною гауссівською складовою (реально використовувалася лише одна складова стану) оцінювався оптимальний фільтр Вінера $H\beta(\theta)$.

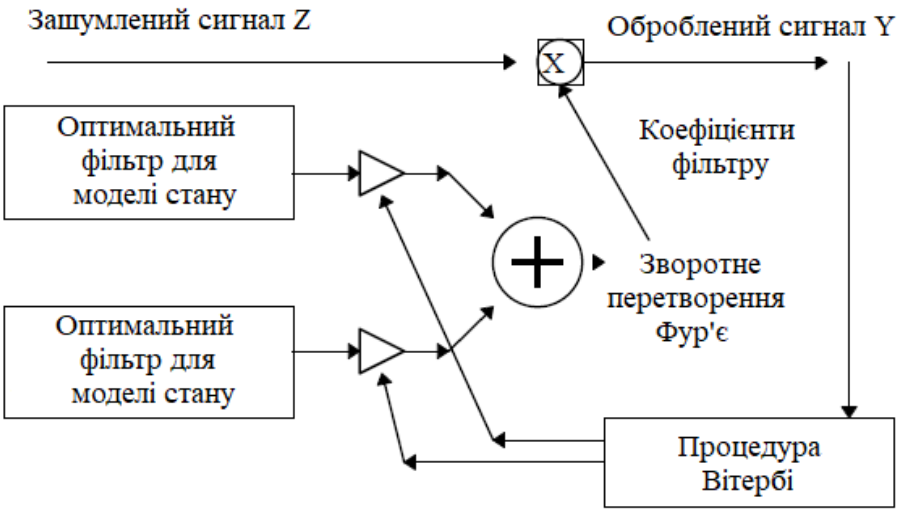


Рисунок 2.3 – Алгоритм фільтрації з використанням апарату прихованих марківських моделей

У процесі обробки сигналу на кожному кроці (кадрі аналізу) за допомогою процедури Вітербі обчислювалися правдоподібності станів відповідно до яких вибиралися вагові коефіцієнти W для кожного фільтра Hβ(θ). Очищення сигналу потім проводилася в частотній області відповідно до виразу

$$y\ (iw,k) = \left[\sum_{\beta} W_{\beta}(k) H_{\beta}^{-1}(iw)\right]^{-1} z(iw,k)\ , \qquad (2.9)$$

де k – номер ітерації (кадра аналізу),

W – вага фільтра H (θ),

z (iw) – спектральна компонента зашумленого сигналу,

y(iw) – спектральна компонента обробленого сигналу.

У реальних експериментах з фільтрації мовного сигналу описаними методами число станів марківської моделі мовного сигналу вибиралося рівним 5, тобто фактично мовний сигнал моделювався як послідовність фонів, що відповідають широким фонетичним категоріям (дзвінкий-галасливий-глухий і т.п.).

Порівняльні дослідження виконувались на різних типах шумів (білий, симуляція шуму гелікоптером, одночасна розмова декількох дикторів)

Об'єктивні вимірювання (зміна відношення сигнал/шум на вході та виході системи) показали очевидну перевагу описаної методики (у середньому покращення +2.5 дБ, в діапазоні від 0 до +20 дБ, причому при відношенні сигнал/шум > 10 дБ. перевага склала в середньому +7 дБ) над стандартною системою фільтрації, побудованою за методом віднімання амплітудних спектрів.

Суб'єктивні тести (оцінка якості звучання обробленого сигналу на слух за п'ятибальною шкалою) також показали перевагу марківських моделей над стандартними методиками. На думку авторів [36], розбірливість мови в результаті обробки також підвищилася, ймовірно внаслідок того, що низькоенергетичні звуки в даному випадку обробляються значно акуратніше.

2.5 Методи, на основі оцінки спектральних характеристик шуму

Найпоширенішими методами, заснованими на використанні спектральних характеристик шуму, є методи, що реалізують різні модифікації алгоритму віднімання амплітудних спектрів [37].

Як обґрунтування цих методів наводяться такі міркування. Якщо стаціонарний сигнал s(t), t=…-1,0,1,.. із спектральною щільністю потужності Pss(iw) спотворений адитивним стаціонарним шумом n(t) зі спектральною щільністю потужності Pnn(iw), який передбачається некорельованим з x(t), то спектральна щільність потужності зашумленого сигналу x(t) – Pxx(iw) дорівнює:

$$Pxx(iw) = Pss(iw) + Pnn(iw), \quad (2.10)$$

отже спектральна густина потужності корисного сигналу s(n) може бути оцінена як:

$$Pss(iw) = Pxx(iw) - Pnn(iw). \quad (2.11)$$

Через нестаціонарність мовних сигналів використовувати співвідношення (2.11) безпосередньо не можна.

На практиці, при обробці промови на досить коротких ділянках, наприклад, квазістаціонарних ділянках голосних звуків, величини Pxx(iw), Pnn(iw) апроксимують за допомогою усереднених квадратів короткочасних амплітудних спектрів сигналу і шуму, що спостерігається. Спектр шуму при цьому повинен оцінюватись у моменти пауз. Отримана таким чином оцінка відповідає квадрату амплітудного спектра сигналу.

Відновлення мовного сигналу у часовій області здійснюється за допомогою зворотного перетворення Фур'є, причому фазовий спектр для відновленого сигналу береться таким же, як і у сигналу, що спостерігається.

У найбільш загальному вигляді операція спектрального віднімання може бути виражена співвідношенням:

$$|S(t,iw)|^2 = \begin{cases} |X_i(t,iw)|^2 - A(t)|N(t,iw)|^2, & \text{якщо} \quad |X_i(t,iw)|^2 \geq (A(t)+B)|N(t,iw)|^2 \\ B|N(t,iw)|^2, & \text{навпаки} \end{cases}$$

Тут коефіцієнт A(t) (фактор переоцінювання), взагалі, залежить від співвідношення сигнал/шум на сегменті аналізу, і має типові значення близькі до 0,7…0,95, а коефіцієнт B (спектральний поріг) – вибирається в діапазоні 0.01…0.1.

Частотна характеристика фільтра шумоподавлення на основі оцінки спектральних характеристик шуму наведена на рис. 2.4.

Дослідження якості та розбірливості мови, що отримується в результаті застосування описаної методики показали {Sondhi, 1982], що в тих випадках, коли шум або перешкоди мають стаціонарний (або квазістаціонарний) характер та їх спектр має гармонійну структуру, досягається значне на слух підвищення як якості і розбірливості промови. Однак, у разі шумів з швидкозмінними спектральними характеристиками така обробка

малоефективна.

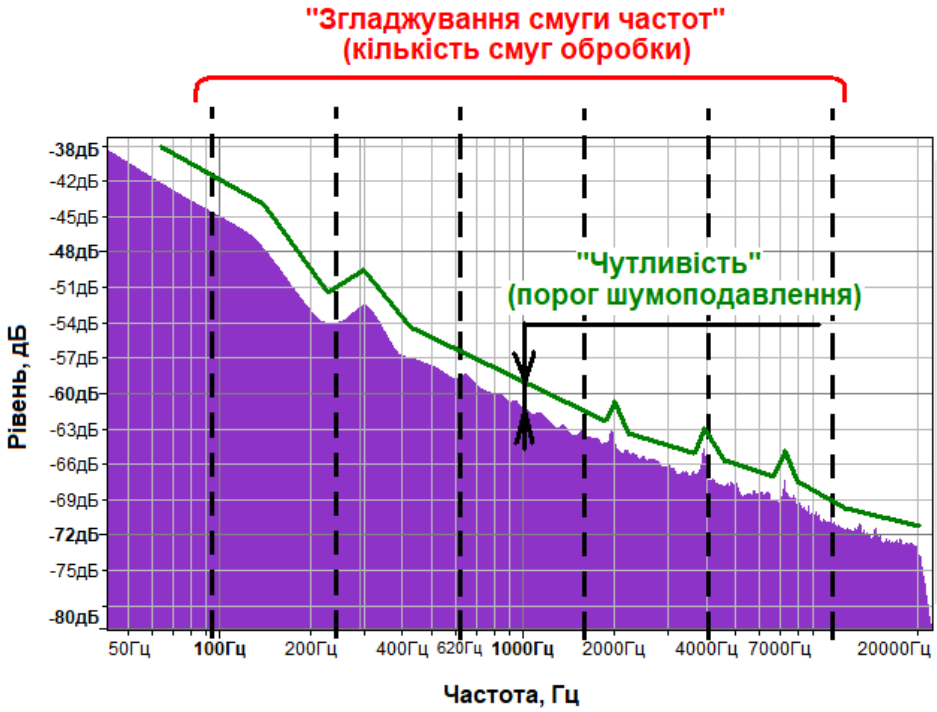


Рисунок 2.4 – Частотна характеристика фільтра шумоподавлення на основі оцінки спектральних характеристик шуму

На думку аудиторів, така мова звучить чистіше та приємніше (попри наявність характерних ефектів обробки – так званих "музичних тонів", полягають у випадкових короткочасних викидах у спектрі обробленого сигналу) ніж до обробки, проте помітного підвищення розбірливості у разі адитивних широкосмугових шумів немає, хоча відношення сигнал/шум підвищується на 6…12 дБ. [38].

2.6 Метод на основі мінімальної середньоквадратичної помилки

Описуваний алгоритм (оригінальна назва Minimum Mean-Square Error estimation) вперше було запропоновано у роботі [39]. Як і віднімання спектрів алгоритм заснований на оцінці амплітудного спектра сигналу.

Серед інших методі фільтрації, що передбачають наявність тільки одного мікрофона, алгоритми, засновані на мінімумі середньоквадратичної помилки є одними з найкорисніших. Їх використання призводить до значного

скорочення рівня шуму в сигналі без внесення залишкових спотворень типу музичних тонів [40].

Припустимо, що s(t) і b(t) позначають відповідно мовний сигнал і адитивний шум, а y(t) – сигнал, що спостерігається, тобто

$$y(t) = s(t)+b(t). \tag{2.12}$$

Нехай також S(iw), B(iw) та Y(iw) позначають відповідно спектральні компоненти мовного сигналу, шуму та зашумленого сигналу, оцінені на інтервалі аналізу, у якому передбачається квазистаціонарність мовного сигналу.

Оцінювач амплітудного спектра сигналу щонайменше середньоквадратична помилка (MMSE) визначається з двох наступних (апостерирного та апріорного) локальних відношень сигнал/шум:

$$SNR_{post}(f) = \frac{|Y(iw)|^2}{E\{|B(iw)|^2\}}$$

і

$$SNR_{prio}(f) = \frac{E\{|S(iw)|^2\}}{E\{|B(iw)|^2\}}.$$

Передаточна функція шумоподавлювача визначається формулою:

$$N(iw) = \frac{\Lambda(iw)}{1+\Lambda(iw)} N_0(iw). \tag{2.13}$$

де Λ(iw) – це узагальнене ставлення правдоподібності, яке бере до уваги величину невизначеності присутності корисного сигналу (промови) в зашумлений сигнал.

Наведені формули виведені при неявному припущенні, що апріорне відношення сигнал/шум відоме. В реальних умовах, однак, цей параметр апріорі невідомий, при цьому пропонується оцінювати його співвідношенням:

$$SNR_{prio}(t,iw) = (1-\beta)P\left[SNR_{post}(t,iw)-1\right] + \beta\frac{\left|S(t-1,iw)\right|^2}{P_B(iw)}, \qquad (2.14)$$

де t – індекс часу,

P[] – позначає операцію клішування напівхвилі.

Параметр β вибирається з емпіричних міркувань і зазвичай β =0,98.

У дослідженнях [41] стверджується, що значною мірою перевага методу оцінювання мінімальної середньоквадратичної помилки над методиками типу Вінерівської фільтрації або віднімання амплітудних спектрів пов'язано саме з введенням апріорної оцінки сигнал/шум у кожній спектральній смузі. У зв'язку з цим були запропоновані модифікації стандартних підходів (вінерівської фільтрації, віднімання амплітудних спектрів та оцінок максимальної правдоподібності).

