НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»
ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ
Кафедра інформаційної безпеки

«До захисту допущено»
В.о. завідувача кафедри

_____ М.В.Грайворонський
(підпис)

“_____”_____2019 р.

Дипломна робота

на здобуття ступеня бакалавра

з напряму підготовки 6.170101 «Безпека інформаційних і комунікаційних систем»

на тему: Генерація цільового голосу людини з використанням нейронних мереж

Виконав: студент 4 курсу, групи ФБ-52
(шифр групи)

Максименко Олег Анатолійович _____
(прізвище, ім’я, по батькові) (підпис)

Керівник к. т. н., доц. каф. ІБ, Родіонов А. М. _____
(посада, науковий ступінь, вчене звання, прізвище та ініціали) (підпис)

Консультант _____ _____ _____
(назва розділу) (посада, вчене звання, науковий ступінь, прізвище, ініціали) (підпис)

Рецензент к. т. н., доц. каф. ТК, Корнага Я. І. _____
(посада, науковий ступінь, вчене звання, науковий ступінь, прізвище та ініціали) (підпис)

Засвідчую, що у цій дипломній роботі немає запозичень з праць інших авторів без відповідних посилань.
Студент _____
(підпис)

Київ - 2019 року

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»
ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ
Кафедра інформаційної безпеки

Рівень вищої освіти – перший (бакалаврський)

Напрям підготовки 6.170101 «Безпека інформаційних і комунікаційних систем»

ЗАТВЕРДЖУЮ
В.о. завідувача кафедри
_____ М.В.Грайворонський
(підпис)
«___»_____2019 р.

ЗАВДАННЯ
на дипломну роботу студенту

Максименку Олегу Анатолійовичу
(прізвище, ім'я, по батькові)

1. Тема роботи:
«Генерація цільового голосу людини з використанням нейронних мереж»,
науковий керівник роботи:
к. т. н., доц. каф. ІБ, Родіонов Андрій Миколайович,
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

затверджені наказом по університету від «27.05 » 2019 р. № 1414-с

2. Термін подання студентом роботи: 10 червня 2019 р.

3. Вихідні дані до роботи: методи машинного навчання для синтезу мовних сигналів за текстовими послідовностями.

Зміст роботи:
- Огляд існуючих підходів до генерації індивідуалізованого мовлення та методів розпізнавання мовця
- Спектральний аналіз мовних сигналів та їх представлення в якості тренувальних даних
- Побудова архітектури нейронної мережі
- Оцінювання та аналіз результатів

5. Перелік ілюстративного матеріалу (із зазначенням плакатів, презентацій тощо):

- Презентація

6. Дата видачі завдання: 7 вересня 2018 р.

Календарний план

№ з/п	Назва етапів виконання дипломної роботи	Термін виконання етапів дипломної роботи	Примітка
1	Аналітичний огляд наукової літератури за темою роботи	08.09.18 - 10.10.18	
2	Аналіз основних підходів до копіювання голосу людини та методів розпізнавання мовця.	10.10.18 - 26.11.18	
3	Написання оглядового розділу дипломної роботи	26.11.18 - 11.12.18	
4	Вивчення способів акустичного аналізу мовних сигналів та їхнього представлення	11.12.18 - 17.01.19	
5	Пошук, обробка та підготовка тренувальних даних	17.01.19 - 29.02.19	
6	Написання другого розділу дипломної роботи	29.02.19 - 13.03.19	
7	Побудова архітектури нейронної мережі та її програмна реалізація	13.03.19 - 15.04.19	
8	Проходження переддипломної практики (Розробка послідовності виконання експериментального дослідження та його проведення)	15.04.19 - 17.05.19	
9	Аналіз, узагальнення та оцінка результатів. Написання третього розділу дипломної роботи	17.05.19 - 30.05.19	
10	Передзахист дипломної роботи	30.05.19	
11	Доопрацювання й оформлення роботи та презентації	30.05.19 - 20.06.19	
12	Захист дипломної роботи	20.06.19	

Студент _____ О. А. Максименко
(підпис) (ініціали, прізвище)

Науковий керівник роботи _____ А. М. Родіонов
(підпис) (ініціали, прізвище)

РЕФЕРАТ

Генерація цільового голосу людини з використанням нейронних мереж // Дипломна робота бакалавра Максименка О.А. Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Фізико-технічний інститут, кафедра інформаційної безпеки, група ФБ-52. – К.:НТУУ «КПІ імені Ігоря Сікорського», 2019. с. – 62, рис. – 9, табл. – 9.

В умовах сьогодення з метою підвищення рівня надійності і безпеки доступу створюються спеціальні біометричні системи ідентифікації та верифікації, що використовують унікальні фізіологічні та поведінкові властивості людини (ДНК, райдужна оболонка ока, обличчя, відбитки пальців тощо). Популярність цих технологій пояснюється тим, що вони менш схильні наражатися на атаки, оскільки організувати атаки на них досить складно.

Метою даної роботи було синтезування мовлення людини з тексту зі збереженням його індивідуальних особливостей за допомогою нейронної мережі. Об’єктом дослідження є нейронні мережі та їхні архітектури. Предметом – синтез мовних сигналів із використанням методів машинного навчання та нейронних мереж.

Досягненню мети і вирішенню поставлених завдань сприяло використання комплексу теоретичних та емпіричних методів дослідження.

У роботі запропоновано нейронну мережу, яка може за короткий проміжок часу переналаштовуватися для генерації нових голосів, потребуючи при цьому невеликий об'єм даних для адаптації. Було досліджено моделі згорткових нейронних мереж з механізмом уваги, які можуть замінити рекурентні нейронні мережі у вирішенні задач синтезу розбірливого і наближеного до природного мовлення людини зі збереженням індивідуального голосу та вимови.

У першому розділі було розглянуто основні переваги та недоліки підходів до моделювання процесів синтезу та модифікації мовних сигналів, а також наведено конкретні приклади реалізацій цих підходів.

У другому розділі було описано способи обробки мовних сигналів з метою використання їх для виділення корисної інформації з сигналів, яку можна використати в якості характерних ознак для тренування нейронної мережі. Також розглянуто алгоритм, який дозволяє навпаки з цих ознак відтворити сигнал.

У третьому розділі детально охарактеризовано процес побудови архітектури нейронної мережі, описано основні організаційні кроки проведення експериментальних досліджень для досягнення поставленої в даній роботі мети. Після цього було здійснено суб’єктивне та об’єктивне оцінювання якості роботи запропонованої нейронної мережі.

Практичне значення одержаних результатів полягає у розробці нейронної мережі синтезу особистісного мовлення за текстом, яку можна використовувати як інтерфейс виведення даних для багатьох систем, а також в якості засобу перевірки біометричних систем ідентифікації за голосом.

Ключові слова: нейронна мережа, машинне навчання, розпізнавання мовця, генерація мовлення, обробка мовних сигналів, голосова конверсія.

ABSTRACT

Generation of target speaker's voice using neural networks // Thesis for bachelor's degree by Maksymenko O. National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Institute of Physics and Technology, Department of Information Security, group FB-52. - K.: NTUU "Igor Sikorsky Kyiv Polytechnic Institute", 2019. p. - 62, fig. - 9, tab. - 9.

Nowadays special biometric identification and verification systems using unique physiological and behavioral properties of a person (DNA, iris, face, fingerprints, etc.) are created in order to increase the level of reliability and security of access. The popularity of these technologies caused by the fact that they are less prone to attack, because it is difficult to compromise systems.

The purpose of this work was to synthesize the person's speech from the text preserving his/her individual characteristics with the help of a neural network.

The object of the study is neural networks and their architectures. The subject is the synthesis of speech signals with the help of machine learning methods and neural networks.

The achievement of the goal and the solution of the tasks has been facilitated by the use of the complex of theoretical and empirical research methods.

The paper proposes a neural network that can be re-configured for a short period of time to generate new voices, while requiring a small amount of data to adapt. We investigated the models of convolutional neural networks with the attention mechanism that can replace recurrent neural networks in solving the problems of synthesizing intelligible and approximate to person's natural speech, preserving individual voice and pronunciation.

In the first chapter, the main advantages and disadvantages of approaches to modeling processes of synthesis and modification of speech signals have been considered, as well as concrete examples of implementation of these approaches.

The second section describes how to process speech signals in order to use them to extract useful information from signals, which can be used as the

characteristic feature for training the neural network. Also we considered the algorithm that reconstructs the signal out of these features.

The third section describes in detail the process of constructing the architecture of the neural network, depicts the main organizational steps of conducting experimental research to achieve the goal set in this paper. A subjective and objective evaluation of the quality of the proposed neural network has been carried out.

The practical value of the results obtained consists in the development of an individual's speech synthesis neural network, which can be used as data output interface in many systems, as well as the means of testing biometric voice recognition systems.

Key words: neural network, machine learning, speaker recognition, speech synthesis, speech signal processing, voice conversion.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

STFT – Short-time Fourier transform – віконне перетворення Фур’є

Мел – психофізична одиниця висоти звуку (тону)

MFC – mel-frequency cepstrum – мел-частотний кепстр

MFCC – mel-frequency cepstral coefficients – коефіцієнти, які в сукупності утворюють мел-частотний кепстр

CNN – convolutional neural network – згорткова нейронна мережа

RNN – recurrent neural network – рекурентна нейронна мережа

GLA – Griffin-Lim algorithm – алгоритм Гріффіна-Ліма.

VAE – variational auto encoder – варіаційний автокодувальник

ReLU – rectified linear unit – випрямлена лінійна функція

GMM – Gaussian Mixture Model модель – суміш Ґаусівських розподілів

DTW – Dynamic Time Warping – алгоритм динамічної трансформації часової шкали

DFW – Dynamic Frequency Warping – алгоритм динамічної трансформації частотної шкали

HMM – hidden Markov model – прихована модель Маркова

ВСТУП

Актуальність роботи. Біометрія є методом, що пов'язує сукупність цифрових даних і унікальних біологічних характеристик людини. Її використання поширюється для організації безпечного та зручного доступу до різноманітних сервісів і захисту цифрових даних. В умовах зростаючої технічної складності навколишнього середовища зручний та безпечний доступ до цифрових даних стає особливо актуальним. З метою підвищення рівня надійності і безпеки доступу створюються спеціальні біометричні системи ідентифікації та верифікації, що використовують унікальні фізіологічні та поведінкові властивості людини (ДНК, райдужна оболонка ока, обличчя, відбитки пальців тощо). Популярність цих технологій пояснюється тим, що вони менш схильні наражатися на атаки, оскільки організувати атаки на них досить складно.

Зокрема, серед біометричних методів розпізнавання є ідентифікація за голосом. Саме цей метод характеризується простотою і зручністю у використанні, адже для реалізації даного методу не потрібна дорога апаратура – досить мікрофона та звукової плати. В сучасних реаліях ця технологія достатньо швидко розвивається і широко використовується в сучасних бізнес-центрах, банківських сервісах, системах голосової пошти тощо. Такі системи розпізнавання можуть бути використані для контролю доступу, телефонного банкінгу, біометричних досліджень, розслідування злочинів. Існує ряд комерційних, організаційних, персональних автоматичних систем розпізнавання мовця, наприклад T-NETIX, ITT, Lernout & Hauspie, Veritel and voice control system [1]. Дослідження засвідчують, що технологія voice FONCARD корпорації Sprint є однією з найбільших за масштабами розгортання біометричних систем [1].

Проте основним недоліком таких систем розпізнавання є їх невисока точність. Наприклад, людину з хворобою система може не розпізнати. Голос людини може змінюватись у залежності від стану її здоров'я, віку, настрою

тощо. Таке різноманіття факторів впливу створює серйозні труднощі при виявленні унікальних властивостей голосу людини і створенні високоякісної біометричної системи. Особливо важливим фактором є імовірність помилки другого роду, яка виникає при атаках на систему, а тому перед експлуатацією технології в якості готового продукту її потрібно ретельно протестувати. Одним із найнебезпечніших типів атак є атака на основі синтезу мовлення, тому доцільною є перевірка системи на здатність виявляти синтезоване мовлення і відрізняти його від природного.

Створення якісного синтезатора, що може відтворювати за текстом індивідуалізоване мовлення людини, є також непростою задачею.

Останнім часом використання нейронних мереж у розробці статистичних моделей для вирішення нетривіальних задач стало дуже поширеним. Генеративні моделі, що ґрунтуються на глибоких нейронних мережах, успішно застосовуються в багатьох галузях досліджень, таких як генерація і класифікація зображень, мовне моделювання тощо. Глибокі нейронні мережі здатні моделювати складні процеси та розподіли даних і можуть бути додатково обумовлені зовнішніми входами для управління змістом і стилем згенерованих вибірок.

Проблемою синтезу мовлення за текстом займались Б. Богерт, З. Ву, Д. Девіс, Б. М. Лобанов, Т. В. Людовик, О. Ф. Кривнов, Я. Стиліану, О. Караалі, Х. Фуджісакі та інші.

У складних завданнях завдяки використанню нейронних мереж стало можливим досягти навіть кращих результатів, аніж за допомогою класичних методів, таких як розпізнавання мовлення, машинний переклад тощо. Головним фактором, що призводить до підвищення якості й ефективності алгоритмів, є здатність мереж узагальнювати.

З огляду на всі ці фактори, було вирішено розробити нейронну мережу в якості засобу перевірки біометричних систем верифікації на здатність виявляти синтезоване мовлення, що приведе до зменшення похибки другого роду.

