**Київський національний університет імені Тараса Шевченка**

Факультет інформаційних технологій

Кафедра програмних систем і технологій

**УДК 004.942**

*На правах рукопису*

# ВИПУСКНА КВАЛІФІКАЦІЙНА БАКАЛАВРСЬКА РОБОТА

**Тема:** “ Дослідження проблеми розпізнавання емоцій голосу в розвитку штучного інтелекту ”

*(назва згідно з наказом ректора)*

**Спеціальність** – 121 “Інженерія програмного забезпечення”

## ПОЯСНЮВАЛЬНА ЗАПИСКА

**БР.ІПЗ - 31.00.00.000**

*(позначення)*

**Студент**

**ІПЗ-43**_____**/Валєрія ГНІДЕНКО/**
(шифр групи) (підпис) (дата)(розшифровка підпису)

**Науковий керівник**

**к.т.н.**_____**/Максим ТКАЧЕНКО/**
(посада) (підпис) (дата) (розшифровка підпису)

**Консультант з питань нормоконтролю**

**фахівець**_____**/Тамара ЧАПОВСЬКА/**
(посада) (підпис) (дата) (розшифровка підпису)

Допускається до захисту
з питань нормоконтролю

**Завідувач кафедри**

**д.т.н., проф.**_____**/Олексій БИЧКОВ/**
(посада) (підпис) (дата) (розшифровка підпису)

Київ – 2021

**Київський національний університет імені Тараса Шевченка**
Факультет інформаційних технологій
Кафедра програмних систем і технологій
Освітньо-кваліфікаційний рівень бакалавр
Спеціальність 121 “Інженерія програмного забезпечення”

ЗАТВЕРДЖЕНО
Зав. кафедри програмних систем і технологій
_____ ( Олексій БИЧКОВ )
(підпис) (ім'я та прізвище)

# ЗАВДАННЯ

## НА ВИПУСКНУ КВАЛІФІКАЦІЙНУ БАКАЛАВРСЬКУ РОБОТУ СТУДЕНТУ

Гніденко Валєрії Сергіївні

(прізвище, ім’я, по-батькові)

**1. Тема бакалаврської роботи “ Дослідження проблеми розпізнавання емоцій голосу в розвитку штучного інтелекту ”**

керівник проекту (роботи) Ткаченко Максим Васильович, к.т.н., асистент кафедри ПСТ

затверджені наказом вищого навчального закладу від “ 11 ” листопада 2020 р. № 6

**2. Строк подання студентом роботи**

**3. Вихідні дані до проекту (роботи)** Теоретичні концепції розпізнавання емоцій, розвитку штучного інтелекту. Мова програмування – Python. Середовище програмування – PyCharm Community Edition 2021.1.1.

**4. Зміст розрахунково - пояснювальної записки(перелік питань, які потрібно розробити)**

1. Висвітлення проблем штучного інтелекту
2. Розробка програми для розпізнавання емоцій голосу
3. Висвітлення проблем розпізнавання емоцій голосу та порад щодо їх вирішення
4. Введення поняття часткового емоційного інтелекту для вирішення проблеми розпізнавання емоцій голосу в розвитку штучного інтелекту

**5. Перелік графічного матеріалу (з точним зазначенням обов’язкових креслень)**

1. Зв'язок штучного інтелекту, машинного навчання, глибинного навчання (рис. 1.5.1, ст. 34)
2. Майже повний список нейронних мереж (рис. 1.5.2, ст. 35)
3. Use case diagram (Figure 1.3.1.3.1, p. 65)
4. Happiness identified (Figure 5.1, p. 70)

**6. Консультанти розділів проекту (роботи)**

| Розділ | Консультант | Підпис, дата | |
|---|---|---|---|
| | | Завдання видав | Завдання прийняв |
| Аналіз обраної теми | Ткаченко Максим Васильович | 11.01.2021 | 05.03.2021 |
| Програмна реалізація | Ткаченко Максим Васильович | 06.03.2021 | 15.05.2021 |
| Оформлення дипломної роботи | Ткаченко Максим Васильович | 16.05.2021 | 30.05.2021 |

**7. Дата видачі завдання** 11 листопада 2020 р.

**Керівник**_____(Максим ТКАЧЕНКО)

**Завдання прийняв до виконання** _____ (Валєрія ГНІДЕНКО)

# КАЛЕНДАРНИЙ ПЛАН

| № п/п | Назви етапів бакалаврської роботи | Строк виконання етапів роботи | Примітка |
|---|---|---|---|
| 1 | Підбір і вивчення літератури | 05.03.2021 | виконано |
| 2 | Аналіз концепцій розвитку штучного інтелекту | 21.03.2021 | виконано |
| 3 | Аналіз концепцій розпізнавання емоцій | 05.04.2021 | виконано |
| 4 | Розробка програми для розпізнавання емоцій голосу | 15.05.2021 | виконано |
| 5 | Опис поняття часткового емоційного інтелекту в рішенні проблем штучного інтелекту | 20.05.2021 | виконано |
| 6 | Оформлення дипломної роботи | 30.05.2021 | виконано |
| 7 | Затвердження пояснювальної записки роботи завідувачем кафедри | | виконано |

**Студент – бакалавр** _____(Валєрія ГНІДЕНКО)

**Керівник роботи** _____(Максим ТКАЧЕНКО)

# АНОТАЦІЯ

**Випускна кваліфікаційна бакалаврська робота:** 70 сторінок, 4 рисунки, 1 додаток, 27 джерел.

**Тема:** Дослідження проблеми розпізнавання емоцій голосу в розвитку штучного інтелекту

**Об'єкт дослідження:** розвиток штучного інтелекту.

**Мета роботи:** дослідження проблеми розпізнавання емоцій голосу в розвитку штучного інтелекту шляхом розробки програми для розпізнавання емоцій голосу. На базі такої програми можна визначити ключові проблеми розпізнавання емоцій голосу та запропонувати концепти їх вирішення.