2.7 Методи на основі штучних нейронних мереж

Розробка апарату штучних нейронних мереж призвела до появи нового типу алгоритмів для обробки зашумлених мовних сигналів, заснованих на використанні моделей нейронних мереж.

Проведемо докладний теоретичний аналіз методів шумоочищення на основі штучних нейронних мереж.

2.7.1 Підхід на основі згорткових нейронних мереж Conv-TasNet

Одним із «проривних» підходів до нейромережевого шумоподавлення та покращення мовного сигналу виявився підхід на основі згорткових

нейронних мереж Conv-TasNet. Багато сучасних підходів шумоподавлення часто порівнюються з його архітектурою. Він ґрунтується на накладанні 1D згорток на чистий сигнал без розкладання на частоти.

Попередник цієї архітектури – TasNet [42]. Архітектура TasNet складається з згорткових енкодерів і декодера з деякими особливостями:

– вихід енкодера обмежений значеннями [0, ∞);

– лінійний декодер конвертує вихід енкодера в звуковий сигнал;

– подібно до багатьох методів-попередників на основі спектрограм, на останньому етапі система апроксимує вагову функцію (в даному випадку LSTM) для кожного моменту часу.

Conv-TasNet – модифікація алгоритму TasNet, яка використовує у якості вагової функції згорткові шари з розширенням (dilation). Ця модифікація була зроблена після того, як згортки з розширенням показали себе ефективним алгоритмом при одночасному аналізі та генерації даних змінної довжини, зокрема, для синтезу таких рішень, як WaveNet [43].

Схема алгоритму Conv-TasNet показана на рис.2.5.

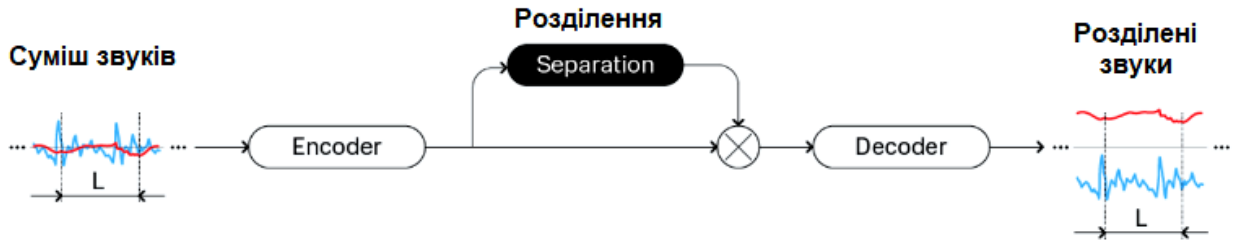


Рисунок 2.5 – Схема алгоритму Conv-TasNet

Підхід для поділу аудіо / шумоподавлення Conv-TasNet складається з 3-х компонентів:

– енкодер,

– розділення,

– декодер.

Основний компонент у схемі – етап розділення. Цей етап вирішує проблему наближеного обчислення джерел, суміш яких ми розглядаємо як «брудні» приклади. Формально припущення про «змішаність» нашого сигналу можна висловити так:

$$x(t) = \sum_{i=t}^{C} s_i(t), \tag{2.15}$$

де $x(t)$ – суміш у певний момент часу,

C – кількість джерел, що несуть внесок у суміш,

$s_i(t)$ – сигнали i-го джерела у певний час t.

Завдання алгоритму машинного навчання – визначити джерела $s_i(t)$, знаючи заздалегідь кількість джерел C та суміш $x(t)$.

Варто відзначити, що розділення в алгоритмі відбувається не відразу, а тільки після отримання ознак сигналу за допомогою «1D блоків» (1-D Conv на схемі). На рис.2.6 представлена схема, як перетворюється сигнал суміші у набір окремих джерел.

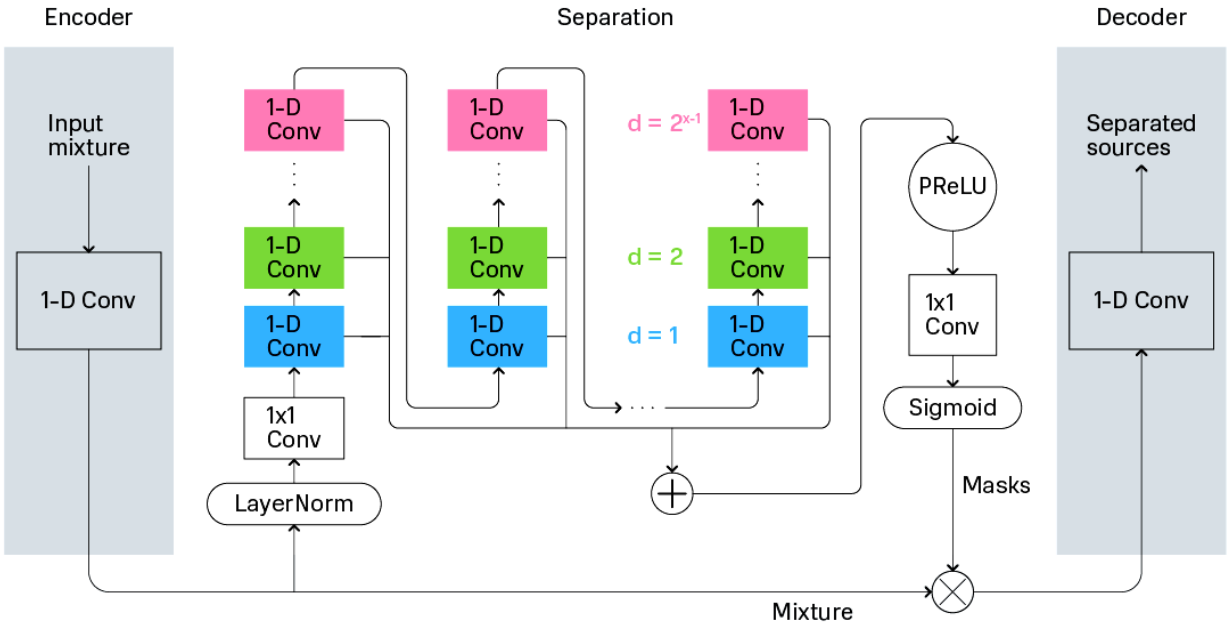


Рисунок 2.6 – Схема перетворення сигналу суміші у набір окремих джерел

1D блок, який використовується як енкодер і декодер, має наступну структуру, показану на рис.2.7.

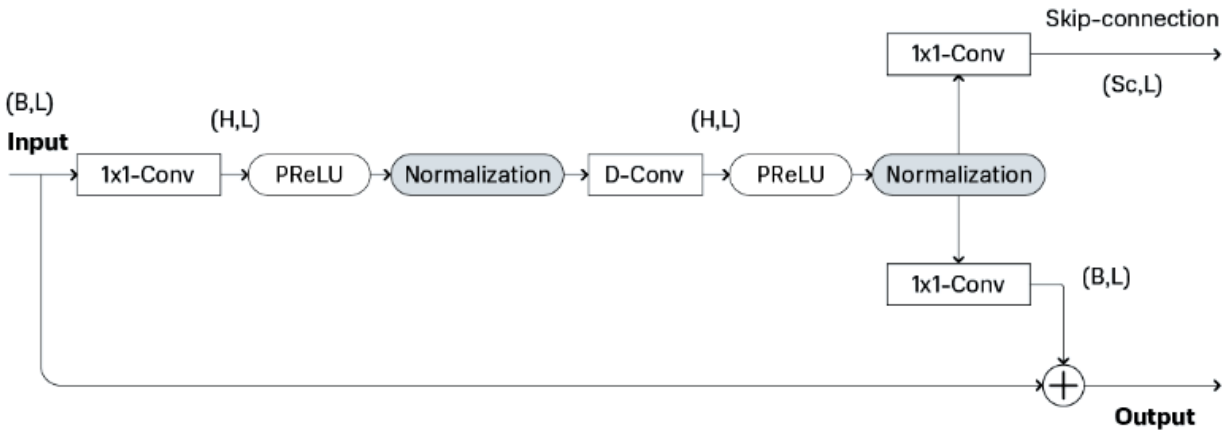


Рисунок 2.7 – Схема 1D блоку

Принцип дії системи Conv-TasNet наступний. Кодер відображає сегмент сигналу суміші у високовимірному представленні, а модуль розділення обчислює мультиплікативну функцію (тобто маску) для кожного з цільових джерел. Декодер реконструює сигнали джерела з замаскованого вигляду. Одновимірний згортковий автокодер моделює форми хвиль і часову згорткову мережу (TCN). Модуль розділення оцінює маски на основі вихідних даних кодера. Кожен 1-D згортковий блок складається з операції 1×1-conv, за якою слідує операція згортання по глибині (D-conv) з нелінійною функцією активації та нормалізацією, доданою між кожними двома операціями згортки. Два лінійних блоки 1×1-conv служать виходом і виходом пропуску з’єднання відповідно.

Детальний опис алгоритму Conv-TasNet та результати експериментів надано у [44].

2.7.2 Алгоритм DEMUCS

Алгоритм DEMUCS або глибоке вилучення музичних джерел (Deep Extractor for Music Sources) також використовується завдання розділення джерел у сигналі і шумоподавления. На відміну від попередника Conv-TasNet, цей алгоритм безпосередньо генерує джерела з вихідного сигналу,

минаючи проміжне прогноз масок.

Автори цього алгоритму надихнулися існуючою архітектурою для сегментації зображень U-Net. U-Net архітектура є кодувальником і декодувальником, між якими знаходиться пляшкова шийка. На відміну від звичайного автокодувальника шари між собою пов'язані «з'єднаннями швидкого доступу», в результаті підсумковий сигнал не погіршується після стиснення. Схема алгоритму U-Net для шумоподавлення виглядає так, як показано на рис.2.8.

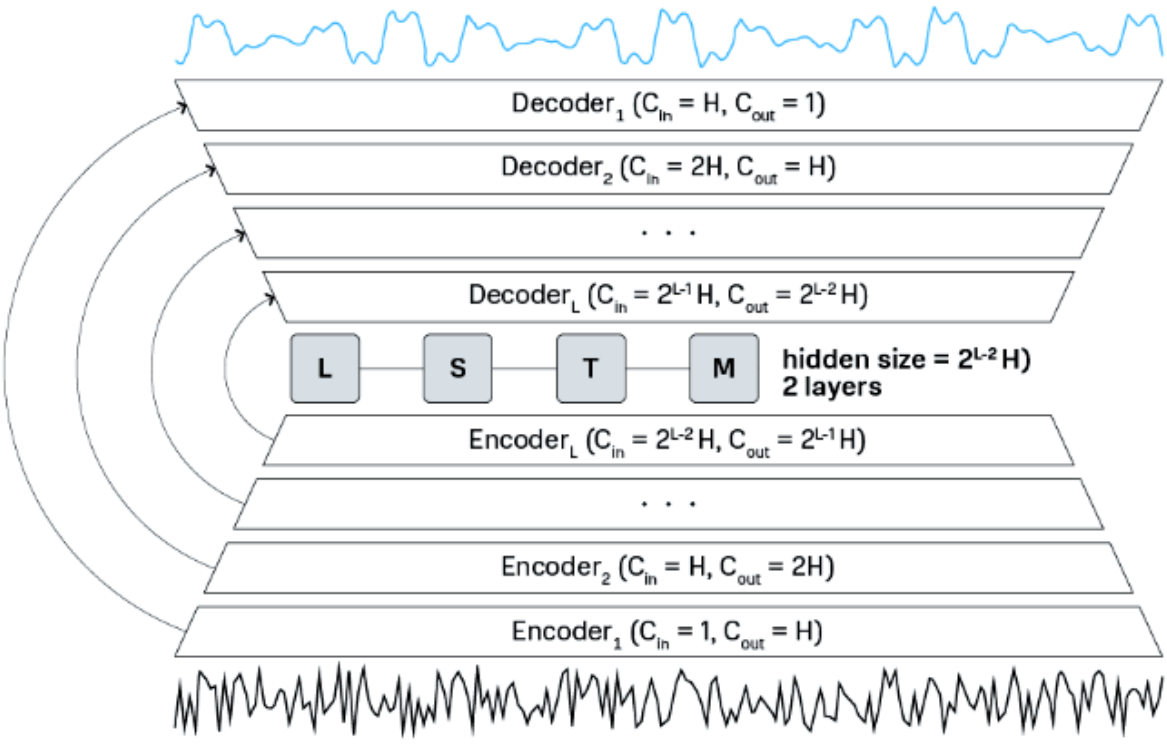


Рисунок 2.8 – Схема алгоритму U-Net для шумоподавлення

Як пляшкова шийка в DEMUCS – односпрямований LSTM шар. Це дозволяє ефективно використовувати алгоритм аналізу потокових даних. Кодувальник та декодувальник сформовані з блоків, які складені зі згорткових шарів (1D, 1x1 та 1D Transpose) та функцій активації (Gated Linear Unit та Rectified Linear Unit). Вони скомпоновані таким чином (рис.2.9).