Метою дослідження є синтезування мовлення людини з тексту зі збереженням його індивідуальних особливостей за допомогою нейронної мережі.

Досягнення визначеної мети дослідження передбачає вирішення таких **завдань:**

- опрацювати наукову літературу за темою роботи;
- дослідити методи спектрального аналізу мовних сигналів;
- зібрати, обробити та підготовити тренувальні дані;
- спроектувати архітектуру нейронної мережі;
- виконати тренувальні експерименти;
- проаналізувати, узагальнити та оцінити результати.

Об’єктом дослідження є нейронні мережі та їхні архітектури.

Предметом дослідження є синтез мовних сигналів із використанням методів машинного навчання та нейронних мереж.

Досягненню мети і вирішенню поставлених завдань сприяло використання комплексу **методів** дослідження:

– *теоретичних*: аналіз основних положень машинного навчання, порівняльний метод і системний аналіз наукових джерел для з’ясування стану розробленості проблеми розробки нейронних мереж та визначення базових понять дослідження; аналіз; синтез, узагальнення й концептуалізація для формування основних положень дослідження; метод моделювання, застосування якого дало змогу розробити модель синтезу особистісного мовлення за текстом;

– *емпіричних*: узагальнення результатів експериментального дослідження синтезованого нейронною мережею мовлення, статистичне оброблення експериментальних даних з використанням методів статистичного аналізу.

Наукова новизна. У роботі запропоновано нейронну мережу, яка може за короткий проміжок часу переналаштовуватися для генерації нових голосів, потребуючи при цьому невеликий об'єм даних для адаптації. Було

досліджено моделі згорткових нейронних мереж з механізмом уваги, які можуть замінити рекурентні нейронні мережі у вирішенні задач синтезу розбірливого і наближеного до природного мовлення людини зі збереженням індивідуального голосу та вимови.

Практичне значення одержаних результатів полягає у розробці нейронної мережі синтезу особистісного мовлення за текстом, яку можна використовувати як інтерфейс виведення даних для багатьох систем, а також в якості засобу перевірки біометричних систем верифікації за голосом.

1 ОГЛЯД ІСНУЮЧИХ МЕТОДІВ КОПІЮВАННЯ МОВЛЕННЯ ТА СИСТЕМ ГОЛОСОВОГО РОЗПІЗНАВАННЯ

Існує досить значна кількість способів створення голосового шаблону людини. Зазвичай, це різні комбінації частотних і статистичних характеристик голосу, наприклад інтонація чи висота тону.

Найпростішим способом підробки голосу є використання диктофону, на якому записано голос обраної людини. Однак сучасні біометричні системи запропонують або сказати деякий невідомий пароль, або промовити певну фразу. В такому випадку для нападу на систему доведеться використовувати синтезатор мови. До основних типів атак на системи розпізнавання по голосу належать:

- запис голосових біометричних характеристик людини з їх подальшим повторенням;
- перетворення мовлення зловмисника в мовлення іншої людини;
- синтез мови.

Від першого типу атаки, як зазначалося вище, легко захиститись шляхом запиту промовити парольну фразу.

Найбільшу небезпеку становлять метод голосової конверсії (модифікації) та метод атаки на основі синтезу мовлення з характеристиками, властивими голосу користувача біометричної системи.

Голосова конверсія (модифікація) – це техніка перетворення мовлення однієї людини (джерела) таким чином, щоб воно звучало подібно до мовлення якоїсь іншої людини (цілі). Для цього модифікується специфічна нелінгвістична інформація мовного сигналу, в той час як лінгвістична інформація зберігається незмінною. Іншими словами, змінюються такі властивості голосу людини як тон, висота, тембр, сила, інтенсивність тощо, але при цьому зміст сказаного (фонеми, склади, слова) залишається сталим.

Синтез мовлення – відновлення мовного сигналу за його параметрами. Найрозповсюдженішою практикою є формування мовного сигналу на основі

друкованого тексту. Загалом способи синтезу мови поділяються на три основні типи: конкатенативний, параметричний, синтез за правилами. Проте кожен підхід має свої недоліки та переваги.

Найпростішим є конкатенативний, який ґрунтується на з’єднуванні початкових елементів синтезу із заздалегідь записаного словника у цілісні повідомлення. Незважаючи на ідейну простоту, цей метод складний для реалізації, оскільки в місцях складання елементів чутно розриви.

При використанні параметричного синтезу, змінюючи характеристики, можна реалізовувати моделювання емоційного забарвлення тексту. Цей тип є більш гнучким внаслідок параметризації на основі дрібних фонетичних одиниць.

Синтез за правилами дозволяє управляти усіма параметрами сигналу, як наслідок, це дає можливість синтезувати мовлення за невідомим заздалегідь текстом. Генерація мови реалізується як моделювання голосового тракту людини з використанням цифрової чи аналогової техніки. Параметри та правила поєднання звуків, отримані при аналізі мовних сигналів зберігаються для подальшого використання. Під час генерування сигналу значення параметрів і правила застосовуються послідовно з певним інтервалом.

1.1 Модифікація і синтез мовних сигналів за допомогою ймовірнісних і статистичних моделей та алгоритмів

Одним зі стандартних підходів для голосової конверсії є використання трансформацій, що ґрунтується на GMM (модель-суміш Ґаусівських розподілів) [2, 3].

Прикладом реалізації такої техніки може слугувати програмне забезпечення “sprocket” [4]. Автори статті Kazuhiro Kobayashi, Tomoki Toda описують принцип роботи такої моделі. Ідея полягає в моделюванні функції

густини ймовірнісного розподілу, що складається з суміші багатьох Ґаусівських розподілів (Gaussian Mixture Model).

Невідомі параметри для шуканого розподілу оцінюються за допомогою методу максимізації функції правдоподібності [5]. Процес моделювання складається з декількох етапів. Перш за все потрібно підготувати тренувальну вибірку (датасет), що містить «паралельні» вирази мовців, тобто такі що мають однакову лінгвістичну інформацію. Авторами був використаний датасет VCC 2016. Далі обраховують такі ознаки мовних сигналів як частота основного тону (форманта), аперіодичність та мел-кепстр, а також такі статистики отриманих ознак як середнє значення, стандартне відхилення логарифму частоти основного тону та глобальну дисперсію (Global variance) мел-кепстру. Для моделювання функції потрібні покадрово вирівняні вектори ознак, які отримують за допомогою алгоритму динамічної трансформації часової шкали (Dynamic Time Warping) – алгоритм, що дозволяє знайти оптимальну відповідність між часовими послідовностями Q і C:

$$DTW(Q,C) = min\left\{\frac{\sum_{k=1}^{K} d(w_k)}{K}\right\} \quad (1.1)$$

$Q = q_1, q_2, ..., q_i, ..., q_n$;

$C = c_1, c_2, ..., c_i, ..., c_m$;

$d(w_k) = d(q_i, c_j)$ – евклідова відстань між точками q_i, c_j;

K – довжина шляху, який мінімізує загальну відстань між Q і C.

DTW вирівнює спектрограми за часом, обчислюючи показники подібності між сегментами мовних сигналів, спотворює часові шкали спектрограм так, що ідентичні мовні події в обох спектрах відбуваються в однакові моменти часу. В результаті для перетворення ознак голосу мовця-джерела в ознаки голосу мовця-цілі використовується змодельована функція (GMM).

Недоліками наведеного методу є потреба в датасеті, що складається з «паралельних» виразів і, що більш важливо, погіршення якості перетворення

мови внаслідок надмірного згладжування перетворених спектрів, чого неможливо уникнути при використанні такого стандартного підходу.

Окрім GMM і DTW, використовують алгоритм динамічної трансформації частотної шкали (Dynamic Frequency Warping). Аналогічно до DTW, DFT використовують разом з GMM, але цей алгоритм знаходить відповідності між частотними послідовностями. Якість модифікації голосу значно вища, ніж у стандартного підходу, про що зазначається у роботах [6, 7]. Також покращені результати отримано за допомогою використання алгоритму названого Weighed Frequency Warping [8].

Окрім цього для модифікації голосу часто використовують синусоїдальну модель [8]:

$$s[n] = \sum_{l=1}^{L} A_l[n] cos(\phi_l[n]) \qquad (1.2)$$

або лінійні перетворення спектру та масштабування тону голосу [10, 11]. Незважаючи на те, що ці методи досить ефективні, вони також вводять артефакти (спотворення мовних сигналів), які виникають через фазову декогерентність (непослідовність), неприродну дисперсію фаз і високу спектральну дисперсію невокалізованих звуків. На практиці система перетворення голосу повинна це враховувати, якщо потрібно зберегти високу якість звучання звуку. У своїй роботі S. Young і Hui Ye описали нові техніки для вдосконалення такої моделі та прибирання (знешкодження) цих артефактів [12].

Для здійснення параметричного синтезу мовлення найчастіше використовуються системи на основі прихованої моделі Маркова (hidden Markov model) [13, 14]. Прихована модель Маркова – це статистична стохастична модель, у якій система, що моделюється, розглядається як марковський процес із прихованими станами (тобто процес, що випадково змінюється, а майбутні стани залежать лише від поточного, а не від послідовності подій у минулому). Використання систем на основі HMM вважається значно ефективним для синтезу мовлення прийнятного рівня якості. Сигнал у таких системах представляється набором параметрів, які

неперервно змінюються, тобто є континуальними. Основними незаперечними перевагами статистичного параметричного синтезу мовлення є гнучкість, яка полягає в можливості зміни мовних характеристик, тривалості та стилю мовлення, емоцій, у підтримці різних мов. До того ж, така система покриває значний акустичний простір для синтезу, оскільки параметри моделі можуть бути змінені за необхідності як завгодно.

Недоліком таких систем є якість синтезованого мовлення, на яку впливають три фактори: вокодери, точність акустичного моделювання та надмірне згладжування. Синтезоване мовлення звучить гулко, оскільки система використовує мел-кепстральні вокодери. Приховані моделі Маркова працюють якісно, за умов припущень, які висуваються при їх використанні, зокрема таких як постійні значення статистик всередині стану, покадрова умовна незалежність від ймовірностей вихідних станів тощо. Оскільки параметри мовлення напряму генеруються з акустичних моделей, то їх точність впливає на якість синтезованого мовлення. Синтезоване мовлення звучить приглушеним у порівнянні з природною мовою, оскільки згенеровані траєкторії мовних параметрів часто надмірно згладжуються, тобто деталізовані характеристики параметрів втрачаються у процесі моделювання і не можуть бути відновлені під час синтезу.

1.2 Модифікація і синтез мовних сигналів за допомогою методів машинного навчання та нейронних мереж

Аналогічно для вирішення задач модифікації [15, 16] і синтезу мовних сигналів [17, 18, 19] використовуються і нейронні мережі, які імітують функціонування та процеси обробки інформації головним мозком живих організмів і складаються з нейронів, які беруть участь у прийнятті рішень. Принцип роботи окремого штучного нейрона досить простий за своїм змістом: він обчислює зважену суму всіх компонент вхідного вектора $\bar{x}$, використовуючи матрицю вагів $\bar{w}$, а також адитивну складову зсуву $\bar{b}$, після

чого до отриманого результату застосовується деяка (зазвичай нелінійна) функція активації σ.

Так само, як і в ймовірнисних моделях, у нейронних мережах в якості вхідних даних використовуються деякі набори ознак (короткочасна енергія, інтенсивність сигналу, спектральні коефіцієнти тощо), які отримуються з мовних сигналів, але за допомогою мережі будується апроксимація функції багатьох змінних, яка б виконувала поставлену задачу. У випадку голосової конверсії – це функція, яка перетворюватиме вхідний мовний сигнал джерела на мовний сигнал цілі, а у випадку синтезу – вхідний друкований текст (або ж інші дані) на мовний сигнал цілі. Для побудови цієї функції знаходять її параметри за допомогою мережі. Цей процес називається тренуванням. Для тренування мережі обирається певна функція втрат, яка оптимізується. Іншими словами, мережа навчається на досвіді D щодо класу задач P у сенсі міри якості Q, якщо при вирішенні задачі P якість, що визначається мірою Q, зростає при демонстрації нового досвіду D. Сама мережа може складатися як з одного прихованого шару нейронів (просто нейронна мережа), так і з багатьох (глибинна нейронна мережа). Нейронні мережі також варіюються за типом і структурою [20]. Так, наприклад для вирішення задач голосової конверсії та синтезу в роботах H. Zen і Z. Wu [18, 19] використовується глибинна нейронна мережа прямого розповсюдження (Deep Feed Forward NN), у роботі T. Nakashika [16] – мережа під назвою Deep Belief Network (DBN), у C. Hsu і S. Aric [15, 21] – варіаційний автокодувальник (Variational Auto Encoder).

Як зазначається в роботі H. Zen [18], підхід із використанням HMM є достатньо ефективним, але має декілька обмежень, зокрема дерева прийняття рішень неефективні для моделювання складного контексту залежностей. Співвідношення між вхідним текстом і його акустичним представленням, що моделюється DNN, може усунути деякі обмеження цього стандартного підходу. Експериментальні результати засвідчують, що системи на основі DNN перевершують HMM-системи з аналогічною кількістю параметрів.