**Предмет дослідження:** розпізнавання емоцій голосу за допомогою методів машинного навчання.

**Результати дослідження:**

1) висвітлено проблеми штучного інтелекту;

2) створено програму для розпізнавання емоцій голосу;

3) висвітлено проблеми розпізнавання емоцій голосу та поради щодо їх вирішення;

4) введено поняття часткового емоційного інтелекту для вирішення проблеми розпізнавання емоцій голосу в розвитку штучного інтелекту.

**Висновок**

В результаті досліджень було запропоновано новий підхід для вирішення проблем штучного інтелекту, що базується на використанні різних видів розпізнавання емоцій, серед яких є розпізнавання емоцій голосу.

ШТУЧНИЙ ІНТЕЛЕКТ, ШТУЧНИЙ ЗАГАЛЬНИЙ ІНТЕЛЕКТ, РОЗПІЗНАВАННЯ ЕМОЦІЙ ГОЛОСУ, ГЛИБИННЕ НАВЧАННЯ, ШТУЧНІ НЕЙРОННІ МЕРЕЖІ, ЧАСТКОВИЙ ЕМОЦІЙНИЙ ІНТЕЛЕКТ.

# АННОТАЦИЯ

**Выпускная квалификационная бакалаврская работа:** 70 страниц, 4 рисунка, 1 приложение, 27 источников.

**Тема:** Исследование проблемы распознавания эмоций голоса в развитии искусственного интеллекта

**Объект исследования:** развитие искусственного интеллекта.

**Цель работы:** исследование проблемы распознавания эмоций голоса в развитии искусственного интеллекта путем разработки программы для распознавания эмоций голоса. На базе такой программы можно определить ключевые проблемы распознавания эмоций голоса и предложить концепты их решения.

**Предмет исследования:** распознавание эмоций голоса с помощью методов машинного обучения.

**Результаты исследования:**

1) представлены проблемы искусственного интеллекта;

2) создана программа для распознавания эмоций голоса;

3) представлены проблемы распознавания эмоций голоса и советы по их решению;

4) введено понятие частичного эмоционального интеллекта для решения проблемы распознавания эмоций голоса в развитии искусственного интеллекта.

**Вывод**

В результате исследований был предложен новый подход для решения проблем искусственного интеллекта, основанный на использовании различных видов распознавания эмоций, среди которых распознавание эмоций голоса.

ИСКУСТВЕННЫЙ ИНТЕЛЛЕКТ, ИСКУСТВЕННЫЙ ОБЩИЙ ИНТЕЛЛЕКТ, РАСПОЗНАВАНИЕ ЭМОЦИЙ ГОЛОСА, ГЛУБОКОЕ ОБУЧЕНИЕ, ИСКУСТВЕННЫЕ НЕЙРОННЫЕ СЕТИ, ЧАСТИЧНЫЙ ЭМОЦИОНАЛЬНЫЙ ИНТЕЛЛЕКТ.

# ANNOTATION

**Graduation qualifying bachelor's work:** 70 pages, 4 applications, 1 appendix, 27 sources.

**Theme:** Research of the problem of recognition of voice emotions in the development of artificial intelligence

**Object of study:** development of artificial intelligence.

**The goal of the work:** research of the problem of voice emotion recognition in the development of artificial intelligence by developing a program for voice emotion recognition. On the basis of such a program, key problems in recognizing the emotions of the voice can be identified and concepts to solve them can be offered.

**Subject of study:** recognition of voice emotions using machine learning methods.

**Results of the research:**

1) the problems of artificial intelligence were presented;

2) a program for recognizing the emotions of the voice was created;

3) the problems of recognizing the emotions of the voice and advices on how to solve them were presented;

4) the concept of partial emotional intelligence to solve the problem of recognizing the emotions of the voice in the development of artificial intelligence was introduced.

**Conclusion**

Research has proposed a new approach to solving problems of artificial intelligence, based on the use of different types of emotion recognition, among which is the recognition of voice emotions.

ARTIFICIAL INTELLIGENCE, ARTIFICIAL GENERAL INTELLIGENCE, RECOGNITION OF VOICE EMOTIONS, DEEP LEARNING, ARTIFICIAL NEURAL NETWORKS, PARTIAL EMOTIONAL INTELLIGENCE.

# ЗМІСТ

# ПЕРЕЛІК ОСНОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, СКОРОЧЕНЬ

БП - багатошаровий персептрон

ГНМ - глибинна нейронна мережа

ЗНМ - згорткова нейронна мережа

ОПМ - обробка природної мови

РЕГ - розпізнавання емоцій голосу

ЧЕІ - частковий емоційний інтелект

ШЗІ - штучний загальний інтелект

# ВСТУП

**Актуальність роботи**

На даний момент штучний інтелект має більшу частину проблем саме через нездатність думати “як людина”. Іншими словами, щось на зразок сарказму вже може зробити його припущення некоректним, адже слова окремо від промови можуть мати різний зміст. Для того, щоб штучний інтелект міг опановувати більш складні поняття та читати мотиви людей залучення розпізнавання емоцій є надзвичайно необхідним. Саме тому розпізнавання емоцій голосу на даний момент часу має невичерпну актуальність та великі перспективи розвитку.