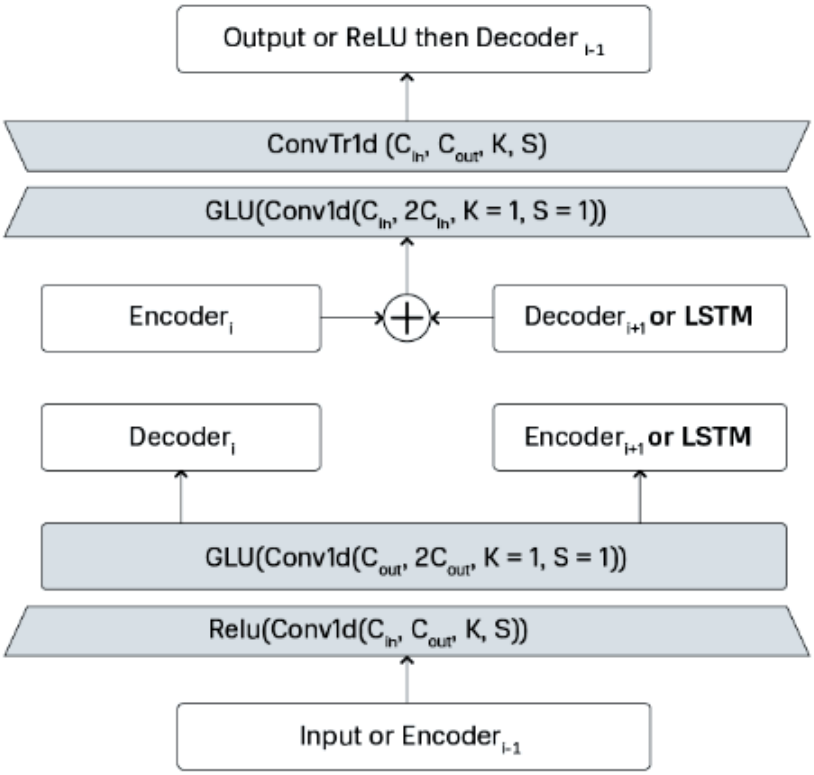


Рисунок 2.9 – Компонування блоків кодера-декодера

В якості функцію втрат при шумоподавленні достатньо використовувати L1 Loss між передбаченим записом і еталонним, але для поліпшення збіжності автори статті використовують також STFT Loss різного масштабу (STFT з різними параметрами при підрахунку функцій втрат), який є сумою двох функцій втрат – збіжності (spectral convergence) та амплітуд (magnitude):

$$L_{STFT}(y,\hat{y}) = L_{sc}(y,\hat{y}) + L_{mag}(y,\hat{y}),$$

$$L_{sc}(y,\hat{y}) = \frac{\| \, |STFT(y)| - |STFT(\hat{y})| \, \|_F}{\| \, STFT(y) \, \|_F}, \qquad (2.16)$$

$$L_{mag}(y,\hat{y}) = \frac{1}{T} \| \log|STFT(y)| - \log|STFT(\hat{y})| \|_1 .$$

де $y, \hat{y}$ – еталонний сигнал і передбачений сигнал відповідно,

T – тривалість сигналу,

$\|\cdot\|_F$ – норма Фробеніуса,

$\|\cdot\|_1$ – L1 "норма" (абсолютна помилка).

Детальний опис алгоритму DEMUCS та результати експериментів надано у [45].

2.7.3 Алгоритм HiFi-GAN

Походи, викладені в пунктах 2.7.1 і 2.7.2, добре генералізуються при вирішенні завдань шумозаглушення, щоб вичленувати мову і позбавлятися немовних подій в аудіо потоці. Але ці алгоритми можуть створювати артефакти в сигналі, які можуть заважати сприйняттю людиною, або псувати якість подальшої автоматичної обробки, наприклад, розпізнавання промови.

Частково позбавитися артефактів допомагає згладжування, але воно часто попутно знижує чіткість всього аудіо. На відміну від попередників, генеративно-стільна мережа високої точності (High Fidelity Generative Adversarial Network) добре справляється з генерацією аудіо подібно до студійного запису без артефактів штучної генерації.

Алгоритм HiFi-GAN складається з трьох основних частин (рис.2.10).

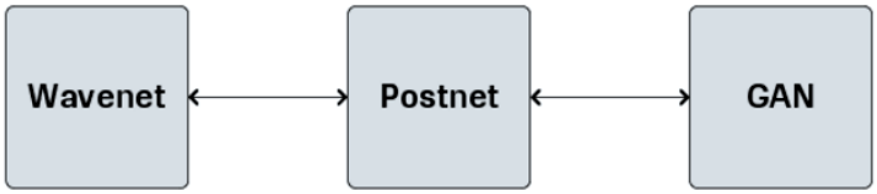


Рисунок 2.10 – Загальна схема алгоритму HiFi-GAN

За генерацію чистого сигналу на основі зашумленого відповідає блок WaveNet, цей алгоритм значно успішно використовувався для синтезу слова (текст → аудіо). При модифікації завдань для аналізу звуку ця архітектура також показала себе ефективною.

Особливість WaveNet-а для шумоподавлення в тому, що генерація нового сигналу відбувається для всього цілого запису, а не для кожного

моменту часу t_n, як це робиться в оригінальному алгоритмі WaveNet. Це дозволяє підвищити швидкість генерації для точної паралелізації процесів, які можуть виконуватися одночасно.

Після генерації сигналу WaveNet-ом проходить через кілька згорткових шарів, цей етап називається Postnet (рис.2.11).

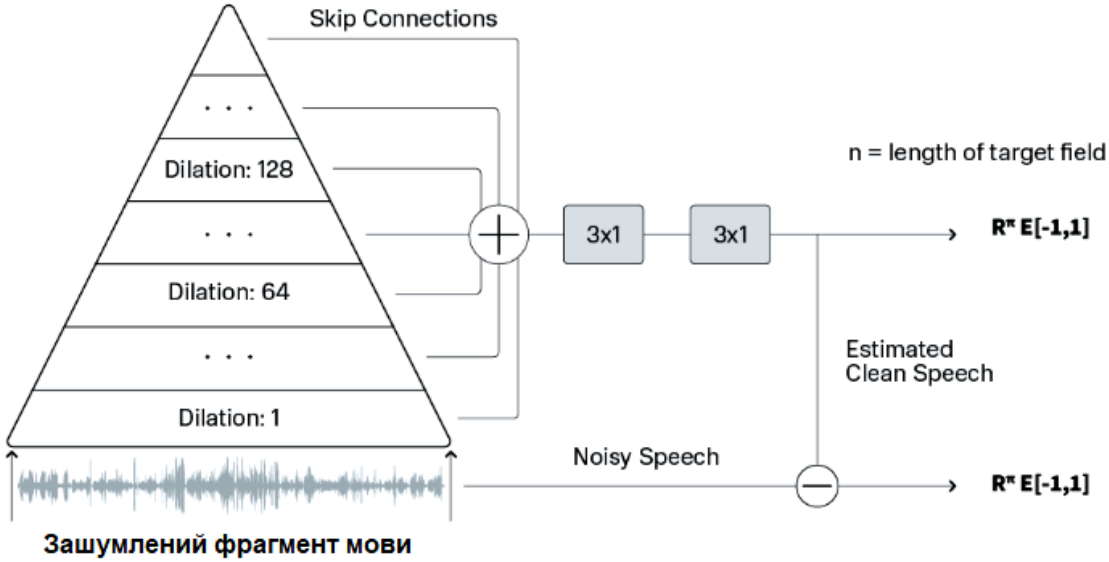


Рисунок 2.11 – Структурна схема етапу Postnet

Postnet потрібен, щоб виправляти і вточнювати грубе і наближене повідомлення WaveNet-а. Крім того, Postnet-регулююча дія додатково надає чотири різних дискримінатора, які навчені виділити чисті оригінальні записи зі створених. Кожен дискримінатор приймає вихід Postnet-а в різному форматі:

– сигнал у вихідному вигляді з різною частотою дискретизації 16 кГц, 8 кГц, 4 кГц;

– Мел-спектрограму сигналу.

Все разом зв'язується в архітектуру, показану на рис.2.12.

У підсумку для навчання використовуються наступні функції потенціалу:

– L1 (абсолютна помилка у сигналі);

– функція потенціалу на лог-спектрограмах передбачених і чистого сигналу після перетворення Фурье з наступними параметрами:

а) розмір вікна 2048 і крок 512,

б) розмір вікна 512 і крок 128,

– змагальна функція потенціалу (adversarial loss) для навчання Postnet-у;

– функція потенціалу глибинних ознак (deep feature loss) для навчання дискримінаторів.

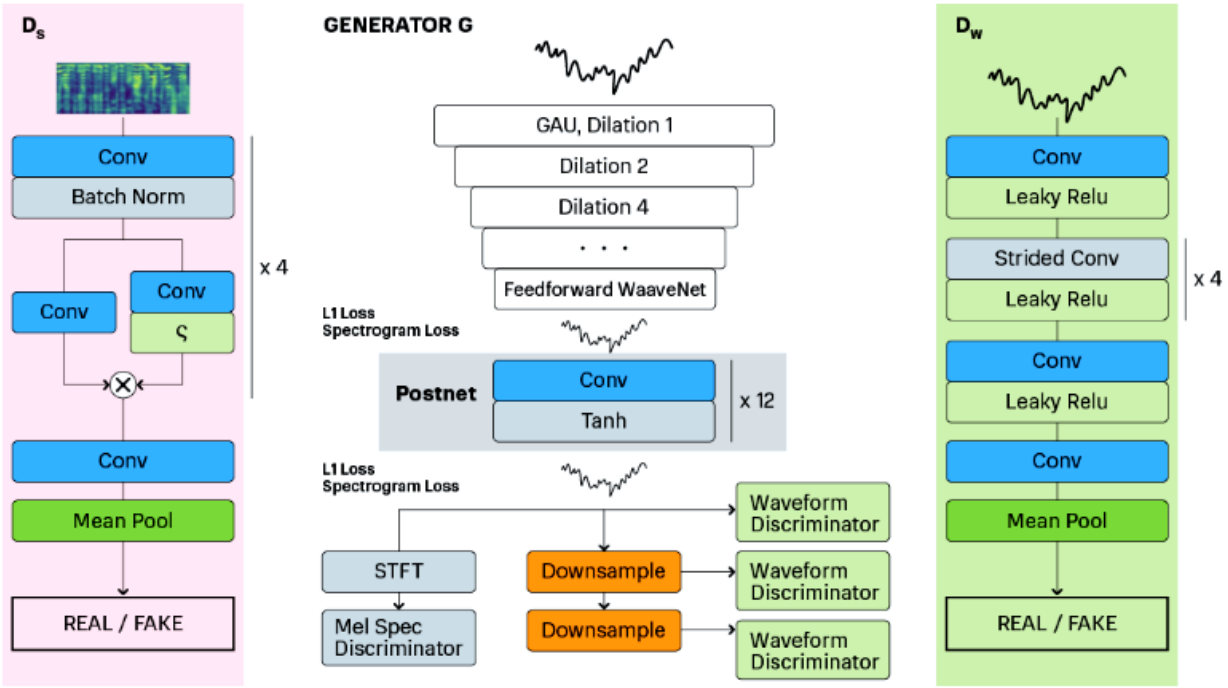


Рисунок 2.12 – Загальна архітектура алгоритму HiFi-GAN

Детально про функції потенціалу, про архітектуру, а також про експерименти можна ознакомитись у статті [46].