1.3 Методи голосового розпізнавання

Розпізнавання мовця – це техніка, призначена для автоматичного визначення мовця на основі ознак, виділених з його мови. Мовний сигнал містить різноманітні види інформації і тому здатен ідентифікувати людину. Методи розпізнавання мовця поділяються на ідентифікацію та верифікацію мовця (Рис 1.1) [1]. Вони дозволяють підтримувати достатній рівень безпеки в конфіденційних сферах діяльності та є надзвичайно корисними для перевірки особистості. Системи верифікації встановлюють відповідність один до одного (1:1), в той час як системи ідентифікації - один до багатьох (1:n). При верифікації отриманий зразок мовлення порівнюється з певною еталонною моделлю голосу людини, а при ідентифікації система намагається співставити невідомого мовця з усією базою даних голосів. В свою чергу системи ідентифікації та верифікації також поділяються на текстозалежні та текстонезалежні (Рис 1.1) [1]. Текстозалежна система вимагає надання того самого тексту або виразу для навчання та тестування, в той час як текстонезалежна система цього не потребує.

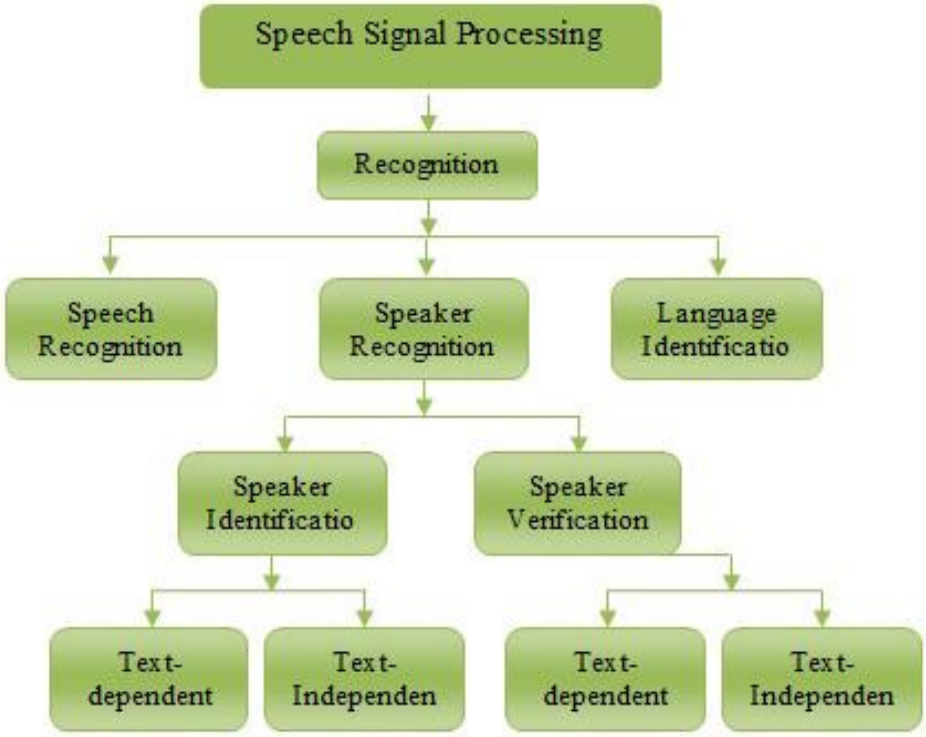

Рисунок 1.1 – Категорії методів розпізнавання мовця

Метою автоматичного розпізнавання є отримання голосу кожного із мовців і створення їх моделей з подальшим порівнянням цих моделей з виразами (зразками мовних сигналів) певного мовця для доведення ідентичності. Зрозуміло, що різні люди мають власні особливості голосу, але навіть голос певної людини може різнитись час від часу. Методи розпізнавання ґрунтуються на здібності людини ідентифікувати голос іншої людини за таких умов:

- людина здатна розпізнати голос будь-якої людини, яку знає і з якою часто спілкується;

- людина може визначити голос іншої людини, з якою вона часто спілкується незалежно від середовища спілкування або фонового шуму;

- людина може впізнати іншу людину, навіть якщо спілкування відбувається після довготривалих перерв (місяці, роки);

- людина здатна відчути і зрозуміти емоційний стан іншої людини, слухаючи її голос.

Високоякісний механізм розпізнавання мовця повинен фокусуватись на факторах, за допомогою яких одна людина впізнає голос іншої. Саме знання того, як люди розпізнають голос, допоможе створювати системи, що будуть все більш і більш точними. Такі системи повинні характеризуватись надійністю, точністю, стійкістю, простотою вимірювання, незалежністю від конкретного голосу і емоційного стану.

За останні шістдесят років було досягнуто значного прогресу в області розпізнавання людини за особливостями її голосу, проте навіть дотепер існує багато проблем, які потребують досконаліших рішень як з технічної так і з теоретичної точок зору, зокрема невідповідність каналів зв’язку та умов запису мовних сигналів. Для розробки ефективної системи потрібно дослідити стабільні в часі параметри голосових ознак, незалежні від різноманіття мовлення, фонового шуму, спотворень, які вносять канали зв’язку.

Загальний огляд досліджень і розвиток в області автоматичного розпізнавання мовця коротко описаний в таблиці 1.1 [1].

Таблиця 1.1 – Розвиток технологій розпізнавання людини за її голосом

Винахідник/ Автор/рік	Організація	База даних	Метод моделювання	Мовні ознаки	Передача голосу	Тип системи	Точність (%)
F. K. Soong, et.al./1985	AT&T Bell Laboratories	50 male and 50 female)	vector quantization (VQ)	short-time spectral	Telephone	Independent	98%
B. S. Atal/ 1974	Bell laboratories	10 speakers	LPC	Cepstrum	Lab	Independent	93%
Colombi, et al./199634	AFIT	138	HMM monophone	Cepstrum	office	Dependent	Error: identification 0.22% (10s) verification 0.28% (10s)
Alfredo Maesa/2012	Voxforge. org	450 speakers	MFCC	spectral subtraction	Audio data- base	Independent/ Identification	>96%
Douglas A. Reynolds/1995	Lincoln Laboratory	49	GMM	Short Utterance	Telephone	Independent/ Identification	96.8%
Rabah W et.al./2004	King Abdulaziz University	20	SVD-based algorithm	LPC/ Cepstral	office	Independent/ Identification	94%
Najim Dehak et.al./2007 35	NIST-2006	NA	GMM-JFA	prosodic features	Lab	language identification	Improvement 8% (all trials) and 12% (English only)

Продовження таблиці 1.1

Винахідник/ Автор/рік	Організація	База даних	Метод моделювання	Мовні ознаки	Передача голосу	Тип системи	Точність (%)
Sharada V. Chougule/2015	Finolex Academy of Management & Technology	97	NDSF	Spectral	Lab	Independent/ Identification	~(98-100)%
Yang Shao et.al./2008	Ohio State University	34(18 male 16 female)	GFCCs	auditory features	Telephone	Independent/ Identification	~99.33 %
Vincent Dubreucq/1994	Digital speech laboratory, RMA	21	HMM	Pitch	Lab	Independent/ Recognition	VER=7.6% RER=7.7%
Douglas A. Reynolds/2001	TIMIT(168), NTIMIT(168), Switchboard (113)	449	GMM	Unconstrained speech	Lab	Dependent/ Recognition	99.7%, 76.2%, 82.8%
Rabah W et.al./2003	King Abdulaziz University	10	SVD-based algorithm	LPC/Cepstral	office	Independent/ Identification	99.5%
P. Krishnamoorthy/ 2011	TIMIT	100	GMM-UBM	MFCC	Lab	Independent/ Identification	80%
Sriram Ganapathy/2014	SRE database (NIST-2010)	random	AR model	FDLP	Lab	Dependent/ Recognition	relative improvements of up to 25%
Hesham Tolba/2011	Arabic speakers	10	HMM/ GHMM	MFCC	Lab	Dependent/ Independent	100%/ 80%

Кінець таблиці 1.1

Винахідник/ Автор/рік	Організація	База даних	Метод моделювання	Мовні ознаки	Передача голосу	Тип системи	Точність (%)
Chih-Hung Chou et. Al./2015	ALTERADE 2-70	16	VQ/GMM-PQ	OOS	Lab	Dependent	Recognition Rate 88.3%
Emmanual Perrin et. Al./1994	E-HERRIOT	60	Acoustical Signature	Vocalic Space	Standard Protocol	Dependent	>90%
Ergun Yucesoy et al./2016	E Gender database INTER SPEECH 2010	299 speakers	GMM-SV	prosodic features	Lab	Dependent	90.4%, 54.1% and 53.5% in gender, age, and age & gender categories
Xuanjing Shen et al./2014	TIMIT speech database	38(19 Female, 19 Male)	LFA-SVM Gaussian kernel	12-order MFCC	Lab	NA	81.52%
Anzar S.M et al./2016	ELDASR	50(Male/ Female)	GMM/ MFCC	MFCC super template	Lab with intra-class variations	NA	Improved (% NA)
Isaias Sanchez-Cortina et al./2016	video Lectures. net, poliMedia	NA	logistic regression model	NB model	Online educational lectures	Dependent	Improvement between 2% and 7%.

Скорочення:

NDSF: Normalized Dynamic Spectral Feature

VER: Verification error rate

RER: Rejection error rate

FDLP: frequency domain linear prediction

SRE :NIST-2010 speaker recognition evaluation database

AR Model: Auto Regressive Model

NB Model : word-dependent naïve Bayes (NB)

JFA: joint factor analysis

Колонка «База даних» містить кількість дикторів, для яких виконувались тести. Колонка «Передача голосу» вказує на те, яким способом було отримано мовні сигнали, в якому середовищі вони поширювались. Колонка «Тип системи» вказує залежність системи від тексту і спосіб перевірки (ідентифікація чи верифікація).

Висновки до розділу 1

Для вирішення задач синтезу мовлення та розпізнавання за голосом використовуються як статистичні математичні моделі, наприклад GMM, HMM, так нейронні мережі. Останній підхід є більш сучасним і, що більш важливо, дозволяє перевершити досягнення та усунути недоліки стандартних підходів, тому доцільним є використання і дослідження можливостей саме нейронних мереж.

Процес моделювання модифікації та синтезу мовних сигналів за допомогою ймовірнисних і статистичних моделей та алгоритмів складається з декількох етапів, що завжди включають в себе виділення характерних ознак з мовних сигналів, знаходження невідомих параметрів шуканого розподілу й оптимізації функції-критерію. Кожен тип моделей модифікації і синтезу голосу має як певні переваги, так і певні недоліки.

За останні десятиліття вченими зроблено значний внесок в розробку методик розпізнавання людини за особливостями її голосу, однак залишається значне поле діяльності для їх удосконалення.

2 АКУСТИЧНИЙ АНАЛІЗ МОВНИХ СИГНАЛІВ

Першим етапом розробки системи синтезу мовлення є акустичний аналіз та обробка мовленнєвого сигналу, який полягає у фільтрації та отриманні набору інформативних ознак до кожного фрагменту сигналу, таких як частотні діапазони формант, висота звуку, тембр, паралінгвістичні властивості голосу, гучність, частота основного тону тощо. Згадані ознаки містять закодовану інформацію про характеристики сигналу на даному фрагменті. Акустична обробка є однією з найважливіших частин створення системи, оскільки від її результатів залежать власне і результати роботи всієї системи. Виділення з сигналу акустичних ознак проводиться з метою зменшення надлишковості сигналу та знаходження найбільш корисної інформації.

Акустичні мовні сигнали, що створюються мовним апаратом людини, містять частоти основних коливань у діапазоні (в середньому) 80 -3000 Гц, а частоти гармонік досягають 8000 Гц. Аудіо дані зазвичай представляються в формі електричних коливань. Вони можуть надходити від мікрофону, що записує акустичні звукові хвилі, або від іншого електронного пристрою. Для перетворення цих коливань у форму, яка може бути оброблена комп’ютером, їх необхідно оцифрувати. Цей процес називається аналогово-цифровим перетворенням. У результаті перетворення неперервний сигнал відображається в ряд дискретних відліків (дискретизація за часом), кожен з яких є цілим числом (квантування за рівнем), що характеризує аналоговий сигнал у даній точці із заданою точністю. Отриманий цифровий аудіо потік кодується за допомогою обраного алгоритму кодування і зберігається на комп’ютері у вигляді файла-контейнера певного формату. Саме такі аудіо файли будуть слугувати даними, з яких будуть виділятися ознаки для тренування нейронної мережі, а саме амплітудна спектрограма та мел-частотний кепстр.

2.1 Спектральний аналіз сигналів

Одним із шляхів дослідження мовного сигналу є спектральний аналіз. Для цього застосовуються алгоритми, що реалізують перетворення Фур’є, а також цифрову фільтрацію.

Метод цифрової фільтрації передбачає найбільш простий і зрозумілий підхід до спектрального розкладу сигналу. Аналізатори, що використовуються в цьому методі, складаються з набору паралельно з’єднаних цифрових полосних фільтрів, кожен з яких займає вузьку смугу, що покриває необхідний діапазон частот, і ці смуги щільно прилягають одна до одної. Для отримання спектру сигналу достатньо зняти значення потужності на виході кожного фільтра і розглядати їх як складові спектру, що відповідають центральним частотам пропускних смуг. Вектор зі спектральними параметрами, що описують мовний сигнал, формується з оцінених значень потужностей, які отримуються покадрово з кожного фільтру. Кожен крок для отримання кадру частково перекриває попередній, що дозволяє відслідковувати всі акустичні явища в мовному сигналі і одночасно забезпечує досить плавну зміну параметрів.