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

У даній роботі широко використовується питання розвитку штучного інтелекту, що тісно пов'язано з безліччю наукових програм, планів та тем, а реалізація технології розпізнавання емоцій голосу відчиняє двері до різних галузей починаючи від мультимедійної та закінчуючи медичною.

**Мета і задачі дослідження**

**Метою** бакалаврської роботи є дослідження проблеми розпізнавання емоцій голосу в розвитку штучного інтелекту шляхом розробки програми для розпізнавання емоцій голосу. На базі такої програми можна визначити ключові проблеми розпізнавання емоцій голосу та запропонувати концепти їх вирішення.

Досягнення мети включало розв’язання таких **задач**:

1) висвітлення проблем штучного інтелекту;
2) створення програми для розпізнавання емоцій голосу;
3) висвітлення проблем розпізнавання емоцій голосу та порад щодо їх вирішення;
4) введення поняття часткового емоційного інтелекту для вирішення проблеми розпізнавання емоцій голосу в розвитку штучного інтелекту.

**Об’єктом дослідження** є розвиток штучного інтелекту.

**Предметом дослідження** є розпізнавання емоцій голосу за допомогою методів машинного навчання.

**Методи дослідження**

Для дослідження проблем розпізнавання емоцій голосу шляхом створення програми для розпізнавання емоцій голосу було використано статистичні методи, штучні нейронні мережі, глибинні нейронні мережі.

**Наукова новизна отриманих результатів**

Запропоновано новий підхід для вирішення проблем штучного інтелекту, що базується на використанні різних видів розпізнавання емоцій, серед яких є розпізнавання емоцій голосу. Через брак великих баз даних та ряд інших специфічних труднощів з поміж усіх видів розпізнавання емоцій розпізнавання емоцій голосу є найскладнішим, найменш дослідженим, а разом з тим й найбільш перспективним. Саме тому проблеми розпізнавання емоцій голосу було досліджено в рамках цієї роботи та надано концепти їх рішення.

**Практичне значення одержаних результатів**

Хоча розпізнавання емоцій голосу розглядалося насамперед як одна з компонентів першої складової часткового емоційного інтелекту для пришвидшення розвитку штучного інтелекту та штучного загального інтелекту, цю технологію можна застосовувати також в перекладі з однієї мови на іншу, інтерактивних онлайн-уроках і курсах, забезпеченні безпеки транспортних засобів, сеансах терапії, стресових і галасливих умовах, наприклад в кабінах літаків, сфері послуг і електронній комерції, кол-центрах, інтерактивних фільмах, і т. д.

**Особистий внесок студента**

Основними результатами є:

1. представлення проблем штучного інтелекту як вичерпних;

2. створення програми, завдяки якій було визначено проблеми розпізнавання емоцій голосу та згодом надано концепти їх вирішення;
3. введення поняття часткового емоційного інтелекту для вирішення проблеми розпізнавання емоцій голосу в розвитку штучного інтелекту (запропоновано концепт, що за допомогою розпізнавання емоцій голосу наряду з іншими компонентами першої складової часткового емоційного інтелекту дозволить вирішити майже всі проблеми штучного інтелекту).

**Апробація результатів випускної кваліфікаційної бакалаврської роботи**

Результати бакалаврського дослідження не були представлені на наукових з'їздах, конференціях, симпозіумах, нарадах.

**Публікації**

Результати наукових досліджень, проведених у бакалаврській роботі, було опубліковано в збірнику тез MSTIoE-8.

**Структура та обсяг роботи**

Робота викладена на 70 сторінках друкованого тексту, який складається із титульної сторінки, завдання на роботу, календарного плану, анотацій на українській, російській та англійській мовах, змісту, переліку основних позначень, символів, скорочень, вступу, двох розділів, висновків, списку використаних джерел (27 найменувань), одного додатку. Робота містить 4 рисунки та 1 додаток, обсягом 9 сторінок.

# РОЗДІЛ 1

# РОЗВИТОК ШТУЧНОГО ІНТЕЛЕКТУ

## 1.1 Штучний інтелект

Штучний інтелект (ШІ) - це інтелект, продемонстрований машинами, на відміну від природного інтелекту, що притаманний людям та тваринам і включає в себе свідомість та емоційність. Відмінність між першою та другою категоріями часто відображається через обрану абревіатуру. "Сильний" ШІ зазвичай позначають як штучний загальний інтелект (ШЗІ), тоді як спроби наслідувати "природний" інтелект називають штучним біологічним інтелектом (ШБІ). Провідні підручники з штучного інтелекту визначають цю галузь як вивчення «інтелектуальних агентів»: будь-який пристрій, який сприймає навколишнє середовище та здійснює дії, що максимізують його шанси на успішне досягнення своїх цілей. У розмовній формі термін "штучний інтелект" часто використовується для опису машин, що імітують "когнітивні" функції, які люди пов'язують з людським розумом, такі як "навчання" та "вирішення проблем".[1]

У міру того, як машини стають все більш працездатними, завдання, які, як вважають, вимагають "інтелекту", часто вилучаються із визначення ШІ. Таке явище відоме як ефект ШІ.[2]  Помітка в теоремі Теслера говорить: "ШІ - це все, що ще не зроблено". Наприклад, оптичне розпізнавання символів часто виключається з речей, які вважаються ШІ, бо воно стало звичною технологією. Сучасні можливості машин, які зазвичай класифікуються як ШІ, включають успішне розуміння людської мови, змагання на найвищому рівні в стратегічних ігрових системах (таких як шахи та Go), а також ігри з недосконалою інформацією, такі як покер, самокеровані машини, інтелектуальну маршрутизацію в мережах доставки вмісту та військові симуляції.