2.8 Висновки по розділу 2

Теоретичний аналіз методів підвищення якості та розбірливості зашумлених мовних сигналів показує, що існує багато різних підходів до обробки зашумленого мовлення. Така різноманітність методів обумовлена як важливістю проблеми і відсутністю досить надійних методів її вирішення.

Об'єктивне порівняння цих методів та вибір найбільш прийнятних зробити дуже важко, тому що перед системами корекції мовних сигналів ставляться різні завдання.

Наприклад, можна як головний критерій використовувати підвищення розбірливості мови, припускаючи при цьому можливість спотворень у тембрі голосу або появу артефактів у вигляді структурованого шуму. Можна поставити за мету зниження стомлюваності аудитора або збереження натуральності голосу диктора, що досягається в основному рахунок підвищення якості мовного сигналу. Зрештою, можуть бути відомі заздалегідь важливі апріорні відомості, наприклад тип або параметри шуму, показники голосу диктора, нарешті, гіпотези про вимовний текст, що також може визначальним чином вплинути вибір методу фільтрації.

Важливо відзначити, що універсальні методи обробки, які однаково добре боролися б із суттєво нестаціонарними та стаціонарними, адитивними та мультиплікативними шумами, суттєво підвищували б якість і одночасно розбірливість мови зараз немає, і можливо не буде.

Як типова (за рідкісними, зазначеними в огляді винятками, спостерігається зворотна тенденція: якщо порівнювати системи обробки зашумленого мовлення за двома показниками – підвищенням якості звучання мовних сигналів та підвищенням розбірливості, то системи, що підвищують якість та натуральність звучання, швидше за все знижують розбірливість і навпаки, підвищення розбірливості призводить до зниження якості та натуральності звучання. Тому, багато з розглянутих методів фільтрації слід розглядати як взаємодоповнюючі, і в ідеальному випадку необхідно мати бібліотеку з кількох методів фільтрації.

Розглядаючи останні тенденції в галузі обробки зашумлених сигналів слід особливо виділити високі результати, отримані за рахунок використання математичних моделей мовних сигналів (авторегресійні, приховані марківські моделі), а також використання нейроподібних структур для фільтрації адитивних стаціонарних шумів.

В останні кілька років тема нейромережевого шумоподавлення стає дедалі популярнішою і поки що не зупиняється на досягнутому. Тому слід очікувати нових проривних результатів в цій галузі.

3 ТЕСТУВАННЯ ПРОГРАМНИХ МЕТОДІВ ШУМООЧИЩЕННЯ, ПІДВИЩЕННЯ ЯКОСТІ ТА РОЗБІРЛИВОСТІ МОВИ

3.1 Постановка задачі

Як показано в розділі 1, можна навести наступну класифікацію шумів за складністю задачі їх подавлення:

– стаціонарні;

– нестаціонарні змінні;

– нестаціонарні переривчасті;

– нестаціонарні імпульсні.

Отже, для тестування нам необхідно підготувати такі заготовки аудіофайлів:

– мова диктора (українська, студійний запис, без шумів, з мінімумом реверберації);

– стаціонарний шум (наприклад, шум працюючого холодильника);

– змінний шум (наприклад, шум вентилятора, що змінює своє положення);

– переривчастий шум (наприклад, шум кухонного блендера, що час від часу включається в сусідній кімнаті);

– імпульсний шум (наприклад, шум клацання комп'ютерної мишки, або шум набору тексту на клавіатурі).

Розрахунок відношення сигнал/перешкода для стаціонарного і змінного шумів доцільно проводити за формулою:

$$SNR = 20 \lg \sqrt{\sum_{t=1}^{T} s^2(t) / \sum_{t=1}^{T} n^2(t)}, \qquad (3.1)$$

де $s(t)$ – відліки корисного сигналу мови диктора;

$n(t)$ – відліки шуму;

t – номер відліку сигналу;

T – тривалість реалізації.

У разі імпульсного шуму формула (3.1) дасть занижений результат, тому доцільно використовувати вираз:

$$PSNR = 20 \lg \sqrt{\sum_{t=1}^{T} s^2(t) / \max[n^2(t)]}. \quad (3.2)$$

Ступінь очищення сигналу від шуму можна оцінювати за об’єктивними показниками, наприклад, виграш у відношенні сигнал-шум:

$$K = 20 \lg \sqrt{\sum_{t=1}^{T} s_0^2(t) / \sum_{t=1}^{T} (s(t) - s_0(t))^2}, \quad (3.3)$$

де s – відновлений сигнал;

s_0 – первинний сигнал без завад.

Більш доцільно буде контролювати суб’єктивні показники:

– ступінь зашумленості спектрограм обробленого сигналу;

– якість обробленого сигналу по сприйняттю на слух.

3.2 Підготовка матеріалів для тестування

Підготовка матеріалів для тестування виконувалася в умовах затишної кухонної кімнати площею 6,3 м^2 з бархатними шторами і м’яким куточком.

Запис вівся на диктофон смартфону Xiaomi Redmi 12, далі оброблявся один зі стереоканалів, мікрофон якого розташований ближче до джерела сигналу.

Відстань до джерела обиралася таким чином, щоби забезпечити максимум рівня записаного сигналу, але уникнути при цьому перевантаження або ефекту задування у мікрофон.

Для додаткового уникнення поп-ефекту на мікрофон смартфону надягався умовний «поп-фільтр» у вигляді шерстяної шкарпетки.

Загальна схема підготовки аудіофрагментів показана на рис. 3.2.

Рисунок 3.2 – Загальна схема підготовки аудіо фрагментів

У якості дикторського тексту взято фрагмент ефіру з чоловічим голосом з Радіо НВ.

На рис. 3.3 показано хвильоформу та спектрограму запису голосу диктора.

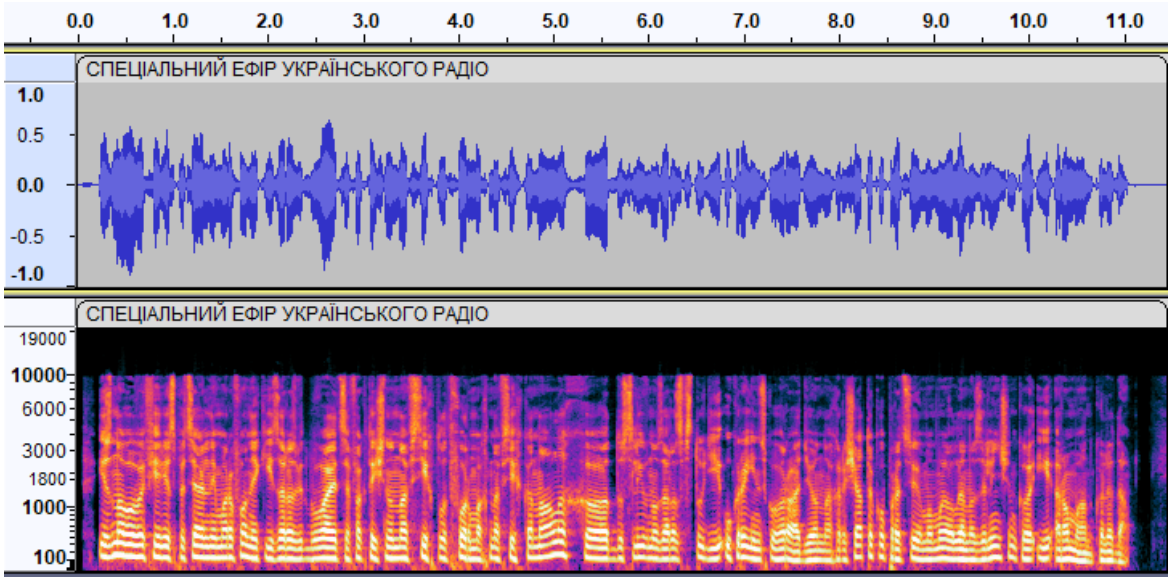


Рисунок 3.3 – Хвильоформа та спектрограма запису голосу диктора

На рис. 3.4 і рис. 3.5 показано часові реалізації (хвильоформи) та спектрограми записаних реалізацій шумів.

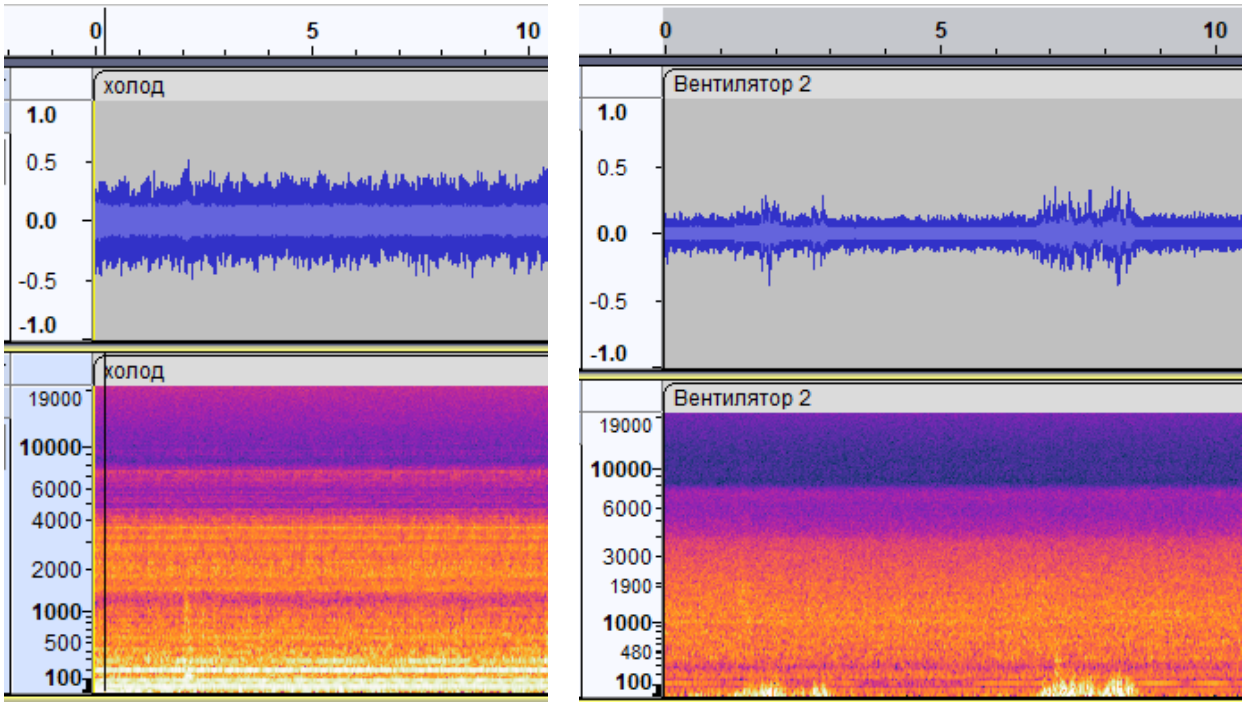


Рисунок 3.4 – Часові реалізації (хвильоформи) та спектрограми записаних реалізацій шумів холодильника (справа – стаціонарний шум) і вентилятора (зліва – змінний шум)