Спектральний аналіз на основі алгоритмів, що реалізують обчислення рядів Фур'є, полягає у застосуванні перетворення Фур'є до мовного сигналу. По суті ця операція ставить у відповідність одній функції дійсної змінної іншу. Отримана функція описує коефіцієнти розкладу вхідної функції на елементарні складові. У випадку з мовним сигналом, який є функцією, залежною від часу, результатом перетворення буде функція, яка представляє розклад сигналу на частотні складові – гармонічні коливання різної частоти.

Перетворення Фур’є – це інтегральне перетворення, що задається такою формулою:

$$f(w) = \int_{-\infty}^{\infty} f(x) e^{-ixw} dx. \qquad (2.1)$$

Оскільки сигнал є дискретним, то він являє собою набір точок із значеннями сигналу і до нього відповідно застосовується дискретне перетворення Фур'є:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-ik\frac{2\pi}{N}n}, (k = 0, \dots, N-1), \quad (2.2)$$

N – кількість значень сигналу, виміряних за період, а також кількість коефіцієнтів розкладу;

x_n – виміряні значення сигналу в дискретних часових точках з номерами n;

X_k – комплексні амплітуди синусоїдальних сигналів, що утворюють вхідний сигнал;

k – індекс частоти.

Окрім цього, для підвищення ефективності та швидкості обчислення шуканих коефіцієнтів розкладу, варто застосовувати швидке перетворення Фур'є. Оптимізація досягається за рахунок використання алгоритму Cooley-Tukey [22].

Проте залишається ще одна перешкода – реальні мовні сигнали не є періодичними послідовностями і постійно змінюються, тому використання перетворення Фур'є безпосередньо до всієї послідовності сигналу буде давати суттєво неточні результати. Вирішити цю проблему можна за допомогою віконного перетворення Фур'є.

STFT (Short-time Fourier transform) – віконне перетворення Фур'є – це різновид перетворення Фур'є, що застосовується для визначення синусоїдної частоти та вмісту фази локальної секції сигналу, що має властивість змінюватися в часі. На практиці процедура обчислення віконного перетворення Фур'є полягає у поділі довгих за часом сигналів на більш короткі сегменти однакової довжини, достатньо малі аби вважатися стаціонарним, а потім обчисленні перетворення Фур'є окремо для кожного з цих сегментів. Надалі можна зобразити зміну спектру як функції від часу.

$$STFT(m,\omega) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-i\omega n}, \qquad (2.3)$$

$w[n]$ – деяка віконна (вагова) функція.

Саме цей тип перетворення буде використано у даній роботі для отримання спектру сигналів тренувальної вибірки. Але, як правило, спектр має достатньо високу частоту дискретизації, тому в результаті застосування перетворення отримується значна кількість відліків спектру, що не дозволяє використовувати їх безпосередньо в якості параметрів опису сигналів. Можна виконати процедуру згладжування або використовувати зважені поканальні значення потужності спектру.

2.2 Мел-частотний кепстр

Мел – психофізична одиниця висоти звуку (тону), яка використовується в музичній акустиці, що базується на сприйнятті звуків людськими органами слуху [23, 24]. Мел-шкала співставляє частоту чистого тону, яка сприймається людиною, з фактично виміряною частотою. Річ у тім, що люди набагато краще розрізняють невеликі зміни висоти звуку на низьких частотах, аніж на високих. Тому використання цієї шкали для зображення значень функцій робить їх більш близькими до того, що чують люди. Формула для перетворення значення частоти звука f (Гц) в значення висоти m (мел):

$$m = 1127ln(1 + f/700). \qquad (2.4)$$

Кількісна оцінка звуку за висотою базується на статистичній обробці великої кількості даних про суб’єктивне сприйняття людиною висоти звукових тонів. Результати досліджень засвідчують, що висота звуку пов’язана в основному з частотою коливань, але залежність нелінійна, особливо на низьких частотах (Рис. 2.1).

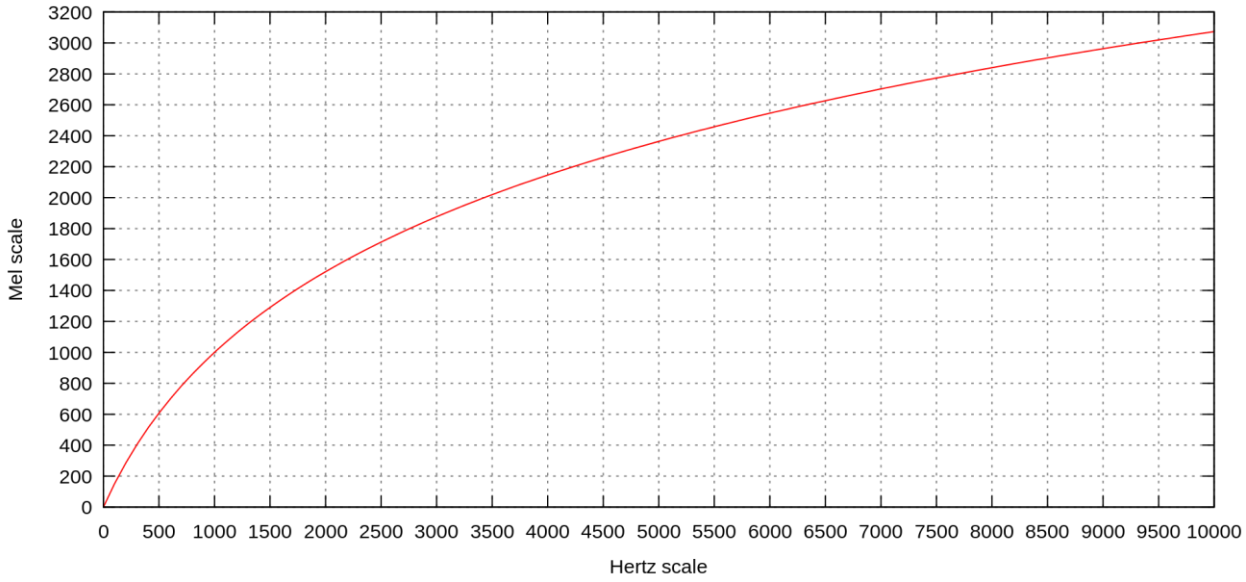

Рисунок 2.1 – Графік залежності висоти звуку від частоти

MFC (Mel-frequency cepstrum) – мел-частотний кепстр [25] – представлення спектра потужності (амплітудного спектра) звуку, засноване на лінійному косинусному перетворенні логарифма спектра потужності на нелінійній меловій шкалі частот. Дане частотне перетворення може забезпечити краще представлення звуку при його стисненні, і в результаті ми отримуємо набагато меншу кількість даних у порівнянні з часовим представленням сигналу чи спектрограмою, яка досить якісно їх замінює.

Для отримання коефіцієнтів, які в сукупності утворюють мел-частотний кепстр (Mel-frequency cepstral coefficients) [26], застосовується такий алгоритм:

1. Сигнал розбивається на короткі кадри і для кожного обчислюється спектр потужності, тобто використовується STFT.

$$S_i(k) = \sum_{n=1}^{N} f_i(n) w(n) e^{-j2\pi kn/N}, 1 \le k \le K, \qquad (2.5)$$

$S_i(k)$ – спектр i-ого кадру;

$f_i(n)$ – i-*ий* кадр;

$w(n)$ – віконна функція;

N – кількість точок в кадрі та віконній функції;

K – довжина перетворення Фур'є.

Відповідно спектр потужності:

$$P_i(k) = |S_i(k)|^2. \quad (2.6)$$

2. Застосувати фільтри, що складаються з трикутних віконних функцій, до спектрограми потужності, щоб розмістити її на мел-шкалі. Знайти суму отриманих значень по кожному фільтру – разом вони утворюють енергетичні коефіцієнти.

Набір із M фільтрів $G_m(k)$:

$$G_m(k) = \begin{cases} 0, & k < f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)}, & f(m-1) \le k \le f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)}, & f(m) \le k \le f(m+1) \\ 0, & k > f(m+1) \end{cases} \quad (2.7)$$

f()- набір частот, розташованих на мел-шкалі.

Енергетичні коефіцієнти для кожного кадру:

$$EC_i = \sum_{k=1}^{K} P_i(k) G_m(k). \quad (2.8)$$

3. Прологарифмувати енергетичні коефіцієнти і застосувати дискретне косинусне перетворення, результатом якого і будуть мел-частотні кепстральні коефіцієнти.

$$MFCC_i = DCT(\log(EC_i)). \quad (2.9)$$

2.3 Алгоритм Гріффіна-Ліма

У загальному випадку перетворення Фур'є є комплекснозначним, а побудована амплітудна спектрограма (і згенерована нейронною мережею) складається з дійсних чисел, тому інформація про фази сигналу втрачається. Незважаючи на те, що перетворення Фур'є є оборотним, внаслідок модифікацій спектрограми і втрати інформації, відновити сигнал до початкового стану не вдасться. Для того, щоб наближено відновити фази, можна застосувати ітеративний алгоритм Гріффіна-Ліма [27] – спеціальний

алгоритм для оцінки сигналу на основі його модифікованого амплітудного спектру (modified STFT magnitude). Суть методу полягає у мінімізації середньоквадратичної похибки між даним модифікованим спектром і спектром оціненого сигналу (метод найменших квадратів).

Висновки до розділу 2

Існує значна кількість напрямків обробки сигналів, що залежать від їхньої природи, і включають в себе, наприклад, відновлення і розмежування інформаційних потоків, прибирання шуму, стиснення, фільтрація, підсилення сигналів тощо. Сигнали можуть бути аналоговими або цифровими та мати різні джерела. Для виділення корисної інформації з сигналу та отримання ознак варто застосовувати спектральний аналіз. Найбільш поширеним способом такого аналізу є використання перетворення Фур’є для отримання спектра частот сигналу чи віконного перетворення Фур’є для спектрограми. Окрім цього, особливий вид представлення спектру потужності, який називається мел-частотний кепстр, досить точно і при цьому компактно характеризує сигнал з точки зору його енергетичного розподілу.

3 ПОБУДОВА АРХІТЕКТУРИ НЕЙРОННОЇ МЕРЕЖІ ТА ПРОВЕДЕННЯ ЕКСПЕРИМЕНТІВ

Глибинні нейронні мережі є дуже потужними моделями машинного навчання, які досягають неперевершених результатів у вирішенні складних задач. Більш того, великі DNN можуть бути натреновані з використанням методу зворотного поширення помилки, якщо маркована тренувальна вибірка має достатньо інформації, щоб визначити параметри мережі. Незважаючи на свою гнучкість і ефективність, DNN можуть застосовуватись лише для вирішення завдань, у яких данні входу та виходу можуть бути закодовані векторами фіксованої розмірності. Це досить суттєве обмеження, адже багато важливих задач виражаються послідовностями змінної розмірності, і довжина яких невідома апріорі. Зокрема, і в поставленій задачі даної роботи вхідними даними є текстові послідовності, які потрібно відобразити на вихідні – мовні сигнали.

Для вирішення цієї проблеми використовуються рекурентні нейронні мережі (RNNs), а частіш за все їхній різновид LSTM (мережа довгострокової та короткострокової пам’яті) [28], який додатково має властивість вчитись на даних з великими часовими залежностями. Але і рекурентні нейронні мережі мають досить суттєвий недолік – вони потребують великих обсягів тренувальних даних і обчислювальних ресурсів, що призводить до значних затрат часу на тренування таких мереж у порівнянні з іншими видами мереж. Більшість дослідників в області досліджень нейронних мереж використовують графічні процесори (GPU) для тренування мереж завдяки перевагам у швидкості та ефективності.

З огляду на всі ці фактори, для побудови моделі в якості базового блоку було обрано згорткові нейронні мережі, які дуже швидко обчислюються на GPU завдяки їх паралелізації.

3.1 Загальна структура та схема нейронної мережі

У задачі синтезу мовлення генеративні моделі можуть обумовлюватися заданим текстом та ідентифікатором диктора. У той час як текст містить лінгвістичну інформацію і контролює зміст мовлення, що генерується, ідентифікатор диктора визначає такі характеристики як висота тону, швидкість мови і акцент. Одним із підходів для створення мультиспікерної мережі є сумісне тренування генеративної моделі та моделі, що конструює представлення диктора, на трійках тексту, аудіо та ідентифікатора. Ідея полягає у кодуванні залежної від диктора інформації у вигляді вектора невеликої розмірності, і в одночасному розподілі більшості параметрів моделі між усіма дикторами. Одним з обмежень такого методу є той факт, що вони можуть синтезувати мовлення лише тих дикторів, які містились у тренувальній вибірці. Такі генеративні моделі можуть бути якісно натреновані за наявності великого обсягу аудіо даних, проте бажаною б була можливість вивчати (копіювати) голос нової людини з невеликої кількості зразків її мовлення і витрачати на це меншу кількість часу. Тому вибір архітектури нейронної мережі був обумовлений цим фактором.