Штучний інтелект був заснований як навчальна дисципліна в 1955 році, і за ці роки пережив кілька хвиль злетів, за якими слідували падіння та втрата фінансування, які потім замінювалися на нові підходи, успіх та оновлене фінансування. Після того, як

AlphaGo успішно перемогла професійного гравця Go в 2015 році, штучний інтелект знову привернув широку увагу світу. Протягом більшої частини своєї історії дослідження ШІ були розділені на різні напрямки, які часто не взаємодіють між собою. Ці напрямки базуються на технічних міркуваннях, таких як конкретні цілі (наприклад, "робототехніка" або "машинне навчання"), використанні певних інструментів ("логіка" або штучні нейронні мережі) або глибоких філософських відмінностях. Напрямки також базуються на соціальних факторах (конкретні установи чи робота певних дослідників).

Традиційні проблеми (або цілі) дослідження ШІ включають міркування, подання знань, планування, навчання, обробку природної мови, сприйняття та здатність рухатись та маніпулювати предметами. ШЗІ є серед довгострокових цілей галузі. Підходи включають статистичні методи, обчислювальний інтелект та традиційний символічний ШІ. У ШІ використовується багато інструментів, включаючи версії пошуку та математичної оптимізації, штучні нейронні мережі та методи, засновані на статистиці, ймовірності та економіці. Галузь штучного інтелекту базується на інформатиці, інформаційній інженерії, математиці, психології, лінгвістиці, філософії та багатьох інших галузях.

ШІ було засновано на припущенні, що людський інтелект "можливо настільки точно описати, що можна створити машину для його імітації". Це викликає філософські дискусії щодо розуму та етики створення штучних істот, наділених людським інтелектом. Такі проблеми досліджувались міфами, художньою літературою та філософією ще з античності. Деякі люди також вважають ШІ небезпекою для людства, якщо він поступово прогресує. Інші вважають, що ШІ, на відміну від попередніх технологічних революцій, створить ризик масового безробіття.

У двадцять першому столітті методи ШІ пережили відродження після одночасного зростання комп'ютерної потужності, великого обсягу даних та теоретичного розуміння. ШІ став важливою частиною технологічної індустрії, допомагаючи

вирішити багато складних проблем в галузі інформатики, програмного забезпечення та досліджень операцій.

## 1.2 Проблеми

### 1.2.1 Обмеження

Когнітивні можливості сучасних архітектур дуже обмежені, адже вони використовують лише спрощену версію того, на що насправді здатний інтелект. Наприклад, людський розум придумав способи міркування поза мірою та логічні пояснення різних явищ у житті. Те що було б легко для людського розуму, могло б стати складністю для обчислювального вирішення. Це породжує два класи моделей: структурні та функціональні. Структурні моделі мають на меті імітувати основні інтелектуальні операції розуму, такі як міркування та логіка. Функціональна модель відноситься до корелюючих даних з їх обчислювальним аналогом.

Загальною метою дослідження штучного інтелекту є створення технології, яка дозволяє комп'ютерам і машинам функціонувати розумно. Загальна проблема моделювання (або створення) інтелекту була розбита на підкатегорії. Вони складаються з особливих рис або можливостей, які дослідники очікують від інтелектуальної системи. Описаним нижче особливостям приділяють найбільшу увагу.

### 1.2.2 Міркування, вирішення проблем

Ранні дослідники розробляли алгоритми, що імітували покрокові міркування, якими користуються люди, коли вони розгадують головоломки або роблять логічні висновки. В кінці 1980-х та 1990-х років дослідження ШІ розробило методи роботи з невизначеною або неповною інформацією, використовуючи концепції ймовірності та економіки.

Цих алгоритмів виявилося недостатньо для вирішення великих міркувань, оскільки вони зазнали "комбінаторного вибуху": вони стали експоненціально повільнішими, оскільки проблеми зростали. Навіть люди рідко використовують

покрокову дедукцію, яку могли б моделювати ранні дослідження ШІ. Вони вирішують більшість своїх проблем, використовуючи швидкі, інтуїтивні судження.

### 1.2.3 Представлення знань

Представлення знань[3] та інженерія знань є центральними для класичних досліджень ШІ. Деякі "експертні системи" намагаються зібрати явні знання, якими володіють експерти в якійсь вузькій області. Крім того, деякі проекти намагаються зібрати "знання здорового глузду",[4] відомі пересічній людині, у базу даних, що містить великі знання про світ. Серед речей, що містять всеосяжну базу знань про здоровий глузд, є: об’єкти, властивості, категорії та відносини між об’єктами; ситуації, події, стани та час; причини та наслідки; знання про знання (те, що ми знаємо про те, що знають інші люди); та багато інших, менш добре досліджених доменів. Представлення "того, що існує" - це онтологія: сукупність об'єктів, відносин, концепцій та властивостей, формально описаних для того, щоб програмні агенти могли їх інтерпретувати. Семантика їх фіксується як поняття логіки опису, ролі та особи, і зазвичай реалізується як класи, властивості та особи на мові веб-онтології. Найбільш загальні онтології називаються верхніми онтологіями, які намагаються створити основу для всіх інших знань,[5] виступаючи посередниками між онтологіями доменів, які охоплюють конкретні знання про конкретну область знань (сферу інтересу або занепокоєння). Такі офіційні подання знань можуть бути використані в індексації та пошуку на основі вмісту, інтерпретації сцени, підтримці клінічних рішень, виявленні знань (видобуток "цікавих" та дійових висновків із великих баз даних) та інших областях.