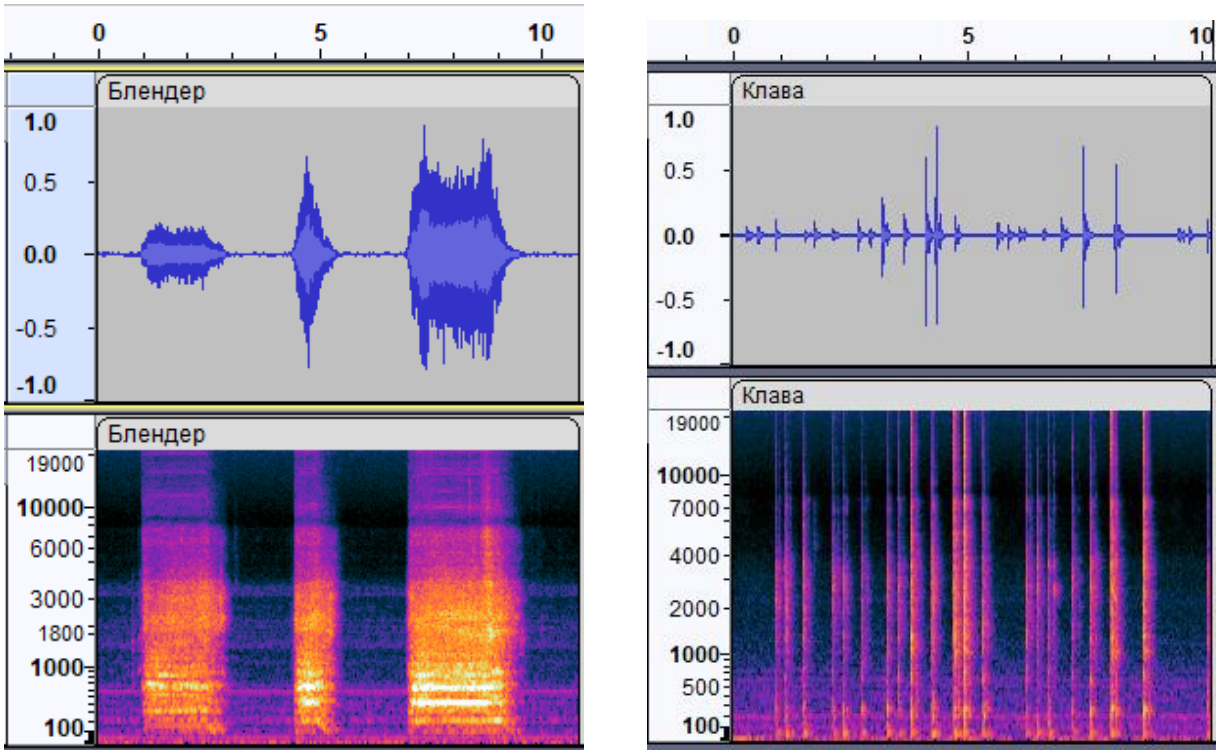


Рисунок 3.5 – Часові реалізації (хвильоформи) та спектрограми записаних реалізацій шумів блендера (справа – переривчастий шум) і клавіатури (зліва – імпульсний шум)

З даних матеріалів мішкувалися зашумлені аудіодоріжки голосу диктора з відношенням сигнал-шум 24 дБ. На рис. 3.6 і рис. 3.7 показано часові реалізації (хвильоформи) та спектрограми зашумлених записів голосу диктора.

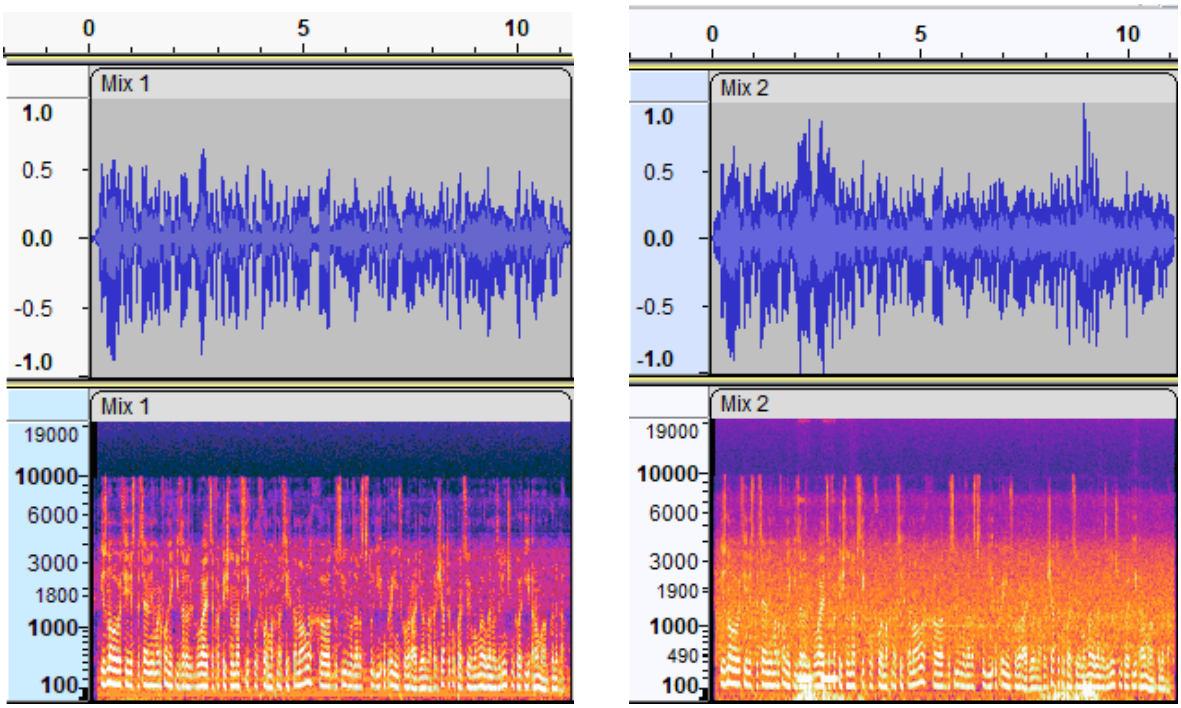


Рисунок 3.6 – Часові реалізації та спектрограми зашумлених записів голосу диктора (зліва–шум холодильника, справа – шум вентилятора)

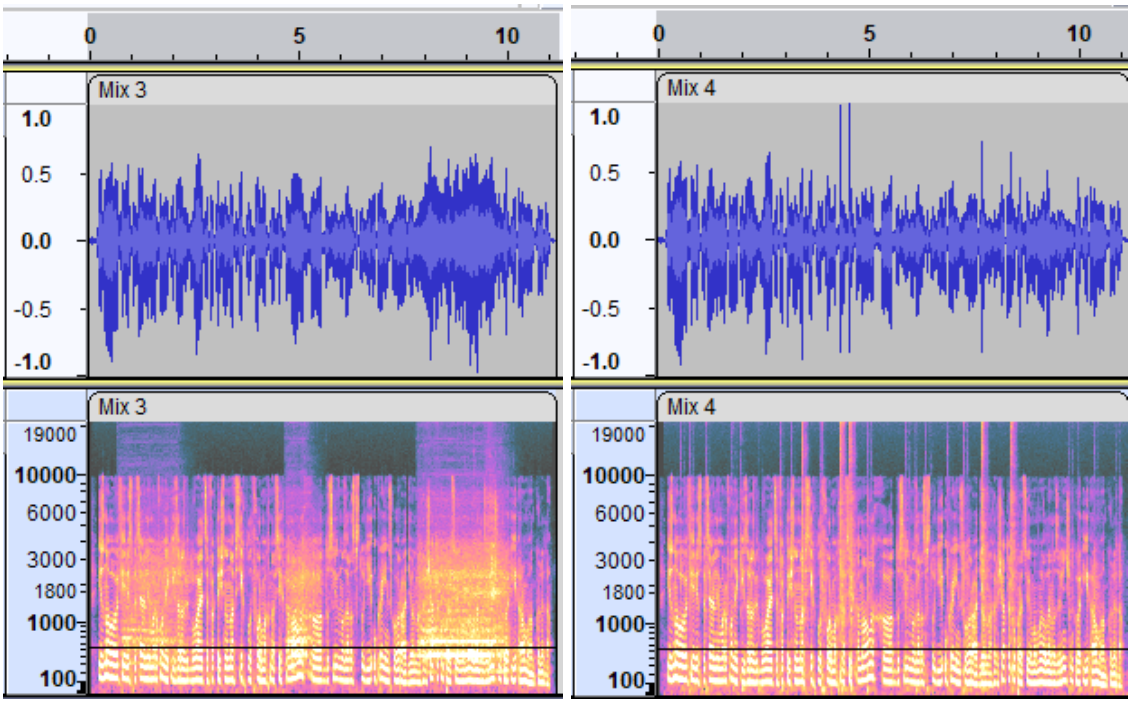


Рисунок 3.7 – Часові реалізації та спектрограми зашумлених записів голосу диктора (справа – шум блендера, зліва – шум клавіатури)

З точки зору класичних завдань радіотехніки – виділення та розпізнавання сигналів – це вагоме значення. Але з точки зору звукозапису та забезпечення якісного звуку для слухачів, це мало. Тобто, умови задачі шумоочищення є коректними.

3.3 Тестування шумоподавлення методом спектрального віднімання

Для тестування шумоподавлення методом спектрального віднімання використано шумоподавлювач Noise Reduction програми Audacity (рис.3.8).

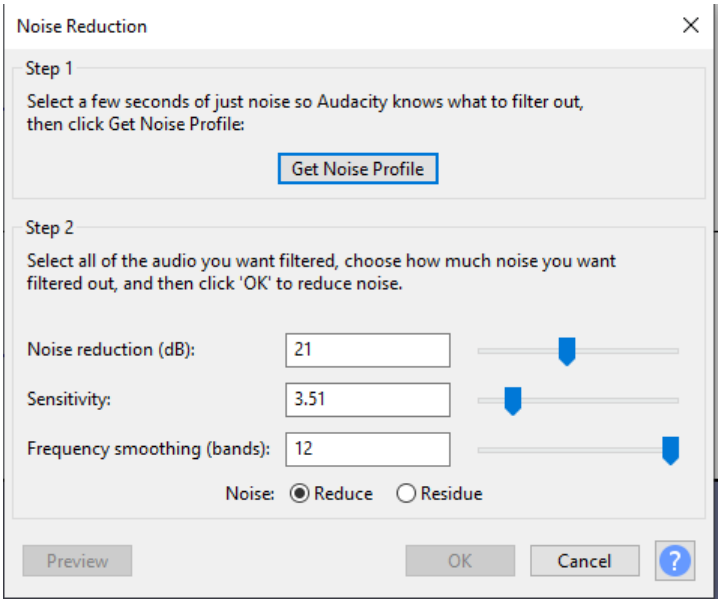


Рисунок 3.8 – Налаштування ефекту Noise Reduction Audacity

Ефект шумоподавлення Noise Reduction складається з двох етапів:

– отримання зразку шуму;

– подавлення шуму з обраними налаштуваннями.

Від налаштувань суттєвою мірою залежить ефективність подавлення шуму за мінімальних спотворень сигналу. Розглянемо більш детально налаштування ефекту Noise Reduction та їх вплив на обробку. На рис.3.9 на спектрі наочно показано вплив налаштувань на процес шумоподавлення.

Вхідна суміш сигналу і шуму поділяється на декілька смуг частот. Число смуг визначається параметром «Frequency smoothing (bands)». Чим

більший цей параметр, тим більше смуг частот, в яких буде проведено незалежне подавлення шуму. На рис.3.9 штриховою лінією показані умовні границі цих смуг частот.

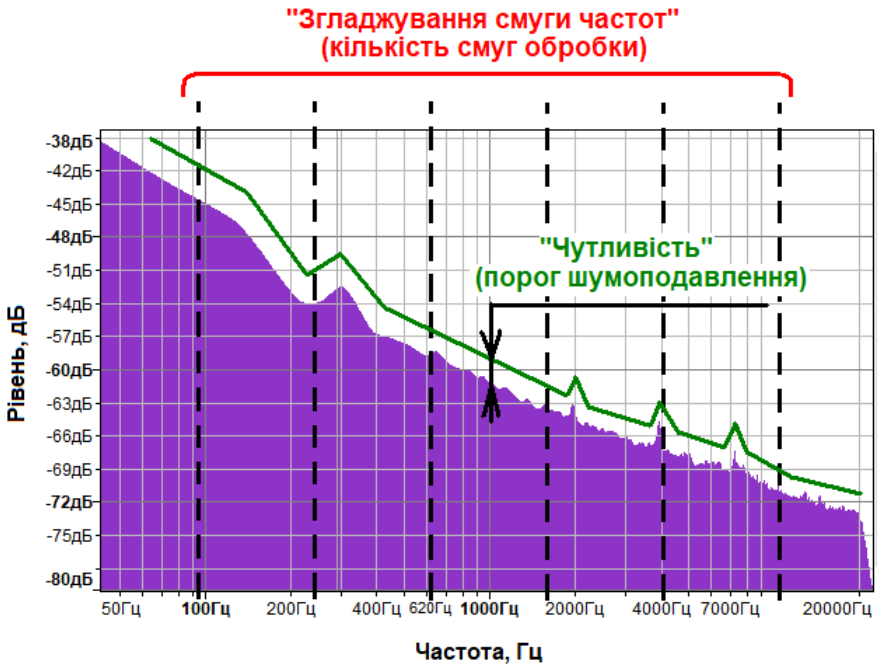


Рисунок 3.9 – Вплив налаштувань ефекту Noise Reduction на процес шумоподавлення

Шумоподавлення в певній смузі частот здійснюється тільки тоді, коли суміш сигналу і шуму в цій смузі буде менше за поріг. Значення порогу над шумом задається налаштуванням «Sensitivity», яке показує, на скільки дБ поріг має перевищувати шум.