В основі структури нейронної мережі лежить варіаційний автокодувальник [29] у поєднанні з механізмом уваги [30] (Рис. 3.1). Автокодувальник складається з двох кодерів Enc_1 та Enc_2, які кодують вхідний текст і мел-частотний кепстр відповідно, та декодера Dec_1, який розкодовує отриманий результат. Механізм уваги використаний для того, щоб нейронна мережа могла вивчити довгострокові часові залежності між послідовностями тексту і мел-частотного кепстру без використання рекурентних мереж. Функція уваги може бути описана як відображення матриці запитів Q і ключів K зі значеннями V у матрицю вагів A, яка і буде представляти собою «карту уваги» між послідовностями:

$$A = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{3.1}$$

d_k - розмірність вектору ключів,

$softmax(x)$ - нормована експоненційна функція.

З цієї причини, кодер Enc_1 кодує текст $T = [t_1, t_2, \dots, t_n]$, що складається з n символів, у вигляді матриці ключів $K \in R^{n \times d}$ та значень $V \in R^{n \times d}$, і додатково використовується кодер Enc_2 мел-частотного кепстру $M \in R^{m \times f}$ (f - кількість фільтрів, використаних для отримання кепстру), який кодує його у вигляді матриці запитів $Q \in R^{m \times d}$. Отримана матриця уваги $A \in R^{m \times d}$ конкатенується із матрицею Q і результат декодується в новий кепстр $M' \in R^{m \times f}$ так, щоб він був максимально близький до того, який подавався на вхід. Кодери і декодер повністю складаються зі згорткових шарів $Conv_i$ та шарів нормалізації $LayerNorm_i$, які розташовані відразу за ними.

Після цього новий кепстр подається в мережу MelSpec, яка перетворює його на амплітудну спектрограму $S' \in R^{m \times F}$ (розширює частотний діапазон). Ця мережа також є глибинною згортковою мережею з шарами нормалізації. До результату роботи мережі – згенерованої амплітудної спектрограми – застосовується алгоритм Гріффіна-Ліма для отримання мовного сигналу.

Детальний опис структури нейронної мережі наведений у Додатку А.

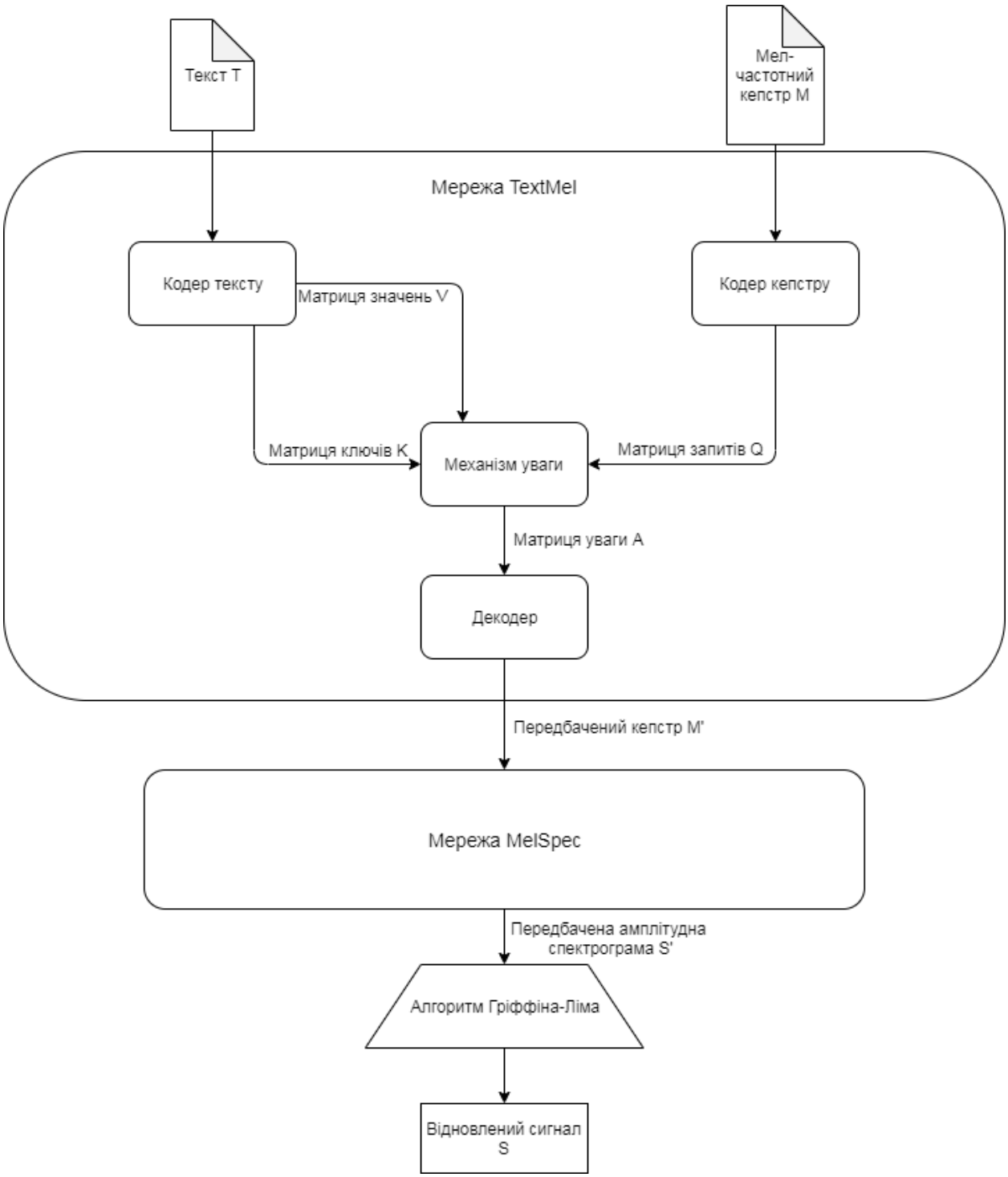

Рисунок 3.1 – Схема нейронної мережі

Окрім цього, в модель нейронної мережі також входять функції активації для того, щоб ввести нелінійність у роботу мережі, але при цьому не змінювати результат її роботи в значних розмірах.

Функція активації – функція, що обчислює вихідний сигнал штучного нейрона. Як аргумент вона приймає сигнал X, одержаний на виході суматора шару. До функцій цієї категорії належать:

- порогова функція (двійковий крок): $f(x) = \begin{cases} 0, x < 0 \\ 1, x \geq 0 \end{cases}$;
- тангенціальна функція: $f(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$;
- softsign: $f(x) = \frac{x}{1+|x|}$;
- арктангенс: $f(x) = tan^{-1}(x)$;
- ELU (exponential linear unit): $f(\alpha, x) = \begin{cases} \alpha(e^x - 1), x < 0 \\ x, x \geq 0 \end{cases}$;
- Ґауссіан: $f(x) = e^{-x^2}$.

В якості функцій активації в побудованій моделі було використано сигмоїду та ReLU.

Сигмоїда – це неперервна монотонна нелінійна функція, що застосовується для згладжування значень заданої величини. Визначається формулою:

$$S(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1}. \quad (3.2)$$

Дана функція відображає вхідні числа на інтервал (0, 1). Однією з причин, через яку сигмоїда використовується в нейронних мережах є простота вираження похідної через саму функцію:

$$S'(x) = S(x) \cdot \big(1 - S(x)\big). \quad (3.3)$$

Це дозволило істотно скоротити обчислювальну складність методу зворотного поширення помилки для функції.

ReLU (rectified linear unit) – передавальна функція, що визначається таким чином:

$$f(x) = \max(0, x). \quad (3.4)$$

Ця функція є найбільш популярною функцією активації для глибинних нейронних мереж і на додачу має біологічне підґрунтя та математичне

обґрунтування [31]. Висока частота використання обумовлена її низкою переваг, серед яких:

- біологічна правдоподібність – функція одностороння на відміну від гіперболічного тангенса;
- висока швидкість обчислення (містить тільки операцію порівняння, у видозмінених формах – додавання та множення);
- прискорення збіжності градієнтного спуску;
- зменшення ймовірності виникнення проблеми затухання градієнту у порівнянні з гіперболічним тангенсом та логістичною функцією;
- розріджена активація (у випадково ініціалізованій мережі лише приблизно половина прихованих нейронів активуються);
- інваріантність відносно масштабування: $\max(0, ax) = a \cdot max(0, x)$.

Проте ReLU має і деякі недоліки:

- нерегулярна в точці 0;
- наявні «мертві» зони – за великих значень градієнта нейрони можуть бути переведені у стан, у якому вони стануть неактивними для всіх точок вхідних даних. Причиною виникнення такого явища зазвичай є занадто високий коефіцієнт швидкості навчання.

Слід також зазначити, що однією з проблем глибинних нейронних мереж є перенавчання – модель чудово справляється тільки з даними, що надходять із тренувальної вибірки, просто запам’ятовує їх, замість того, щоб навчатись, і втрачає здатність узагальнювати. Оскільки побудована мережа також є глибинною, то існує висока ймовірність виникнення такої загрози і на це потрібно звернути увагу.

Для вирішення цієї проблеми застосовується регуляризація. Існує чимало способів регуляризації (L1 та L2 регуляризація, max norm constraints та ін.), проте один із них перевершує решту – це метод розріджування (dropout). Принцип роботи цього методу проілюстровано на Рис. 3.2. Його

ідея полягає у виключенні з мережі нейронів із ймовірністю p, таким чином ймовірність того, що нейрон залишиться в мережі становитиме $q = 1 - p$. Під виключенням нейрона мається на увазі, що він повертатиме значення 0 за будь-яких вхідних даних чи параметрів. Виключені нейрони не роблять внеску в процес навчання на жодному з етапів алгоритму зворотного поширення помилки, тому виключення хоча б одного нейрона еквівалентно навчанню нової нейронної мережі.

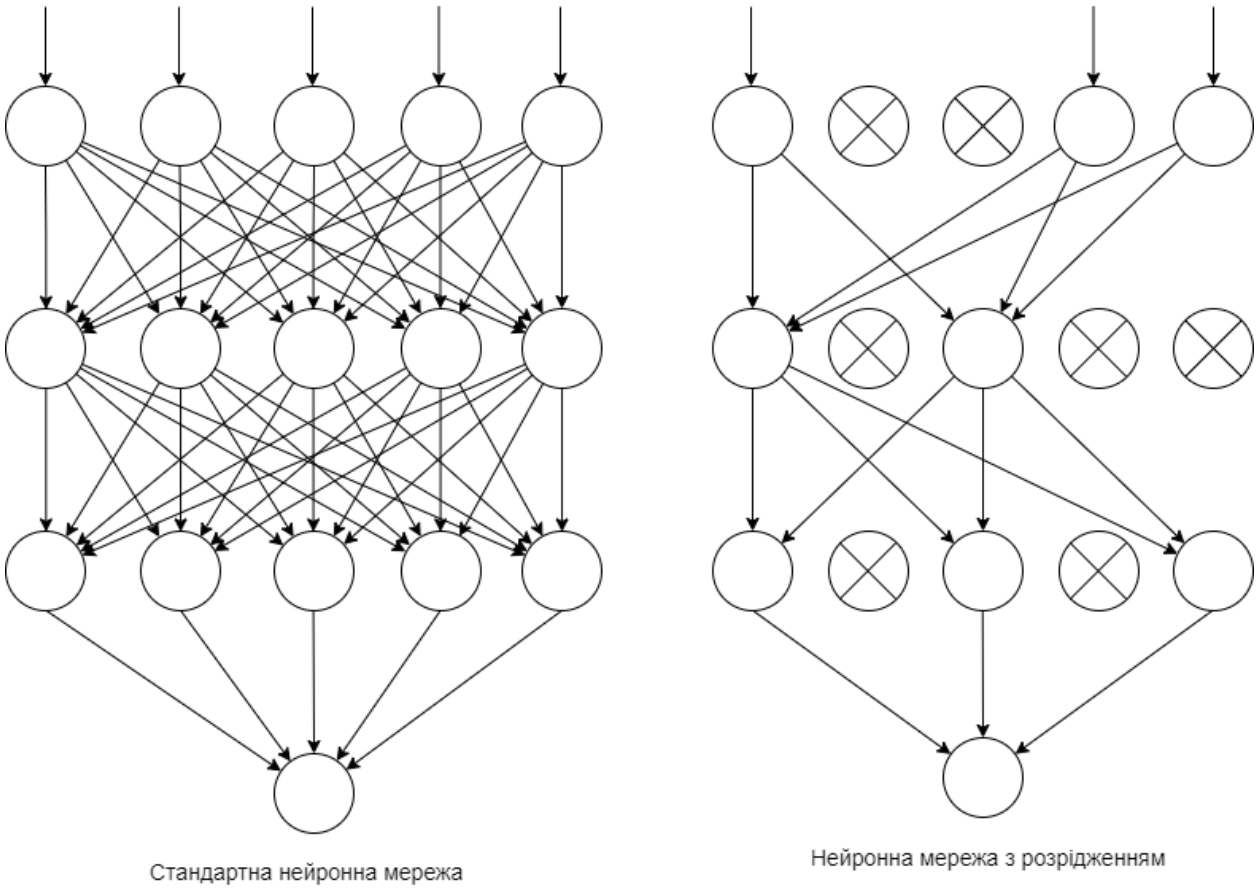

Рисунок 3.2 – Регуляризація dropout

3.2 Організація експериментів

3.2.1 Збір та підготовка тренувальних даних

База даних, яка використовується для тренування нейронної мережі, є дуже важливою складовою і безпосередньо впливає на ефективність та якість результатів її передбачень, тому даному етапу слід також приділити увагу.