Серед найскладніших проблем представлення знань є:

- Міркування за замовчуванням та проблема кваліфікації

  Багато речей, які люди знають, мають форму "робочих припущень". Наприклад, якщо в розмові згадується птах, то люди зазвичай представляють тварину розміром з кулак, яка співає і літає. Ніщо з цього не притаманно абсолютно всім птахам. Джон Маккарті визначив цю проблему в 1969 році як проблему кваліфікації: для будь-якого правила здорового глузду, яке

дослідники ШІ хочуть представляти, існує величезна кількість винятків. Майже ніщо не є просто істинним чи хибним, як того вимагає абстрактна логіка. Дослідження ШІ досліджувало ряд рішень цієї проблеми.

- Обсяг здорового глузду

  Кількість атомних фактів, які знає пересічна людина, дуже велика. Дослідницькі проекти, які намагаються створити повну базу знань здорового глузду (наприклад, Cyc), вимагають величезних обсягів кропіткої онтологічної інженерії - їх потрібно будувати вручну, по одній складній концепції за раз. Через те, що така база вимагає вкладання великих ресурсів ці проекти не можуть одразу ж процвітати.

- Підсвідома форма деяких знань здорового глузду

  Багато з того, що люди знають, не представляється як "факти" чи "твердження", які вони могли б висловити усно. Наприклад, шаховий майстер уникатиме певної шахової позиції, оскільки він "відчуває себе занадто відкритим", або мистецтвознавець може один раз поглянути на статую і зрозуміти, що це підробка. Це несвідомі твердження та підсвідома інтуїція чи звички в мозку людини. Такі знання інформують, підтримують та забезпечують контекст для символічного, свідомого знання. Як і у відповідній проблемі підсимволічних міркувань, очікується, що одного разу розміщений ШІ, обчислювальний інтелект або статистичний ШІ забезпечать способи представлення цих знань.

### 1.2.4 Планування

Розумні агенти повинні вміти ставити цілі та досягати їх.[6] Їм потрібен спосіб візуалізації майбутнього - уявлення про стан світу і можливість робити прогнози щодо того, як їх дії змінять його - і мати можливість робити вибір, який максимізує корисність (або "цінність") доступного вибору.

У класичних задачах планування агент може припустити, що це єдина система, яка діє у світі, що дозволяє агенту бути впевненим у наслідках своїх дій. Однак, якщо агент не є єдиною дійовою особою, тоді йому потрібно, щоб він міг міркувати в

умовах невизначеності. Це вимагає агента, який може не тільки оцінювати своє середовище та робити прогнози, але й також оцінювати свої прогнози та адаптуватися на основі своєї оцінки.

Багатоагентне планування використовує співпрацю та конкуренцію багатьох агентів для досягнення заданої мети. Таку поведінку, що виникає, використовують еволюційні алгоритми та інтелект роїв.

### 1.2.5 Навчання

Машинне навчання[7] (МН), основне поняття досліджень ШІ з моменту створення галузі, - це вивчення комп’ютерних алгоритмів, які автоматично вдосконалюються завдяки досвіду.

Навчання без нагляду - це здатність знаходити закономірності у потоці вхідних даних, не вимагаючи, щоб людина спочатку позначала вхідні дані. Навчання під контролем включає як класифікацію, так і чисельну регресію, що вимагає від людини першого маркування вхідних даних. Класифікація використовується, щоб визначити, до якої категорії щось належить, і відбувається після того, як програма побачить ряд прикладів речей з декількох категорій. Регресія - це спроба створити функцію, яка описує взаємозв'язок між входами та виходами та передбачає, як повинні змінюватися виходи при зміні входів. Класифікатори та регресивні учні можуть розглядатися як "апроксиматори функцій", які намагаються вивчити невідому (можливо неявну) функцію; наприклад, класифікатор спаму можна розглядати як вивчення функції, яка переводить текст електронного листа в одну з двох категорій, "спам" або "не спам". Теорія обчислювального навчання може оцінювати учнів за складністю обчислень, за складністю вибірки (скільки даних потрібно) або за іншими поняттями оптимізації. При навчанні через підкріплення агент винагороджується за хороші відповіді та карається за погані. Агент використовує цю послідовність нагород та покарань, щоб сформувати стратегію роботи у своєму проблемному просторі.

### 1.2.6 Обробка природної мови

Обробка природної мови[8] (ОПМ) дозволяє машинам читати та розуміти людську мову. Досить потужна система обробки природних мов підтримувала б користувальницький інтерфейс на природній мові та могла б отримувати знання безпосередньо з написаних людиною джерел, таких як тексти новин. Деякі прямі програми обробки природних мов включають пошук інформації, видобуток тексту, відповіді на запитання та машинний переклад. Багато сучасних підходів використовують частоти спільних випадків слів для побудови синтаксичних подань тексту. Стратегії пошуку ключових слів - популярні та масштабовані, але не надто розумні; пошуковий запит "собака" може збігатися лише з документами з буквальним словом "собака" і пропускати документ зі словом "пудель". Стратегії "лексичної спорідненості" використовують появу таких слів, як "нещасний випадок", для оцінки настрою документа. Сучасні статистичні підходи ОПМ можуть поєднувати всі ці стратегії, як і інші, і часто досягати прийнятної точності на рівні сторінки чи абзацу. Окрім семантичної ОПМ, кінцевою метою "наративної" ОПМ є втілення повного розуміння міркувань здорового глузду. В 2019 році трансформаторні архітектури глибокого навчання вже могли генерувати зв’язний текст.