Якщо суміш сигналу і шуму в заданій смузі перевищить поріг, то подавлення в цій смузі не відбудеться. Якщо суміш сигналу і шуму знаходитимесь нижче порогу, то в даній смузі буде ослаблення на значення налаштування «Noise reduction» в дБ.

На рис.3.10 і рис.3.11 показано спектрограми сигналів до і після шумоподавлення для різних поєднань сигналу і шуму.

Зліпок шуму для нестаціонарних шумів брався з тих часових проміжків, де сигнал був мінімальний, а шум найбільш представницький.

Величина шумоподавлення обиралась так, щоби забезпечити достатнє подавлення шуму на всіх ділянках, де він присутній.

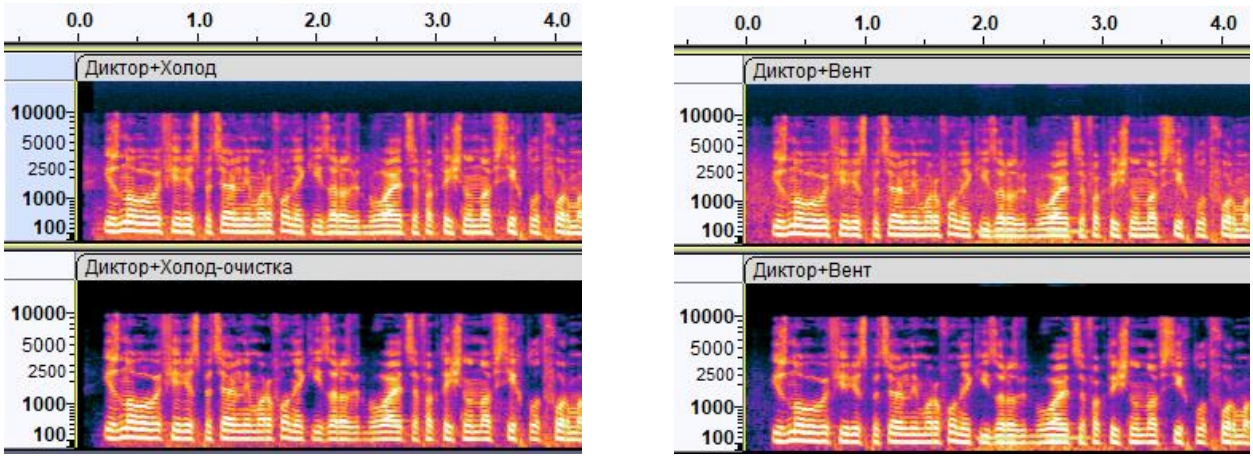


Рисунок 3.10 – Спектрограми зашумлених та очищених записів голосу диктора (зліва–шум холодильника, справа – шум вентилятора)

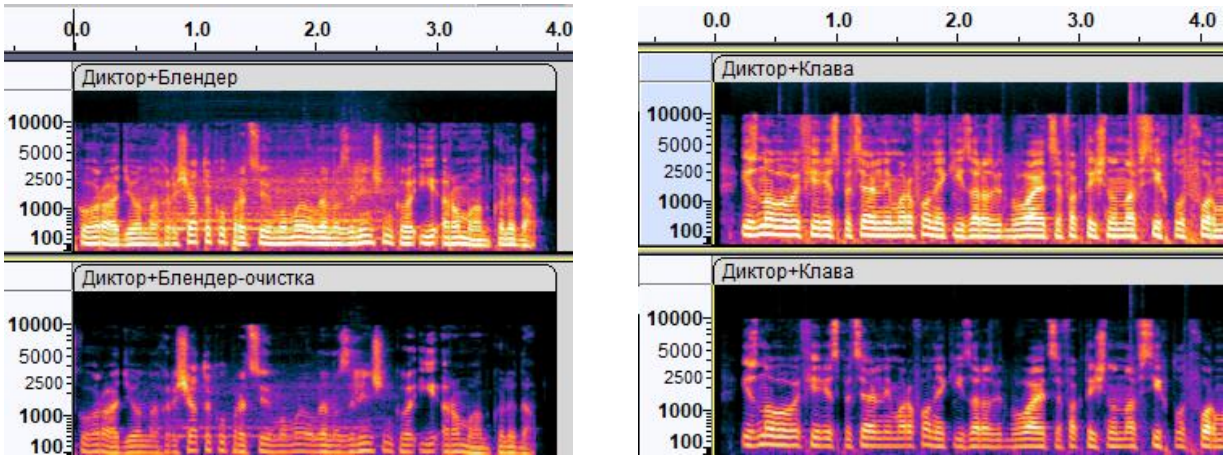


Рисунок 3.11 – Спектрограми зашумлених та очищених записів голосу диктора (справа – шум блендера, зліва – шум клавіатури)

Як слідує з аналізу наведених спектрограм, найкраще шумоподавлення методом спектрального віднімання спостерігається для стаціонарного і змінного шуму. При цьому корисний сигнал зазнає мінімальних спотворень.

Найгірша ситуація з імпульсними шумами. Щоби забезпечити їх помітне подавлення, необхідно обирати дуже високий поріг в усіх смугах частот, що призводить до сильних втрат оброблюваного сигналу (чорні області на спектрограмі), а імпульсний шум все одно при цьому залишається помітним.

Дослідимо шумоподавлення за допомогою нейронної мережі.

3.4 Тестування шумоподавлення за допомогою нейронної мережі

Дослідимо шумоподавлення за допомогою нейронної мережі Adobe Podcast.

Онлайн сервіс Adobe Podcast орієнтований на спеціалістів, які займаються створенням подкастів. Технологічним фундаментом Adobe Podcast є аудіо редактор Project Shasta, представлений наприкінці 2021 року.

За ствердженнями Adobe, сервіс здатний доводити зроблені на мікрофон середнього рівня записи на стільки, що вони звучатимуть, наче записувалися у професійній студії. Користуватися сервісом Adobe Podcast можна безкоштовно, але для взаємодії з ним потрібен обліковий запис Adobe. Інтерфейс сервісу Adobe Podcast показано на рис.3.12.

Рисунок 3.12 – Інтерфейс сервісу Adobe Podcast

Сюди можна завантажувати файли у форматах MP3 та WAV тривалістю до 1 год. та обсягом до 1 ГБ. Залежно від обсягу файлу процес обробки може складати кілька хвилин. В нашому випадку обробка тривала приблизно так само, як і його тривалість.

На рис.3.13 і рис.3.14 показано спектрограми сигналів до і після шумоподавлення для різних поєднань сигналу і шуму. Ступінь ефекту шумоподавлення використано максимальний з того, що надає мережа Adobe Podcast.

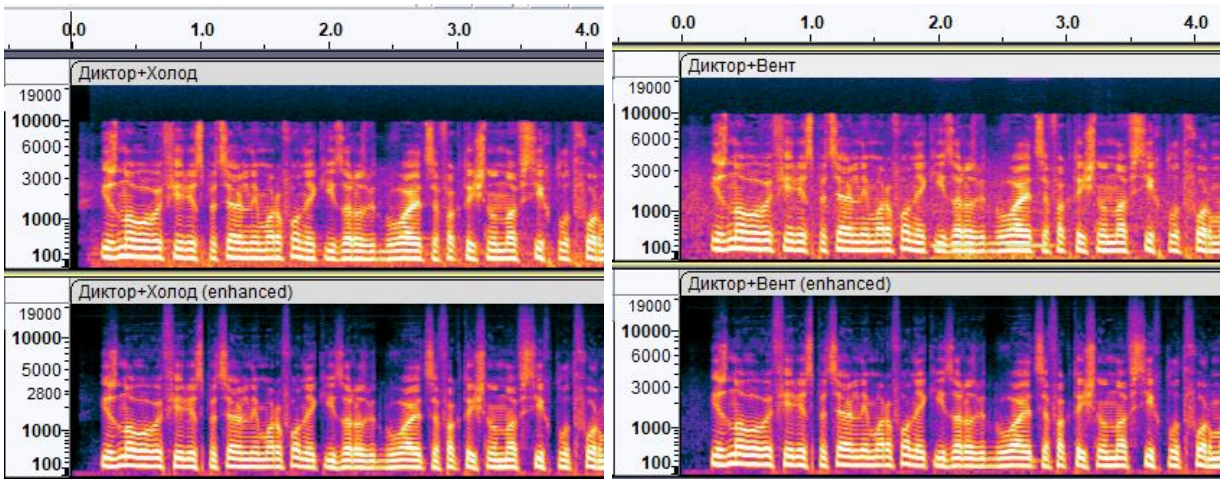


Рисунок 3.13 – Спектрограми зашумлених та очищених Adobe Podcast записів голосу диктора (зліва–шум холодильника, справа – шум вентилятора)

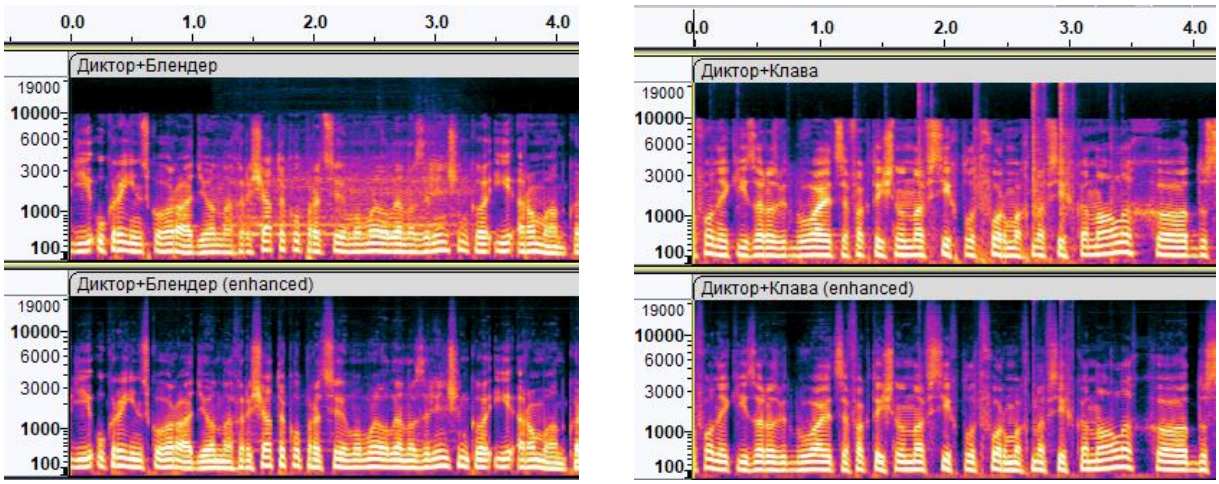
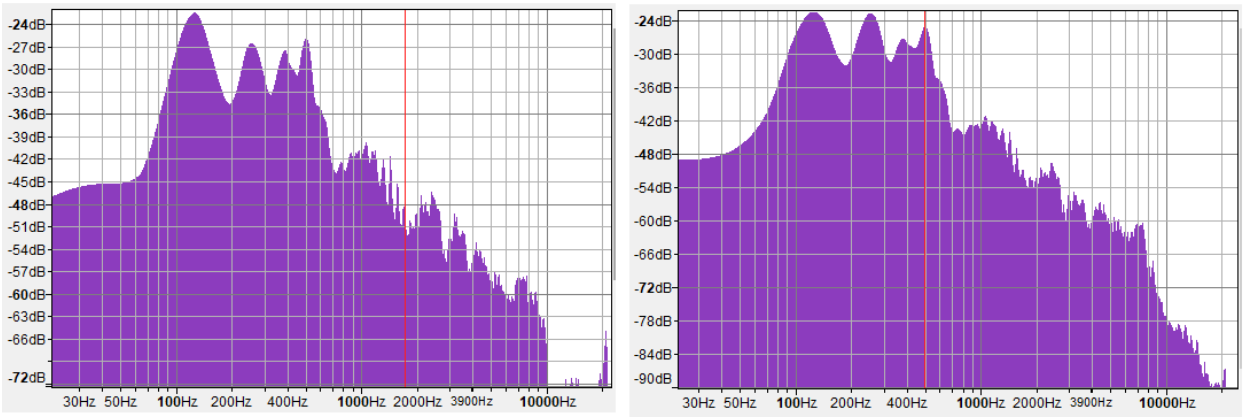