Для тренування нейронної мережі було обрано базу даних The LJ Speech [32], в якості тестової вибірки — VCTK Corpus [33]. The LJ Speech складається 13 тисяч аудіо записів одного мовця, які сумарно становлять близько 24 годин аудіо. До кожного запису наявний відповідний текст. VCTK Corpus містить записи 109 мовців із різним акцентом, приблизно по 400 аудіо файлів для кожного з наявним текстом. Кожен набір містить якісно записані і різноманітні дані з точки зору максимального використання контекстного та фонетичного покриття.

В якості векторів ознак для нейронної мережі буде використано представлення тексту, мел-частотний кепстр, а також амплітудну спектрограму.

Для прикладу, оберемо фразу «But for the beauty of the earlier work they might have seemed tolerable» та відповідний аудіо запис (Рис. 3.3).

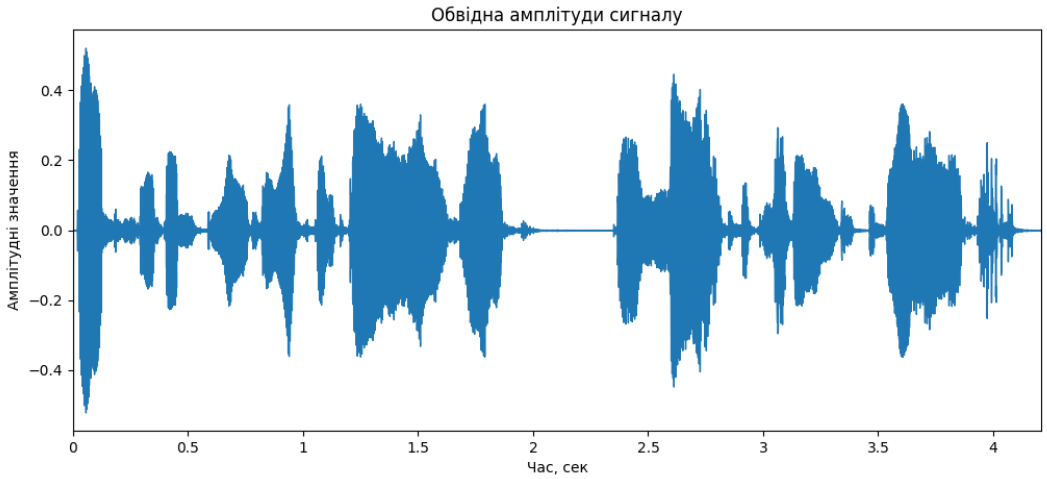

Рисунок 3.3 – Обвідна амплітуди сигналу

Для представлення тексту буде використано простий словник, у якому кожній літері англійського алфавіту ставиться у відповідність ціле число: {‘a’: 1, ‘b’: 2, ...}, а потім в процесі навчання кожній літері підбиратиметься певний вектор ознак з оптимальними значеннями.

Для отримання амплітудної спектрограми аудіо файлу використаємо STFT з вікном Хеннінга, поділом на кадри по 50 мс, зсувом 12 мс, роздільною здатністю 10 Гц на смугу (довжина перетворення F = 2048 при

частоті дискретизації 22050 Гц) з подальшим знаходженням абсолютних значень коефіцієнтів (Рис. 3.4) і нормалізацією.

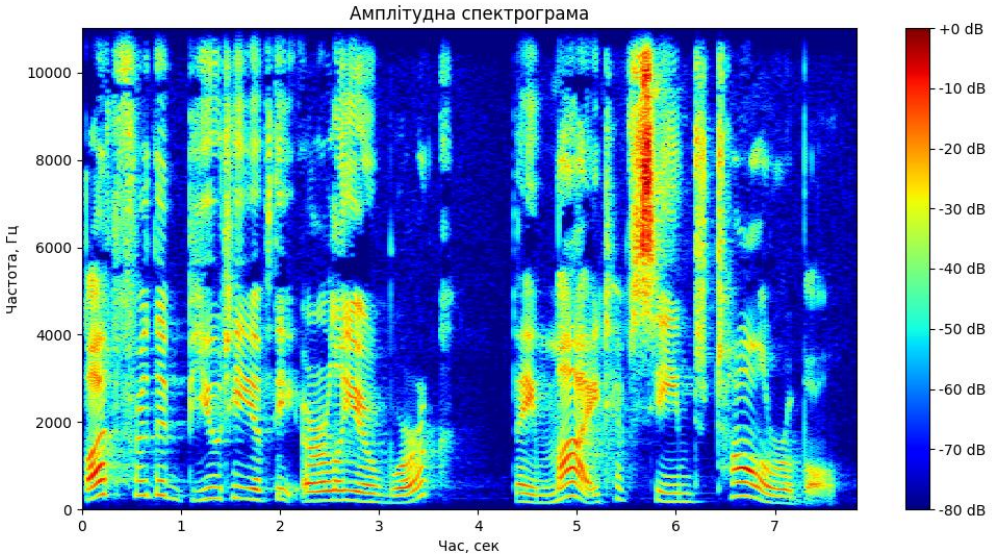

Рисунок 3.4 – Амплітудна спектрограма

Для отримання мел-частотного кепстру (Рис. 3.5) використано наведений раніше (див. розділ 2.2) алгоритм з f = 80 фільтрами.

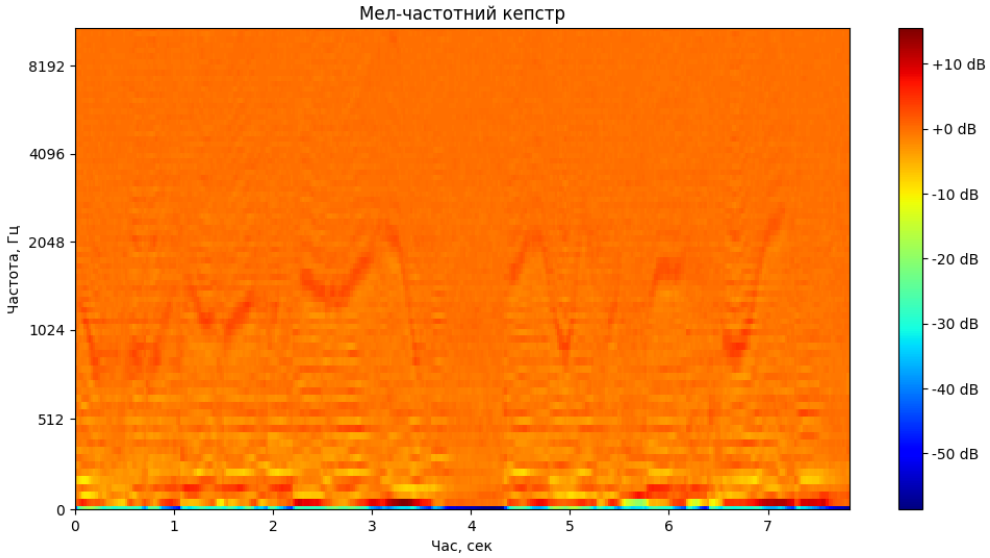

Рисунок 3.5 – Мел-частотний кепстр

На практиці для реалізації програми, що обчислює усі вектори ознак, використано бібліотеки Numpy та LibROSA мови програмування Python3.

3.2.2 Вибір та оптимізація функції втрат нейронної мережі

Для тренування нейронної мережі необхідно обрати функцію, яка буде характеризувати втрати при неправильному прийнятті рішень на основі спостережуваних даних. Шляхом оптимізації такої функції будуть оновлюватись параметри мережі і вона навчатиметься приймати правильні рішення, іншими словами виконувати поставлену задачу.

В якості функції втрат було обрано суму функції найменших модулів (Least absolute deviations або ж L_1 norm) H_1 та функції бінарної перехресної ентропії (Binary Cross Entropy) H_2.

$$H_1(x, y) = \sum_{i=1}^{n} |y_i - f(x_i)|. \tag{3.5}$$

$$H_2(x, y) = -\sum_{i=1}^{n} w_i(y_i \log(x_i) + (1 - y_i) \log(1 - x_i)), \tag{3.6}$$

x – вихідний вектор ознак для нейронної мережі;

y – справжній вектор ознак із тренувальної вибірки.

Така функція втрат зазвичай використовується для вимірювання похибки реконструкції, наприклад в автокодувальниках.

Від вибору алгоритму оптимізації для моделі залежатиме кількість часу необхідного для отримання хорошого результату, і різниця може варіюватись суттєво - від годин до днів.

Adam [34] (adaptive moment estimation) – алгоритм оптимізації, який є розширенням процедури класичного стохастичного градієнтного спуску (метод зворотного поширення помилки), і може бути використаний замість нього для оновлення вагів мережі.

Був обраний саме цей оптимізатор завдяки його численним перевагам:

- простота реалізації;
- обчислювальна ефективність;
- невеликі потреби в пам’яті;

- інваріантність до діагонального масштабування градієнтів;
- чудово підходить для задач, які є великими з точки зору обсягу даних та кількості параметрів;
- підходить для нестаціонарних цілей оптимізації та задач з дуже зашумленими чи розрідженими градієнтами;
- гіперпараметри мають інтуїтивно зрозумілу інтерпретацію та зазвичай потребують незначних налаштувань (tuning).

Стохастичний градієнтний спуск містить лише один коефіцієнт швидкості навчання для оновлення усіх параметрів мережі і цей коефіцієнт не змінюється упродовж тренування. На відміну від нього, Adam використовує для кожного значення ваги свій коефіцієнт і окремо його адаптує по мірі просування в навчанні, базуючись на оцінках першого і другого моментів градієнтів. Цей алгоритм не тільки поєднує переваги інших двох оптимізаторів водночас – AdaGrad та RMSProp – але й вдосконалює їх. Так, наприклад, замість адаптації коефіцієнтів швидкості навчання лише на основі середнього значення першого моменту (математичне очікування) як в RMSProp, Adam також використовує середнє значення другого моменту градієнта (дисперсія). Зокрема, алгоритм обчислює експоненціальне рухоме середнє значення градієнта і його квадрата, а параметри β_1 та β_2 контролюють швидкість затухання цих рухомих значень.

Окрім цього, вибір був зроблений на користь цього алгоритму, оскільки в статті демонструються результати емпіричних дослідів, пов’язаних з оцінкою запропонованого методу при використанні в глибинних згорткових нейронних мережах з великими датасетами (Рис. 3.6) [34].

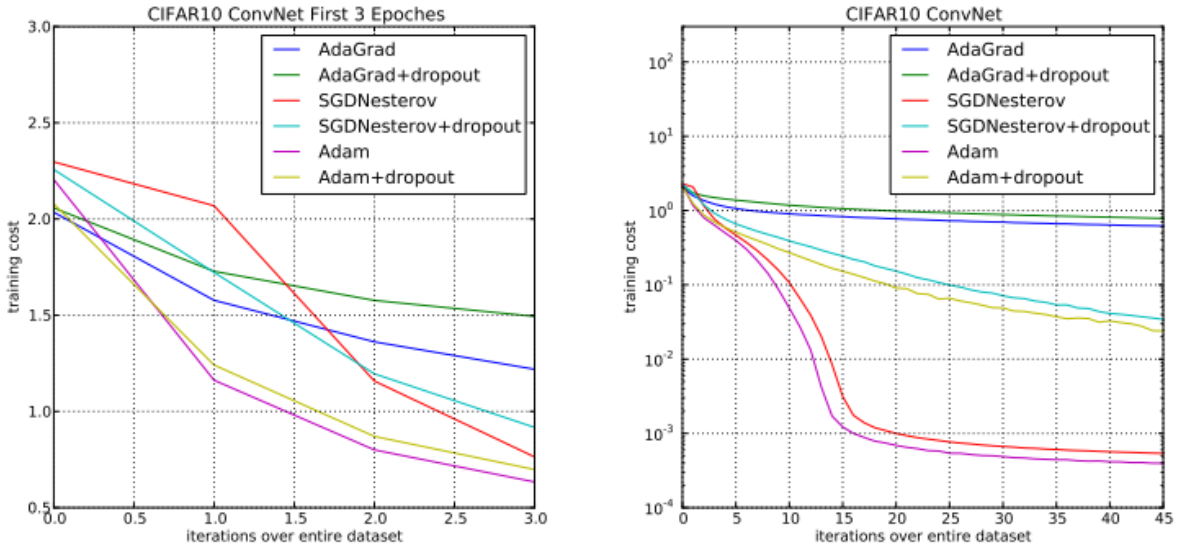

Рисунок 3.6 – Convolutional neural networks training cost. (left) Training cost for the first three epochs. (right) Training cost over 45 epochs. CIFAR-10 with c64-c64-c128-1000 architecture.

Наведена архітектура мережі має три почергово змінних етапи зі згортковими фільтрами 5x5 та max pooling фільтрами 3x3, за якими розташовано повнозв'язний шар, який складається з 1000 прихованих випрямлених лінійних функцій (ReLU).

Для побудованої моделі був використаний Adam зі стандартними параметрами:

$$\alpha = 0.001,$$
$$\beta_1 = 0.9,$$
$$\beta_2 = 0.999,$$
$$\varepsilon = 10^{-8}.$$

Тренування основної моделі на великій базі даних (The LJ Speech) тривало близько 34 години з використанням одного графічного процесора.