### 1.2.7 Сприйняття

Машинне сприйняття[9] - це здатність використовувати вхідні сигнали від датчиків (таких як камери (видимий спектр або інфрачервоний діапазон), мікрофони, бездротові сигнальні пристрої та активні лідарні, сонарні, радіолокаційні та тактильні датчики) для простежування аспектів світу. До програм належать розпізнавання мови, розпізнавання обличчя та розпізнавання об’єктів. Комп’ютерний зір - це здатність аналізувати візуальні дані. Таке введення, як правило, неоднозначне; гігантський 50-метровий пішохід на далекій відстані може видавати ті самі пікселі, що й сусідній пішохід нормального розміру, вимагаючи від ШІ оцінки відносної ймовірності та обґрунтованості різних інтерпретацій, наприклад, використовуючи свою "об'єктну модель", щоб оцінити, що 50-метрових пішоходів не існує.

### 1.2.8 Рух та маніпуляція

ШІ широко використовується в робототехніці.[10] Удосконалені роботизовані озброєння та інші промислові роботи, що активно задіються на сучасних фабриках, можуть навчитися з власного досвіду як ефективно рухатися, незважаючи на наявність тертя та ковзання механізмів. Сучасний мобільний робот, маючи невелике, статичне та видиме середовище, може легко визначити своє місце розташування та скласти карту свого оточення; однак динамічне середовище, таке як (при ендоскопії) внутрішня частина дихального тіла пацієнта, становить більший виклик. Планування руху - це процес розбиття рухового завдання на «примітиви», такі як окремі рухи суглобами. Такий рух часто передбачає поступливий рух - процес, коли рух вимагає підтримання фізичного контакту з предметом. Парадокс Моравека узагальнює, що сенсомоторні навички низького рівня, які люди сприймають як дане, важко запрограмувати для робота. Даний парадокс названий на честь Ганса Моравека, який заявив у 1988 р., що "порівняно легко змусити комп'ютери демонструвати показники рівня дорослих на тестах інтелекту або граючи в шашки, і важко або неможливо дати їм навички однорічного віку, коли справа стосується сприйняття та рухливості". Це пояснюється тим, що, на відміну від шашок, фізична спритність була безпосередньою ціллю природного відбору протягом мільйонів років.

### 1.2.9 Соціальний інтелект

Парадокс Моравека можна поширити на багато форм соціального інтелекту. Розподілена багатоагентна координація автономних транспортних засобів залишається складною проблемою. Афективні обчислення[11] - це міждисциплінарне різноманіття, що включає системи, які розпізнають, інтерпретують, обробляють або імітують афекти людини. Помірний успіх, пов’язаний з афективними обчисленнями, включає аналіз текстових настроїв та, з нещодавнього часу, мультимодальний аналіз афектів, де ШІ класифікує афекти, які відображаються на відеозаписі.

Зрештою, соціальні навички, розуміння людських емоцій та теорії ігор були б цінними для соціального агента. Здатність передбачати дії інших, розуміючи їх мотиви та емоційні стани, дозволить агенту приймати кращі рішення. Деякі

комп’ютерні системи імітують людські емоції та вирази, щоб виглядати більш чутливими до емоційної динаміки людської взаємодії або іншим чином сприяти взаємодії людини з комп’ютером. Так само деякі віртуальні помічники запрограмовані говорити в розмовному стилі або навіть жартівливо глузувати; це, як правило, дає наївним користувачам нереальне уявлення про те, наскільки насправді розумними є існуючі комп'ютерні агенти.

### 1.2.10 Загальний інтелект

Історично такі проекти, як база знань Cyc та масштабна японська ініціатива комп’ютерних систем п’ятого покоління, намагалися охопити широту людського пізнання. Вони не змогли уникнути обмежень кількісних символічних логічних моделей і, ретроспективно, значно недооцінили труднощі міждоменного ШІ. В даний час більшість сучасних дослідників ШІ натомість працюють над відслідковуваними "вузькими ШІ" (такими як медична діагностика або автомобільна навігація). Багато дослідників передбачають, що така робота "вузького ШІ" в різних окремих областях врешті-решт буде інтегрована в машину зі штучним загальним інтелектом (ШЗІ), поєднуючи більшість вузьких навичок і в певний момент навіть перевищуючи людські здібності в більшості або в усіх цих сферах.[12] Багато досягнень мають загальне, міждоменне значення. Одним з гучних прикладів є те, що DeepMind у 2010-х роках розробив "узагальнений штучний інтелект", який міг самостійно вивчати багато різноманітних ігор Atari, а пізніше розробив варіант системи, який досягає успіху в послідовному навчанні. Окрім трансферного навчання, гіпотетичні прориви ШЗІ можуть включати в себе розробку рефлексивних архітектур, які матимуть змогу брати участь в мета-міркуваннях, заснованих на теорії прийняття рішень, і з'ясовуванні того, як "витягти" всеосяжну базу знань з усієї неструктурованої мережі. Деякі стверджують, що якийсь (на даний момент нерозкритий) концептуально простий, але математично складний "головний алгоритм" може призвести до появи ШЗІ. Нарешті, кілька «розвиваючихся» підходів дуже уважно розглядають моделювання людського інтелекту і вважають, що антропоморфні особливості, такі

як штучний мозок або моделювання розвитку дитини, можуть одного дня досягти критичної точки, коли з’являється загальний інтелект.

## 1.3 Підходи

### 1.3.1 Кібернетика та моделювання мозку

У 1940-х та 1950-х роках ряд вчених досліджував зв'язок між нейробіологією, теорією інформації та кібернетикою.[13] Деякі з них побудували машини, які використовували електронні мережі для демонстрування елементарного інтелекту, такі як черепахи В. Грея Уолтера та звір Джонса Хопкінса. Багато з цих дослідників збиралися на засідання Телеологічного товариства в Принстонському університеті та клубу Ratio в Англії. До 1960 р. від даного підходу здебільшого відмовились, хоча його елементи були відроджені у 1980-х роках.