Рисунок 3.14 – Спектрограми зашумлених та очищених Adobe Podcast записів голосу диктора (справа – шум блендера, зліва – шум клавіатури)

Як слідує з аналізу наведених спектрограм, гарне шумоподавлення спостерігається для всіх видів шуму, як стаціонарного, так і нестаціонарного, включаючи імпульсний. При цьому зберігаються всі особливості корисного сигналу, як на слух, так і на спектрограмі. Крім того, звертає на себе увагу, що нейронна мережа робить спектр мови більш рівномірним, але залишає всі його характерні особливості. Різкий обрив спектру на 10 кГц в обробленому варіанті перетворюється на екстрапольований спектр, який плавно спадає до 20 кГц. Причому ця екстраполяція відбувається не завжди, а там де треба, тобто на різких транзієнтах мовного сигналу.

На рис.3.15 показані спектри звукозапису до обробки (а) і після обробки (б).



а) б)

Рисунок 3.15 – Спектри звукозапису до обробки (а) і після обробки (б) Adobe Podcast

На слух шумоочищення нейронною мережею відрізняється від шумоподавлення методом спектрального віднімання. По-перше, відсутній музичний шум, а звук стає схожим на той, який записано в студійних умовах.

3.5 Висновки по розділу 3

Для тестування заготовлено аудіофайли мови диктора (українська, студійний запис, без шумів, з мінімумом реверберації); стаціонарний шум (шум працюючого холодильника); змінний шум (шум вентилятора, що змінює своє положення); переривчастий шум (шум кухонного блендера, що час від часу включається в сусідній кімнаті); імпульсний шум (шум набору тексту на клавіатурі).

Підготовка матеріалів для тестування виконувалася в умовах затишної кухонної кімнати площею 6,3 м^2 з бархатними шторами і м’яким куточком.

Запис вівся на диктофон смартфону Xiaomi Redmi 12, далі оброблявся один зі стереоканалів, мікрофон якого розташований ближче до джерела сигналу.

Відстань до джерела обиралася таким чином, щоби забезпечити максимум рівня записаного сигналу, але уникнути при цьому перевантаження або ефекту задування у мікрофон.

Для додаткового уникнення поп-ефекту на мікрофон смартфону надягався умовний «поп-фільтр» у вигляді шерстяної шкарпетки.

З даних матеріалів в програмі Audacity мішкувалися зашумлені аудіодоріжки з відношенням сигнал-шум 24 дБ. Для стаціонарного і змінного шумів контролювалося середньоквадратичне відношення сигнал/шум, для імпульсного і переривчастого шумів – пікове відношення сигнал/шум.

З точки зору класичних завдань радіотехніки 24 дБ це вагоме значення. Але з точки зору звукозапису та забезпечення якісного звуку для слухачів, це мало. Тобто, умови задачі шумоочищення обрані коректними.

Для тестування шумоподавлення методом спектрального віднімання використано шумоподавлювач Noise Reduction програми Audacity. Зліпок шуму для нестаціонарних шумів брався з тих часових проміжків, де сигнал був мінімальний, а шум найбільш представницький.

Величина шумоподавлення обиралась так, щоби забезпечити достатнє подавлення шуму на всіх ділянках, де він присутній. Як слідує з аналізу отриманих спектрограм, найкраще шумоподавлення методом спектрального віднімання спостерігається для стаціонарного і змінного шуму. При цьому корисний сигнал зазнає мінімальних спотворень.

Найгірша ситуація з імпульсними шумами. Щоби забезпечити їх помітне подавлення, необхідно обирати дуже високий поріг в усіх смугах частот, що призводить до сильних втрат оброблюваного сигналу (чорні області на спектрограмі), а імпульсний шум все одно при цьому залишається помітним.

Досліджено шумоподавлення за допомогою нейронної мережі Adobe Podcast. Користуватися даним онлайн сервісом можна безкоштовно.

Як слідує з отриманих спектрограм, гарне шумоподавлення спостерігається для всіх видів шуму, як стаціонарного, так і нестаціонарного,

включаючи імпульсний. При цьому зберігаються всі особливості корисного сигналу, як на слух, так і на спектрограмі. Крім того, звертає на себе увагу, що нейронна мережа робить спектр мови більш рівномірним, але залишає всі його характерні особливості. Різкий обрив спектру на 10 кГц в обробленому варіанті перетворюється на екстрапольований спектр, який плавно спадає до 20 кГц. Причому ця екстраполяція відбувається не завжди, а там де треба, тобто на різких транзієнтах мовного сигналу.

На слух шумоочищення нейронною мережею відрізняється від шумоподавлення методом спектрального віднімання. По-перше, відсутній музичний шум, а звук стає схожим на той, який записано в студійних умовах.

ВИСНОВКИ

Сьогодні висуваються все вищі і вищі вимоги до технічної якості медіаконтенту. Окремі вимоги існують до якості звукової доріжки. Коли запис проходить не в студії, часто звукозапис виходить зачумленим, причому характеристик шуму змінюються в часі.

Для поліпшення якості звукові сигнали піддають шумоочищенню. В даний час існує кілька класів методів шумоочищення та шумозаглушення, проте усі вони мають недоліки. Для задач звукозапису алгоритм має ефективно подавляти шуми, вносити мінімум спотворень в сигнал і бажано мати невелику затримку.

Мета кваліфікаційної роботи – огляд, аналіз та тестування нових адаптивних методів та алгоритмів шумозаглушення для мовлення, визначення на основі цього аналізу та тестування найбільш ефективних алгоритмів детектування та придушення шуму.

В першому розділі складено класифікацію шумів за статистичними, спектральними, часовими характеристиками. Коли ми хочемо позбавитися шумів у записі мови, варто в першу чергу класифікувати шуми за часовими характеристиками. Тут можна виділити: стаціонарні шуми, змінні у часі шуми, переривчасті та імпульсні шуми.

Якщо скласти піраміду складності задач подавлення зазначених шумів, то серед нестаціонарних найскладнішим є випадок імпульсного шуму, оскільки він має найвищу ентропію. Найпростішим є повільно змінний шум, оскільки його можна розглядати як стаціонарний на обмеженому інтервалі часу.

Умовно методи всі методи подавлення шуму можна розділити на аналітичні (традиційні), і передові – нейромережеві методи. Якщо задачі нагорі піраміди можна вирішити обчислювальними методами, то задачі в нижній частині піраміди можна вирішити лише методами машинного навчання. Якщо обчислювальні методи вирішують задачі позбавлення

сигналу певного шуму, то нейромережеві методи навчаються вирішувати задачі виділення лише релевантної мовної інформації з усього аудіопотоку.

В другому розділі проведено теоретичний аналіз робіт з дослідження методів подавлення шуму у звукозаписах. Адаптивні компенсатори перешкод дозволяють значно покращити якість зашумлених сигналів – на кілька десятків децибелів, але вимога наявності опорного сигналу сильно обмежує їх галузь застосування. У багатьох випадках, наприклад, під час реставрації архівних записів або в криміналістиці, опорного сигналу у явному вигляді немає. Тому опорний сигнал часто доводиться формувати на основі непрямих міркувань, що ускладнює систему і погіршує результат.

Одними з перспективних аналітичних методів є статистична фільтрація у часовій області. Фільтрування моделюється авторегресією, наприклад, за допомогою побудови оптимального лінійного фільтра. Відомі експериментальні випробування на мовному сигналі в суміші з адитивним білим шумом показують збільшення відношення сигнал/шум на 4…6 дБ Авторегресійна модель не має виражених дефектів як музичний шум, проте, артефакти обробки також мають місце.

Є методи, що застосовують в процесі очищення мовного сигналу від шумів його статистичну модель у вигляді прихованого марківського кола, що пов'язане з фонетичною структурою сигналу. Спочатку, по записам незашумленого мовного сигналу будуються статистичні моделі одиниць мовного потоку (фонів чи ширших класів звуків). Після того, за цією моделлю можна розрахувати оптимальний фільтр Вінера. Відомі експерименти на різних типах шумів (білий, шум гелікоптера, одночасна розмова декількох дикторів) показали збільшення відношення сигнал/шум в середньому на +7 дБ. Оцінка якості звучання обробленого сигналу на слух за п'ятибальною шкалою показала результат «добре».

Поширеним методом, заснованими на використанні спектральних характеристик шуму, є алгоритм віднімання амплітудних спектрів. Обробка відбувається в два етапи: на першому отримується амплітудний спектр шума

з тієї ділянки, де сигнал відсутній. Другий – відбувається віднімання отриманого спектра від спектра суміші сигнала з шумом, фазовий спектр при цьому лишають незмінним. Варіації цього методу відрізняються частотним розрізненням, порогом і величиною шумоподавлення в окремих смугах.

На думку аудиторів, оброблена мова звучить чистіше та приємніше (попри наявність характерних ефектів обробки – так званих «музичних шумів») ніж до обробки, проте помітного підвищення розбірливості у разі адитивних широкосмугових шумів немає, хоча відношення сигнал/шум підвищується на 6…12 дБ.

В останні роки тема нейромережевого шумоподавлення стає дедалі популярнішою і поки що не зупиняється на досягнутому. Підхід для шумоподавлення нейромережею складається з 3-х компонентів: кодер, розділення, декодер. Етап розділення вирішує задачу наближеного обчислення джерел, суміш яких ми розглядаємо як «забруднений» сигнал.

Автори багатьох алгоритмів використовують існуючі архітектури для сегментації зображень. В цьому випадку кодер – це обчислювач спектрограми, а декодер – перетворювач із спектрограми в хвильоформу.

Існують алгоритми, основані на розпізнаванні наявності мовної події і генерації, тобто синтезі, на основі цієї події чистої мови зі спектрально-часовими характеристиками, наближеними до оригіналу.

В третьому розділі проведено тестування програмних методів шумоочищення, підвищення якості та розбірливості мови. Заготовлено аудіофайли мови диктора (студійний запис, без шумів, з мінімумом реверберації); стаціонарний шум (працюючий холодильник); змінний шум (вентилятор, що змінює положення); переривчастий шум (кухонний блендер, що інколи включається в сусідній кімнаті); імпульсний шум (набор тексту на клавіатурі).

З даних записів в програмі Audacity мішкувалися зашумлені доріжки з відношенням сигнал-шум 24 дБ. Для стаціонарного і змінного шумів контролювалося середньоквадратичне відношення сигнал/шум, для

імпульсного і переривчастого шумів – пікове відношення сигнал/шум.

З точки зору класичних завдань радіотехніки – 24 дБ це вагоме значення. Але з точки зору звукозапису та забезпечення якісного звуку для слухачів, це мало. Тобто, умови задачі шумоочищення обрані коректними.

Для шумоподавлення методом спектрального віднімання використано шумоподавлювач Noise Reduction програми Audacity. Зліпок шуму для нестаціонарних шумів брався з тих часових проміжків, де сигнал був мінімальний, а шум найбільш представницький.