Решта експериментів проводилась з метою адаптації натренованої мережі під новий голос людини на базі датасету VCTK Corpus. У ході тренування й адаптації коригувалися значення гіперпараметрів (початковий коефіцієнт швидкості тренування, параметри шарів мережі, значення

ймовірності розрідження тощо). Прогрес оптимізації функції втрат зображено на Рис. 3.7.

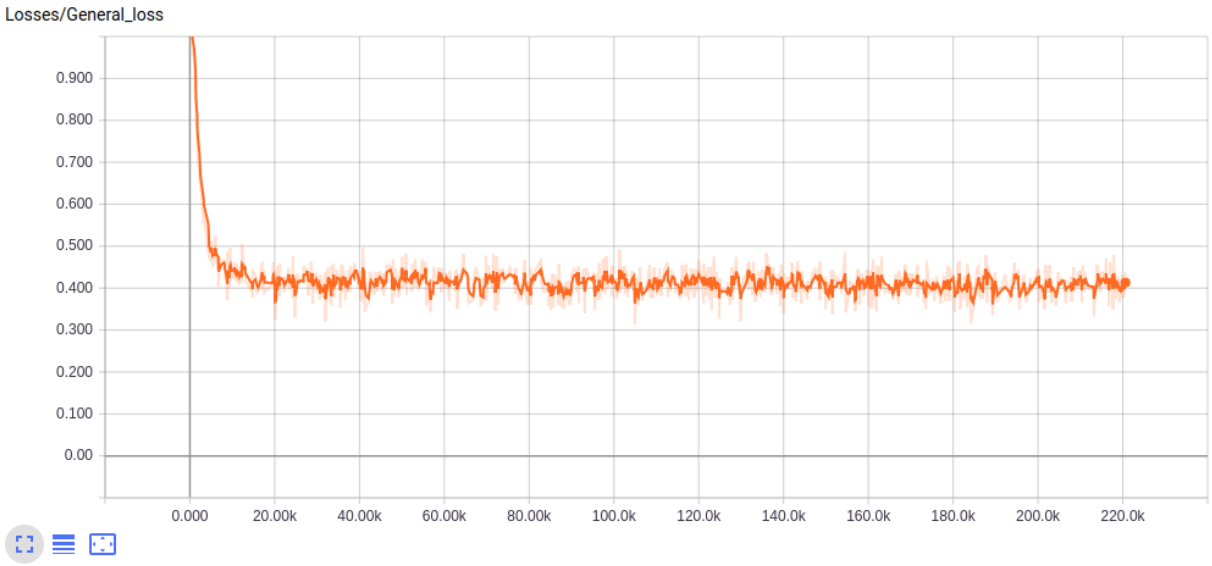

Рисунок 3.7 – Оптимізація функції втрат мережі

3.3 Результати генерації цільового голосу людини та їх оцінювання

У результаті проведення численних експериментів було виявлено, що натренованій нейронній мережі достатньо близько 15 хвилин записів мовлення нового диктора і 35-45 хвилин для адаптації.

Для адаптації під кожного нового диктора з бази даних VCTK було використано лише частину наявних аудіо файлів (близько 25-30%). Решту файлів нейронна мережа не використовувала, тому вони були взяті за сигнали для проведення оцінювання.

Недолік роботи нейронної мережі, який був виявлений у ході проведення експериментів, полягає в тому, що іноді проміжки згенерованого сигналу звучать нечітко з точки зору вимови, з невдалими окремими звуками. За отриманою статистикою трапляються такі випадки в середньому з імовірністю 8%. Це явище можна пояснити неточною роботою механізму уваги.

Методи оцінювання якості синтезу поділяються перш за все на два основні типи: суб’єктивні (MOS-оцінки) та об’єктивні (інструментальні).

До першого типу належать різноманітні тести та опитування, які заповнюються експертами - спеціалістами чи звичайними слухачами. При створенні та проведенні таких оцінювань зазвичай використовують рекомендації, що містяться в стандартах ITU-T або ITU-R. Тут використовується MOS-оцінка за п’ятибальною шкалою в різних категоріях: загальне враження, природність, слухові зусилля, рівень зрозумілості змісту повідомлень, темп, розбірливість тощо.

Для прискорення процесу оцінювання створюються різноманітні об’єктивні методи оцінки якості синтезу. Такі методи ґрунтуються як на автоматичному порівнянні синтезованого мовлення зі справжнім для одного і того ж диктора за допомогою різних мір близькості, так і на побудові незалежних від мовця моделей природного мовлення і різних методах оцінки того, наскільки синтезована мова близька до них.

3.3.1 Суб’єктивне оцінювання

1. Тест на виявлення синтезованого мовлення

Метою проведення даного тесту є отримання оцінки природності звучання та схожості синтезованого нейронною мережею мовлення до справжнього мовлення диктора. Для кожного з 10 дикторів було створено набір зразків аудіо, який складався з 25 справжніх записів із бази даних і 25 синтезованих. Синтезовані зразки були отримані з тексту, який не входив до тренувальної вибірки, на якій навчалась нейронна мережа. Аудіо файли були розташовані у випадковому порядку, і задачею кожного слухача було визначити, який це запис – справжній чи згенерований мережею. Оцінка проводилась кожним слухачем окремо від інших.

У таблиці 3.1 наведено результати тесту – кількість записів для кожного диктора, які були сприйняті кожним слухачем як справжні, але насправді були синтезованими.

Таблиця 3.1 – Результати тесту на виявлення синтезованого мовлення

Слухач \ Диктор	p225	p226	p227	p228	p229
1	21	19	21	20	21
2	18	20	23	21	20
3	20	20	21	19	23
4	22	21	20	22	20
5	19	18	21	20	19
Середнє значення, %	80.0	78.4	84.8	81.6	82.4

Слухач \ Диктор	p230	p231	p232	p233	Rus
1	20	20	19	21	22
2	18	22	21	20	23
3	20	21	19	22	21
4	19	23	22	19	22
5	19	20	20	20	20
Середнє значення, %	76.8	84.8	80.8	81.6	86.4

Кількість синтезованих зразків серед усіх дикторів, які не були виявлені коливається в діапазоні 76.8% до 86.4 %.

У середньому для всіх дикторів ця кількість становить 82.76 %

2. MOS-оцінка якості синтезованого мовлення

Найбільш широко використовуваним методом суб’єктивного оцінювання якості є метод оцінки категорії, при якому слухачі ставлять

оцінку якості тестового сигналу за п’ятибальною числовою шкалою (таблиця 3.2).

Таблиця 3.2 – Шкала оцінювання MOS

Бали	Якість мовлення	Рівень спотворень
5	Excellent (відмінно)	Imperceptible (непомітний)
4	Good (добре)	Just perceptible, but not annoying (ледь помітний, але не дратуючий)
3	Fair (задовільно)	Perceptible and slightly annoying (помітний і трохи дратуючий)
2	Poor (незадовільно)	Annoying, but not objectionable (дратуючий, але прийнятний)
1	Bad (погано)	Very annoying and objectionable (дуже дратуючий і неприйнятний)

Такий метод рекомендований Комітетом Інституту інженерів з електротехніки та електроніки з питань суб’єктивних методів вимірювань (IEEE Subcommittee on Subjective Measurements), а також Міжнародним союзом електрозв’язку (ITU). Виміряна якість тестового сигналу отримується шляхом знаходження середнього арифметичного оцінок, отриманих від усіх слухачів. Цей середній бал і називається MOS (Mean Opinion Score).

Детальний опис та рекомендації щодо процедури проведення тесту міститься в стандарті ITU-R BS.562-3, зокрема він включає в себе інструкції щодо:

- проведення тестової процедури та її тривалості. Зразки мовних сигналів (оригінальні та модифіковані) повинні бути представлені слухачам у випадковому порядку, тривалість сесії не повинна перевищувати 20 хвилин без перерв;

- вибору пристрою відтворення звуку. Рекомендується надавати перевагу навушникам, аніж гучномовцям, оскільки відтворення звуку в

навушниках не залежить від геометричних та акустичних властивостей приміщення, у якому проводиться випробування.

У даному опитуванні кожному слухачеві надавалось 20 зразків синтезованого мовлення кожного з 10 дикторів. Аналогічно синтезовані зразки були отримані з тексту, який не входив до тренувальної вибірки.

Таблиця 3.3 – Результати оцінювання за шкалою MOS

Диктор / Слухач	p225	p226	p227	p228	p229
1	3.45	3.75	3.9	4.1	3.85
2	3.9	3.8	4.15	4.05	3.9
3	3.85	3.95	3.8	3.9	3.75
4	4.0	3.7	3.85	4.15	3.8
5	3.9	3.8	3.95	4.0	3.75
MOS	3.82	3.8	3.93	4.04	3.81

Диктор / Слухач	p230	p231	p232	p233	rus
1	3.6	4.05	3.9	3.85	4.1
2	3.8	4.2	3.65	3.7	3.95
3	3.65	4.1	3.85	4.05	4.25
4	3.7	3.9	3.8	3.95	4.0
5	3.75	4.1	4.0	3.8	4.2
MOS	3.7	4.07	3.84	3.87	4.1

Значення MOS серед усіх дикторів коливається в діапазоні від 3.7 до 4.1. Для порівняння мережа Deep Voice 3 [21] має MOS в межах 3.78 - 4.12.

3.3.2 Об’єктивне оцінювання

Для об’єктивного оцінювання роботи нейронної мережі було обрано ймовірнісну статистичну модель [35], яка використовує PLDA - ймовірнісний лінійний дискримінативний аналіз. Ця модель може використовуватися як

для ідентифікації так і для верифікації мовців. Її тренування відбувалось на великій за обсягом базі даних - більше 1200 дикторів і десятки тисяч зразків мовлення. Як зазначають автори роботи [35], в задачі верифікації похибка склала 7.8%, тому цю модель можна вважати досить якісною для вирішення цієї задачі.

Для проведення тестування в якості еталонного зразку випадковим чином обирався аудіо файл (5 спроб) справжнього голосу мовця з бази даних (10 різних дикторів), а верифікація проводилась зі 100 зразками мовлення синтезованого нейронною мережею для кожного диктора відповідно. В таблиці 3.4 наведено кількість синтезованих зразків, які були успішно верифіковані моделлю як голос справжнього диктора.

Таблиця 3.4 – Результати об’єктивного оцінювання

Диктор / Зразок	p225	p226	p227	p228	p229
1	39	48	45	54	49
2	53	54	55	58	52
3	46	47	53	65	48
4	58	45	57	63	52
5	56	52	56	57	55
Середнє значення, %	50.4	49.2	53.2	59.4	51.2

Диктор / Зразок	p230	p231	p232	p233	rus
1	43	45	57	38	56
2	38	43	53	44	62
3	45	43	60	47	58
4	47	46	59	48	64
5	45	50	64	45	63
Середнє значення, %	43.6	45.4	58.6	44.4	60.6

У середньому кількість неправильних випадків верифікації становить від 43.6% до 60.6%. Такий розмах значень можна пояснити двома факторами: по-перше точність верифікації підвищується, якщо довжина еталонного зразка більша, оскільки в такому випадку модель отримує більше інформації про справжній голос (а в якості еталона обирались зразки різної довжини), хоча іноді додаткові дані не надають нової інформації; по-друге якість синтезованих зразків теж залежить від точності роботи нейронної мережі і може варіюватись в залежності від конкретного випадку (тексту).

Висновки до розділу 3

З’ясовано, що рекурентні нейронні мережі можуть бути ефективно замінені згортковими мережами у поєднанні з механізмом уваги для вирішення задач, які вимагають роботи з послідовностями змінної довжини. Такий підхід може перевершити навіть можливості мереж LSTM, які вважаються одними з найкращих у вирішенні такої задачі, запам’ятовувати довгі часові залежності.

Побудована архітектура нейронної мережі дозволяє зменшити обсяг необхідних обчислювальних ресурсів та даних для успішного навчання і потребує значно менше часу тренування.

Результати проведених оцінювань дають підстави вважати, що синтезоване нейронною мережею мовлення має достатньо високий рівень якості, природності звучання і схожості з оригінальним голосом диктора як із суб’єктивної, так і з об’єктивної точок зору.

ВИСНОВКИ

Розглянуто наукову літературу за темою роботи, в результаті чого виявлено основні підходи до моделювання процесів синтезу та модифікації мовних сигналів, розпізнавання людини за особливостями її голосу. Досліджено переваги та недоліки цих підходів, а також наведено конкретні приклади їх реалізацій. Це дало змогу обрати найкращий і найсучасніший з них для дослідження.

Проаналізовано найпоширеніші способи обробки мовних сигналів з метою використання їх для виділення корисної інформації з сигналів. Встановлено, що доцільним є застосування віконного перетворення Фур'є і дискретного косинусного перетворення для отримання представлення сигналів (спектрограма і мел-частотний кепстр), яке може бути використано в якості характерних ознак для тренування нейронної мережі.

З'ясовано, що рекурентні нейронні мережі можуть бути ефективно замінені згортковими мережами у поєднанні з механізмом уваги для вирішення задач, які вимагають роботи з послідовностями змінної довжини і запам'ятовування залежностей між ними. Такий підхід може перевершити навіть можливості мереж LSTM запам'ятовувати довгострокові часові залежності.

Подано детальний опис процесу побудови архітектури нейронної мережі, яка дозволяє зменшити обсяг необхідних обчислювальних ресурсів та даних для успішного навчання і потребує значно менше часу тренування в порівнянні з рекурентними мережами. Запропоновано основні організаційні кроки проведення експериментальних досліджень для досягнення поставленої в даній роботі мети.