### 1.3.2 Символічний підхід

Коли в середині 1950-х років доступ до цифрових комп’ютерів став можливим, почалося дослідження вірогідності того, що людський інтелект може бути зведений до маніпуляцій із символами.[14] Дослідження було зосереджено у трьох закладах: Університеті Карнегі-Меллона, Стенфорді та Массачусетському технологічному інституті, і, як описано нижче, кожен розробив свій власний стиль дослідження. Джон Хаугленд назвав ці символічні підходи до ШІ "старомодним ШІ". Протягом 1960-х символічні підходи досягли великих успіхів у моделюванні "мислення" на високому рівні в невеликих демонстраційних програмах. Підходи, засновані на кібернетиці або штучних нейронних мережах, були відкинуті або відсунуті на другий план. Дослідники 1960-х та 1970-х років були впевнені, що символічні підходи в результаті зможуть створити машину зі штучним загальним інтелектом, і вважали це метою своєї галузі.

#### 1.3.2.1 Когнітивне моделювання

Економіст Герберт Саймон та Аллен Ньюел вивчали людські навички вирішення проблем та намагалися їх формалізувати. Їх робота заклала основи галузі штучного

інтелекту, а також когнітивної науки, досліджень операцій та науки управління. Дослідницька група Саймона та Ньюелла використовувала результати психологічних експериментів для розробки програм, що імітували методи, які люди використовували для вирішення проблем. Ця традиція, зосереджена в Університеті Карнегі-Меллона, врешті-решт завершилася розвитком архітектури Soar у середині 1980-х.

#### 1.3.2.2 На основі логіки

На відміну від Саймона та Ньюелла, Джон Маккарті вважав, що машинам не потрібно імітувати людські думки, а натомість слід намагатися знайти суть абстрактних міркувань та вирішення проблем, незалежно від того, застосовували люди однакові алгоритми чи ні. Його лабораторія в Стенфорді (SAIL) зосередилася на використанні формальної логіки для вирішення найрізноманітніших проблем, включаючи представлення знань, планування та навчання. Логіка також була в центрі уваги роботи в Единбурзькому університеті та в інших місцях Європи, що призвело до розвитку мови програмування Prolog та науки про логічне програмування.

#### 1.3.2.3 Анти-логічний або “брудний”

Дослідники з Массачусетського технологічного інституту (такі як Марвін Мінський та Сеймур Паперт) виявили, що вирішення складних проблем із зором та обробкою природних мов вимагає спеціальних рішень - вони стверджували, що жоден простий і загальний принцип (як логіка) не охоплює всіх аспектів розумної поведінки. Роджер Шенк описав їхні "антилогічні" підходи як "брудні" (на відміну від "чистих" парадигм в Університеті Карнегі-Меллона та Стенфорді). Бази знань про здоровий глузд (наприклад, Сус Дуга Лената) - це приклад "брудного" ШІ, оскільки їх потрібно будувати вручну, по одній складній концепції за раз.

#### 1.3.2.4 На основі знань

Коли приблизно в 1970 р. стали доступні комп’ютери з великою пам’яттю, дослідники з усіх трьох традицій почали вбудовувати знання в програми ШІ. Ця "революція знань" призвела до розробки та впровадження експертних систем

(запроваджених Едвардом Фейгенбаумом), першої справді успішної форми програмного забезпечення для штучного інтелекту. Ключовим компонентом архітектури системи для всіх експертних систем є база знань, яка зберігає факти та правила, що ілюструють ШІ. Революція знань також була зумовлена усвідомленням того, що багато простих програм ШІ потребують величезних обсягів знань.

### 1.3.3 Підсимволічний підхід

До 1980-х років прогрес у символічному ШІ, здавалося, зупинився, і багато хто вірив, що символічні системи ніколи не зможуть наслідувати всі процеси людського пізнання, особливо сприйняття, робототехніку, навчання та розпізнавання зразків. Ряд дослідників почав розглядати "субсимволічні" підходи до конкретних проблем ШІ. Субсимволічні методи дають змогу наблизитися до інтелекту без конкретних представлень знань.

#### 1.3.3.1 Втілений інтелект

Сюди входять втілений, локалізований, заснований на поведінці та новітній ШІ. Дослідники із суміжної галузі робототехніки, такі як Родні Брукс, відкинули символічний ШІ та зосередилися на основних інженерних проблемах, які дозволять роботам рухатися та виживати. Їхня робота відродила несимволічну точку зору дослідників ранньої кібернетики 1950-х років і відновила використання теорії управління в ШІ. Це співпало з розвитком тези про втілення розуму у відповідній галузі когнітивної науки: ідеї про те, що аспекти тіла (такі як рух, сприйняття та візуалізація) потрібні для вищого інтелекту.

У рамках розвитку робототехніки розробляються підходи до розвиваючого навчання, щоб дозволити роботам накопичувати репертуар нових навичок шляхом автономного самодослідження, соціальної взаємодії з вчителями-людьми та використання механізмів керівництва (активне навчання, дорослішання, синергія руху тощо).

#### 1.3.3.2 Обчислювальний інтелект та м’які обчислення

Інтерес до нейронних мереж та "коннекціонізму" був відроджений Девідом Румельхартом та іншими в середині 1980-х. Штучні нейронні мережі є прикладом м'яких обчислень – вони є рішенням проблем, які неможливо вирішити з повною логічною визначеністю, і де приблизного рішення часто буває достатньо. Інші м’які обчислювальні підходи до ШІ включають нечіткі системи, теорію систем Грея, еволюційні обчислення та багато статистичних інструментів. Застосування м'яких обчислень до ШІ вивчається колективно за допомогою нової дисципліни обчислювального інтелекту.