Як слідує з аналізу отриманих спектрограм, найкраще шумоподавлення методом спектрального віднімання спостерігається для стаціонарного і змінного шуму. При цьому корисний сигнал зазнає мінімальних спотворень.

Найгірша ситуація з імпульсними шумами. Щоби забезпечити їх помітне подавлення, необхідно обирати дуже високий поріг в усіх смугах частот, що призводить до сильних втрат оброблюваного сигналу (чорні області на спектрограмі), а імпульсний шум все одно при цьому залишається помітним.

Досліджено шумоподавлення за допомогою нейронної мережі Adobe Podcast. Користуватися даним онлайн сервісом можна безкоштовно.

Як слідує з отриманих спектрограм, гарне шумоподавлення спостерігається для всіх видів шуму, як стаціонарного, так і нестаціонарного, включаючи імпульсний. При цьому зберігаються всі особливості корисного сигналу, як на слух, так і на спектрограмі. Крім того, звертає на себе увагу, що нейронна мережа робить спектр мови більш рівномірним, але залишає всі його характерні особливості. Різкий обрив спектру на 10 кГц в обробленому варіанті перетворюється на екстрапольований спектр, який плавно спадає до 20 кГц. Причому ця екстраполяція відбувається не завжди, а там де треба, тобто на різких транзієнтах мовного сигналу.

На слух шумоочищення нейронною мережею відрізняється від шумоподавлення методом спектрального віднімання. По-перше, відсутній музичний шум, а звук стає схожим на той, який записано в студійних умовах.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Lukin A.S. AES San Francisco 2008: Tutorial T3. Broadband noise reduction: theory and application // Audio Engineering Society. October 2–5, 2008. [Електронний ресурс]. URL: https://www.aes.org/events/125/tutorials/session.cfm?code=T3 (дата звернення – 08.11.2023).

2. Pascual S., Bonafonte S., Serra J. SEGAN: speech enhancement generative adversarial network // Interspeech 2017. Stockholm: ISCA, 2017. P. 3642–3646.

3. Gabbay A., Shamir A., Peleg Sh. Visual speech enhancement // The Hebrew University of Jerusalem. 2018. P. 1170–1174.

4. Schoenenberg K., Raake A., Koeppe J. Why are you so slow? Misattribution of transmission delay to at-tributes of the conversation partner at the far-end // International Journal of Human-Computer Studies. 2014. Vol. 72. Issue 5: May. P. 477–487.

5. Волощук Ю.І. Сигнали та процеси у радіотехніці: Підручник для студентів вищих навчальних закладів, том 1. - Харків: «Компанія СМІТ», 2003. – 580 с.

6. Волощук Ю.І. Сигнали та процеси у радіотехніці: Підручник для студентів вищих навчальних закладів, том 2. - Харків: «Компанія СМІТ», 2003. – 444 с.

7. Audio, NTi. Unattended Noise Monitoring (PDF). [Електронний ресурс]. URL: www.nti-audio.com (дата звернення – 08.11.2023).

8. Alan V. Oppenheim, Ronald W. Schafer. Digital Signal Processing. Prentice-Hall, 1975. – 585 p.

9. Khan F., Milner B.P. Speaker separation using virtually-derived binary masks. Auditory-Visual Speech Processing. Annecy, ISCA, 2013, pp. 215–220.

10. Heymann J., Drude L., Haeb-Umbach R. Neural network based spectral mask estimation for acoustic beamforming // Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing. Shanghai, China, 2016. P. 196–200.

11. Brandstein M., Ward D. Microphone Arrays: Signal Processing Techniques and Applications. Springer, 2001. 398 p.

12. Curtis R.A. Niederjohn R.E. Several Frequency Domain Processing Methods for Enhansing the Intelligibility of Speech in Wideband Random Noise, Proc.1978 IEEE Int Conf on ASsP, pp. 602-605.

13. Drucker H. Speech Processing in a High Ambient Noise Environment. IEEE Trans. On ASSP, ICASSP-76, pp.251-253.

13. Cappe O. Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. IEEE Trans Speech Audio Process., Vol 2., No.2, pp.345-349.

14. В.Н. Олейников, О.В. Зубков, В.М. Карташов, И.В. Корытцев, С.И. Бабкин, С.А. Шейко, И.С. Селезнев. Экспериментальная оценка эффективности алгоритмов пеленгования беспилотных летательных аппаратов по акустическому излучению. Радиотехника: Всеукр. межвед. науч.-техн. сб. – 2019. – Вып. 199. – С. 29 – 37.

15. V. Kartashov, V. Oleynikov, I. Koryttsev, S. Sheiko, O. Zubkov, S. Babkin, I. Selieznov. Use of Acoustic Signature for Detection, Recognition and Direction Finding of Small Unmanned Aerial Vehicles. 2020 IEEE 15th International Confer-ence on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET). 2020. 4 p.

16. Kartashov V.M., Oleynikov V.N, Zubkov O.V., Korytsev I.V., Babkin S. I., Sheiko S.A., Kolendovskaya M.M. Spatial-temporal Processing of acoustic Signals of Unmanned Aerial Vehicles/ Telecommunications and Radio Engineering. – New York. – 2020. – Vol. 79, №9. – P.769-780.

17. V. Kartashov, V. Oleynikov , I. Koryttsev, S. Sheiko, O. Zubkov, S. Babkin. Processing of Wide Band Acoustic Signals During Detection of Unmanned Aerial Vehicles // 2020 IEEE Ukrainian Microwave Week (UkrMW). Kharkiv, Ukraine, September 21 - 25, 2020. Volume 1 on 2020 IEEE 12th International Conference on Antenna Theory and Techniques (ICATT). pp. 35-39.

18. V.M. Kartashov, G.I. Sidorov, S.A. Sheiko, M.M. Kolendovskaya, O.Yu. Sergienko. Principles of construction and assessment of technical characteristics of multi-frequency atmospheric sodar in the humidity measurement mode. Telecommunications and Radio Engineering. Vol. 79. N.4. 2020. – pp. 323-333.

19. S. Sheiko. Study of the method for assessing atmospheric turbulence by the envelope of sodar signals // Eastern-European Journal of Enterprise Technologies. – 2/5 (92). – April, 2018. – p. 33–40.

20. Сідоров Г.І., Шейко С.О., Шаповалов С.В., Полонська А.С., Дмитренко А.І. Акустичний метод вимірювання турбулентного стану атмосферного прикордонного шару // Радиотехніка: Всеукр. міжвід. наук.-техн. зб. 2018. – Вип. 192. – С. 46–50.

21. Valerii V. Semenets, V. M. Kartashov, V. I. Leonidov. Registration of refraction phenomenon in the problem of acoustic sounding of atmosphere in airports zone. Telecommunications and Radio Engineering. Volume 77, Issue 5, 2018. – P. 461-468.

22. Бабак К.В. Технічні аспекти створення електронної музичної композиції // 27-й Міжнародний молодіжний форум «Радіоелектроніка та молодь у XXI столітті». Зб. матеріалів форуму. Т. 3. – Харків: ХНУРЕ. 2023. – с. 57-58.

23. Свірідок М.С. Технічні аспекти створення музичної композиції // 27-й Міжнародний молодіжний форум «Радіоелектроніка та молодь у XXI столітті». Зб. матеріалів форуму. Т. 3. – Харків: ХНУРЕ. 2023. – с. 104-105.

24. Курдиш В.В. Алгоритм синхронізації звуку і відео в інтерв'ю // 27-й Міжнародний молодіжний форум «Радіоелектроніка та молодь у XXI столітті». Зб. матеріалів форуму. Т. 3. – Харків: ХНУРЕ. 2023. – с. 129-130.

25. Древальський Р.В. Дослідження методу корекції звука для компенсації впливу приміщення /25-й Міжнародний молодіжний форум «Радіоелектроніка та молодь у XXI столітті». Зб. матеріалів форуму. Т. 3. – Харків: ХНУРЕ. 2021. – с. 119-120.

26. Удовік Д.В. Дослідження методів зменшення еквівалентної реверберації в звукозаписі: кваліфікаційна робота на здобуття освітнього ступеня магістра. – Х.: ХНУРЕ. – 2022 р. – 65 с.

27. Тарусін В.Ю. Дослідження методів компенсації спотворень в звукових трактах: кваліфікаційна робота на здобуття освітнього ступеня магістра. – Х.: ХНУРЕ. – 2022 р. – 78 с.

28. Мезенцев І.О. Дослідження алгоритмів автоматизованої еквалізації звукозапису голосу: кваліфікаційна робота на здобуття освітнього ступеня магістра. – Х.: ХНУРЕ. – 2022 р. – 69 с.

29. Widrow B., at al. Adaptive Noise Candelling: Principles and Applications. Proc. IEEE, Vol. 63, No. 12, 1975, pp.1672-1716.

30. Whipple G. Low Residual Noise Speech Enhancement Utilizing Time-Frequency Filtering, Proc of ICASSP'94, p5-9.

31. McWhirer J.S., Palmer K.J., Robers J.B. A Digital Adaprive Noise-Canceller Based on a Stabilizer Version Version of the Widrow L.M.S. Algorithms, Proc. 1982, IEEE Int. Conf. ASSP, pp.1384-1387.

32. Гурьев Ю.Ю. Прохоров Ю.Н. Алгоритм рекуррентной фильтрации речевых сигналов. Материалы Всесоюзного семинара АРСО-12. Киев, 1982, с. 39-42.

33. Andrew P. Sage, James L. Melsa. Estimation Theory with Applications to Communications and Control. McGraw-Hill, 1971. – 529 p.

34. Lee K.Y., Lee B.-G., Song I. etc. Recursive Speech Enhancement Using the EM Algorithm with Initial Conditions Trained by HHM's. Int Conf on Acoustics, Speech and Signal Proc, ICASSP-96, 1996, pp.621-624.

35. Hansen G.H.L, Pellom B.L. Text-directed speech enhancement employing phone class parsing and feature map constrained vector quantizaion. Speech Communication, Vol. 21, 1997, pp. 169-189.

36. Sheikhzadeh H., Sameti H., Deng L. Comparative Performance of Spectral Subrtraction and HMM Based Speech Enhancement Strategies with Application to Hearing Aid Design. Proc. ICASSP-94, p. 13-17.

37. Boll S.F. Suppression of Acoustic Noise in Seech Using Spectral Subtraction. IEEE Trans. ASSP, Vol. 27, No.2, 1979, pp. 113-120.

38. Hoy L.D., etc. Noise Suppression Methods for Speech Applications. Proc. 1983 IEEE Int. Conf. ASSP, ICASSP-83, pp. 1133-1136.

39. Ephraim Y., Malah D. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator, IEEE Trans Acoust, Speech and Signal Process, 1984 IEEE Int. Conf. ASSP, ICASSP-84, pp.1109-1121.

40. Cappe O. Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. IEEE Trans Speech Audio Process., Vol. 2., No.2, pp.345-349.

41. Scalart P. Speech Enhancement Based on a Priori Signal to Noise Estimation. Proc. Int. Conf on Acoustics, Speech and Signal Proc. ICASSP-96, 1996, pp.629-632.

42. Y. Luo and N. Mesgarani. TaSNet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation // 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 696-700.

43. Oord, A.V., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A.W., & Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio. ArXiv, abs/1609.03499.

44. Yi Luo and Nima Mesgarani. Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation. IEEE/ACM Trans. Audio, Speech and Lang. Proc. 27, 8 (August 2019), 1256–1266.

45. Défossez, Alexandre, Gabriel Synnaeve and Yossi Adi. Real Time Speech Enhancement in the Waveform Domain. ArXiv abs/2006.12847 (2020).

46. Su, Jiaqi, Zeyu Jin and Adam Finkelstein. HiFi-GAN: High-Fidelity Denoising and Dereverberation Based on Speech Deep Features in Adversarial Networks. ArXiv abs/2006.05694 (2020).

47. Методичні вказівки з виконання атестаційної магістерської роботи за спеціальністю 8.05090102 «Апаратура радіозв'язку, радіомовлення і телебачення». Освітньо-кваліфікаційний рівень – магістр / Упоряд. В.М. Карташов, В.А. Тихонов, І.В. Савченко – Харків: ХНУРЕ, 2012. – 68 с.