Отже, в ході роботи було виконано всі поставлені завдання, здійснено суб'єктивне та об'єктивне оцінювання якості роботи запропонованої нейронної мережі. За результатами проведених оцінювань і тестувань можна зробити висновок, що запропонована модель нейронної мережі може бути

використана в якості інструменту для перевірки біометричних систем верифікації на здатність виявляти синтезоване мовлення. До того ж вона є ефективною з практичної точки зору, оскільки не потребує великих обсягів даних і затрат часу для навчання.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАНЬ

1. Singh N. Automatic Speaker Recognition: Current Approaches and Progress in Last Six Decades [Text] / Singh N., Agrawal A., R. A. Khan // Global Journal of Enterprise Information System. – 2017. – V. 9, №3. – P. 38–45.

2. Stylianou Y. Continuous probabilistic transform for voice conversion. Speech and Audio Processing [Text] / Stylianou Y., Cappé O., Moulines E. // Proceedings. IEEE Transactions on Acoustics, Speech, and Signal Processing. – Vol. 6, no. 2 (Mar). – 1998. – P. 131–142.

3. Toda T. Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter [Text] / Toda T., Black A. W., Tokuda K. // Proceedings. (ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing. – Vol. 1 (Mar). – 2005. – P. 9–12.

4. Kobayashi K. sprocket : Open-Source Voice Conversion Software / Kobayashi K., Toda T. – Proc. Odyssey. –2018. – P. 203–210.

5. Dempster A.P. Maximum Likelihood from Incomplete Data via the EM Algorithm [Text] / Dempster A. P.; Laird N. M.; Rubin D. B. // Journal of the Royal Statistical Society, Series B (Methodological). – Vol. 39 (1). – 1977. – P 1–38.

6. Godoy E. Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora [Text] / Godoy E., Rosec O., Chonavel T. // Proceedings. IEEE Transactions on Acoustics, Speech, and Signal Processing. – Vol. 20, no. 4 (May). – 2012. – P. 1313–1323.

7. Agiomyrgiannakis Y. Voice morphing that improves TTS quality using an optimal dynamic frequency warping-and-weighting transform [Електронний ресурс] / Y. Agiomyrgiannakis, Z. Roupakia. – Режим доступу: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/44861.pdf

8. Erro D. Voice conversion based on weighted frequency warping [Text] / D. Erro, A. Moreno, A. BonafoLante // Proceedings. IEEE IEEE

Transactions on Acoustics, Speech, and Signal Processing. – Vol. 18, no. 5 (July). – 2010. – P. 922–931.

9. Macon M. W. Applications of sinusoidal modeling to speech and audio signal processing [Електронний ресурс] / M. W. Macon, D. D. J. Blumenthal, D. M. A. Clements, D. R. M. Mersereau. – Режим доступу : http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.51.2790&rep=rep1&type=pdf

10. Kain A. High resolution voice transformation [Text] : PhD dissertation in Computer Science and Engineering; Oregon Graduate Institute / Alexander Kain. – Portland, OR. – 2001. – 129 p.

11. Ye H. Perceptually Weighted Linear Transformation for Voice Conversion [Text] / Ye H. Young S // Eurospeech. – 2003. – P. 2409–2412.

12. Young S. High quality voice morphing [Електронний ресурс] / S. Young. – Режим доступу: https://www.researchgate.net/publication/4087475_High_quality_voice_morphing

13. Zen H. Review: Statistical parametric speech synthesis [Text] / H. Zen, K. Tokuda, A. W. Black // Speech Communication. – Vol. 51, no. 11 (Nov). – 2009. – P. 1039–1064.

14. Yamagishi J. Analysis of speaker adaptation algorithms for hmm-based speech synthesis and a constrained smaplr adaptation algorithm [Text] / J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, J. Isogai // IEEE Transactions on Audio, Speech, and Language Processing. – Vol. 17, no. 1 (Jan). – 2009. – P. 66–83.

15. Hsu C. Voice Conversion from Non-parallel Corpora Using Variational Auto-encoder [Електронний ресурс] / C. Hsu, H. Hwang, Y. Wu, Y. Tsao, H. Wang. – Режим доступу : https://arxiv.org/pdf/1610.04019.pdf – 13.10.2016.

16. Nakashika T. Voice conversion in high-order eigen space using deep belief nets [Text] / T. Nakashika, R. Takashima, T. Takiguchi, Y. Ariki // Interspeech, ISCA. – 2013. – P. 369–372.

17. Karaali O. Speech synthesis with neural networks [Text] / O. Karaali, G. Corrigan, I. A. Gerson // CoRR. – Vol. cs.NE/9811031. – 1998. – P. 45–50.

18. Zen H. Statistical parametric speech synthesis using deep neural networks [Text] / H. Zen, A. Senior, M. Schuster // Proceedings. ICASSP. IEEE, 2013. – P. 7962–7966.

19. Wu Z. A study of speaker adaptation for DNN-based speech synthesis [Text] / Z. Wu, P. Swietojanski, C. Veaux, S. Renals, S. King // International Speech Communication Association. – 2015; Date of Acceptance: 01/06/2015.

20. Mehta A. A Complete Guide to Types of Neural Networks [Електронний ресурс] /A. Mehta. – Режим доступу : https://www.digitalvidya.com/blog/types-of-neural-networks/ - 25.01.2019 р.

21. Arık S. Ö. Neural Voice Cloning with a Few Samples [Електронний ресурс] / S. Ö. Arık. J. Chen, K. Peng, W. Ping, Y. Zhou. – Режим доступу : https://papers.nips.cc/paper/8206-neural-voice-cloning-with-a-few-samples.pdf - 14.02.2018 р.

22. Cooley J. W. An algorithm for the machine calculation of complex Fourier series [Text] / Cooley J. W., Tukey J. W. Math // Comput. – 19 (90). – 1965. – P. 297–301.

23. Гельфанд С.А. Слух: введение в психологическую и физиологическую акустику [Текст] / С. А. Гельфанд. – М.: Медицина, 1984. – 352 с.

24. Пат. RU № 2196508 C2 Российская Федерация, Способ установления единицы высоты тона (мел) и определение её физического смысла [Текст] / Овчинников Е.Л. // Патент RU № 2196508 C2 РФ от 20.04; 2003 по заявке № 99128019 от 31.12.1999.

25. Bogert B. P. The Quefrency Alanysis [sic] of Time Series for Echoes: Cepstrum, Pseudo Autocovariance, Cross-Cepstrum and Saphe Cracking [Text] / B. P. Bogert, M. J. R. Healy, J. W. Tukey // Proceedings of the Symposium on Time Series Analysis (M. Rosenblatt, Ed). – Chapter 15. – New York: Wiley, 1963. – P. 209–243.

26. Davis S. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences [Text] / Davis S., Mermelstein P. // Proceedings. IEEE Transactions on Acoustics, Speech, and Signal Processing. –Vol. 28, No. 4. – 1980. – P. 357– 366.

27. Griffin D. Signal estimation from modified short-time fourier transform [Text] / D. Griffin, J. Lim // Proceedings. IEEE Transactions on Acoustics, Speech, and Signal Processing. – 1984. – P. 236–243.

28. Sutskever I. Sequence to Sequence Learning with Neural Networks [Електронний ресурс] / I. Sutskever, O. Vinyals, Q. V. Le. – Режим доступу : https://arxiv.org/pdf/1409.3215.pdf -14.12.2014 р.

29. Kingma D. P. Auto-Encoding Variational Bayes [Електронний ресурс] / D. P. Kingma, M. Welling. – Режим доступу : https://arxiv.org/abs/1312.6114 - 01. 05. 2014 р.

30. Vaswani A. Attention Is All You Need [Електронний ресурс] / N. Shazeer, N. Parmar, J. Uszkoreit, L Jones, A. N. Gomez, Ł. Kaiser. – Режим доступу : https://arxiv.org/abs/1706.03762 - 06.12.2017 р.

31. Hahnloser R. Permitted and Forbidden Sets in Symmetric Threshold-Linear Networks [Text] / R. Hahnloser, H. S. Seung // NIPS Conference. – 2001. – P. 217–223.

32. The LJ Speech Dataset [Електронний ресурс]. – Режим доступу : https://keithito.com/LJ-Speech-Dataset/ - 08.07.2017 р.

33. CSTR VCTK Corpus. English Multi-speaker Corpus for CSTR Voice Cloning Toolkit [Електронний ресурс]. – Режим доступу : https://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html

34. Kingma D. P Adam: A Method for Stochastic Optimization [Text] / D. P Kingma, J. Ba // Proceedings of the 3rd International Conference on Learning Representations (ICLR). – arXiv preprint arXiv:1412.6980, 2014. P. 1–15.

35. Nagrani A. VoxCeleb: a large-scale speaker identification dataset [Text] / A. Nagrani, J. S. Chung, A. Zisserman // Proceedings of INTERSPEECH, 2017. – P. 2616–2620.

Додаток А

Таблиця А.1 – Детальна структура кодера тексту

Послідовність елементів мережі	Кількість вхідних каналів	Кількість вихідних каналів	Розмір ядра (фільтра)	Зсув (stride)	Розтягнення (dilation)
Emb	1	128	—	—	—
Conv+LayerNorm	128	512	1	1	1
ReLU	—	—	—	—	—
Conv+LayerNorm	512	512	1	1	1
Conv+LayerNorm	512	512	3	1	1
Sigmoid	—	—	—	—	—
Conv+LayerNorm	512	512	3	1	3
Sigmoid	—	—	—	—	—
Conv+LayerNorm	512	512	3	1	9
Sigmoid	—	—	—	—	—
Conv+LayerNorm	512	512	3	1	27
Sigmoid	—	—	—	—	—
Conv+LayerNorm	512	512	3	1	1
Sigmoid	—	—	—	—	—
Conv+LayerNorm	512	512	3	1	1
Sigmoid	—	—	—	—	—
Conv+LayerNorm	512	512	1	1	1
Sigmoid	—	—	—	—	—
Conv+LayerNorm	512	512	1	1	1
Sigmoid	—	—	—	—	—

Таблиця А.2 – Детальна структура кодера мел кепстру

Послідовність елементів мережі	Кількість вхідних каналів	Кількість вихідних каналів	Розмір ядра (фільтра)	Зсув (stride)	Розтягнення (dilation)
Conv+LayerNorm	80	256	1	1	1
ReLU	—	—	—	—	—
Conv+LayerNorm	256	256	1	1	1
ReLU	—	—	—	—	—
Conv+LayerNorm	256	256	1	1	1
Conv+LayerNorm	256	256	3	1	3
Sigmoid	—	—	—	—	—
Conv+LayerNorm	256	256	3	1	3
Sigmoid	—	—	—	—	—

Таблиця А.3 – Детальна структура декодера

Послідовність елементів мережі	Кількість вхідних каналів	Кількість вихідних каналів	Розмір ядра (фільтра)	Зсув (stride)	Розтягнення (dilation)
Conv+LayerNorm	512	256	1	1	1
Conv+LayerNorm	256	256	3	1	1
Sigmoid	—	—	—	—	—
Conv+LayerNorm	256	256	3	1	3
Sigmoid	—	—	—	—	—
Conv+LayerNorm	256	256	3	1	9
Sigmoid	—	—	—	—	—
Conv+LayerNorm	256	256	3	1	27
Sigmoid	—	—	—	—	—
Conv+LayerNorm	256	256	3	1	1
Sigmoid	—	—	—	—	—

Кінець таблиці А.3

Послідовність елементів мережі	Кількість вхідних каналів	Кількість вихідних каналів	Розмір ядра (фільтра)	Зсув (stride)	Розтягнення (dilation)
Conv+LayerNorm	256	256	3	1	1
Sigmoid	—	—	—	—	—
Conv+LayerNorm	256	256	1	1	1
ReLU	—	—	—	—	—
Conv+LayerNorm	256	256	1	1	1
ReLU	—	—	—	—	—
Conv+LayerNorm	256	80	1	1	1
ReLU	—	—	—	—	—

Таблиця А.4 – Детальна структура мережі MelSpec

Послідовність елементів мережі	Кількість вхідних каналів	Кількість вихідних каналів	Розмір ядра (фільтра)	Зсув (stride)	Розтягнення (dilation)
Conv+LayerNorm	80	512	1	1	1
Conv+LayerNorm	512	512	3	1	1
Sigmoid	—	—	—	—	—
Conv+LayerNorm	512	512	3	1	3
Sigmoid	—	—	—	—	—
Conv+LayerNorm	512	512	3	1	1
Sigmoid	—	—	—	—	—
Conv+LayerNorm	512	512	3	1	3
Sigmoid	—	—	—	—	—
Conv+LayerNorm	512	1024	1	1	1
Conv+LayerNorm	1024	1024	3	1	1
Sigmoid	—	—	—	—	—

Кінець таблиці А.4

Послідовність елементів мережі	Кількість вхідних каналів	Кількість вихідних каналів	Розмір ядра (фільтра)	Зсув (stride)	Розтягнення (dilation)
Conv+LayerNorm	1024	1024	3	1	1
Sigmoid	—	—	—	—	—
Conv+LayerNorm	1024	1025	1	1	1
Conv+LayerNorm	1025	1025	1	1	1
ReLU	—	—	—	—	—
Conv+LayerNorm	1025	1025	1	1	1
ReLU	—	—	—	—	—