### 1.3.4 Статистичний підхід

Значна частина традиційних "старомодних ШІ" загрузла в спеціальних виправленнях для символічних обчислень, які працювали на їх власних іграшкових моделях, але не могли узагальнити їх на реальні результати. Однак приблизно в 1990-х роках дослідники ШІ застосували складні математичні інструменти, такі як приховані моделі Маркова (ПММ), теорія інформації та нормативна теорія рішення Байєса для порівняння або об'єднання конкуруючих архітектур. Спільна математична мова дозволила забезпечити високий рівень співпраці з більш усталеними галузями (такими як математика, економіка чи дослідження операцій). Порівняно з "старомодним ШІ", нові методи "статистичного навчання", такі як ПММ та нейронні мережі, набували вищих рівнів точності у багатьох практичних областях, таких як видобуток даних, без необхідності отримувати семантичне розуміння наборів даних. Посилення успіхів із реальними даними призвело до посилення акценту на порівнянні різних підходів із даними спільних тестів, щоб побачити, який із підходів є найкращим у ширшому контексті, ніж той, що забезпечує ідіосинкратичні іграшкові моделі; дослідження ШІ ставали все більш науковими. У наш час результати експериментів часто суворо вимірюються, та іноді (з труднощами) відтворюються. Різні методи статистичного навчання мають різні обмеження; наприклад, основні ПММ не можуть моделювати нескінченно можливі поєднання природної мови. Критики зазначають, що перехід від "старомодного ШІ" до статистичного навчання часто також є відходом від пояснюваного ШІ. У дослідженні ШЗІ деякі вчені

застерігають від надмірної залежності від статистичного навчання та стверджують, що для досягнення загального рівня інтелекту все ж будуть необхідні постійні дослідження "старомодного ШІ".

### 1.3.5 Інтеграція підходів

#### 1.3.5.1 Парадигма інтелектуального агента

Інтелектуальний агент - це система, яка сприймає навколишнє середовище та здійснює дії, що максимізують шанси на успіх. Найпростіші інтелектуальні агенти - це програми, що вирішують конкретні проблеми. Більш складні агенти включають людей і людські організації (наприклад, фірми). Парадигма дозволяє дослідникам безпосередньо порівнювати або навіть комбінувати різні підходи до ізольованих проблем, запитуючи, який агент найкраще максимізує дану "цільову функцію". Агент, який вирішує конкретну проблему, може використовувати будь-який підхід, який працює - деякі агенти є символічними та логічними, деякі є субсимволічними штучними нейронними мережами, а інші можуть використовувати нові підходи. Парадигма також надає дослідникам спільну мову для спілкування з іншими галузями - такими, як теорія рішень та економіка, - які теж використовують поняття абстрактних агентів. Створення повноцінного агента вимагає від дослідників вирішення реалістичних проблем інтеграції; наприклад, оскільки сенсорні системи дають невизначену інформацію про навколишнє середовище, системи планування повинні мати можливість функціонувати за наявності невизначеності. Парадигма інтелектуального агента набула широкого визнання протягом 1990-х.

#### 1.3.5.2 Агентські архітектури та когнітивні архітектури

Дослідники розробили системи для побудови інтелектуальних систем із взаємодіючих інтелектуальних агентів у мультиагентній системі. Ієрархічна система управління забезпечує міст між субсимволічним ШІ на найнижчому, реактивному рівнях та традиційним символічним ШІ на найвищих рівнях, де ослаблені часові обмеження дозволяють планувати та моделювати світ. Деякі когнітивні архітектури створені на замовлення для вирішення вузької проблеми; інші, наприклад Soar,

призначені для імітації людського пізнання та надання розуміння загального інтелекту. Сучасні розширення Soar - це гібридні інтелектуальні системи, що включають як символічні, так і субсимволічні компоненти.

## 1.4 Глибинне навчання

### 1.4.1 Основні поняття та призначення

Штучні нейронні мережі (ШНМ) є одним з обчислювальних інструментів для ШІ наряду з пошуком та оптимізацією, логікою, імовірнісними методами для невизначених міркувань, класифікаторами та статичними методами навчання.[15] ШНМ можна поділити на ациклічні або прямі нейронні мережі (де сигнал проходить лише в одному напрямку) та рекурентні нейронні мережі (які забезпечують зворотній зв'язок та короткочасні спогади про попередні вхідні події). Серед найбільш популярних мереж прямого пересилання є персептрони, багатошарові персептрони та радіальні базисні мережі. В свою чергу глибинне навчання є використанням ШНМ, які мають кілька шарів нейронів між входами і виходами мережі. Глибинне навчання трансформувало багато важливих підгалузей штучного інтелекту, включаючи комп'ютерний зір, розпізнавання мови, обробку природної мови та інші.

### 1.4.2 Історія успіху глибинного навчання

Незважаючи на те що термін «глибинне навчання» з'явився в науковому співтоваристві машинного навчання тільки в 1986 році після роботи Ріни Дехтер, перший загальний робочий алгоритм для глибинних багатошарових перцептронів прямого поширення був опублікований в книзі радянських вчених Олексія Григоровича Івахненка і Валентина Григоровича Лапи «Кібернетичні пророкуючі пристрої».

Інші глибинні архітектури, особливо ті, які спеціалізуються на розпізнаванні образів, беруть свій початок з неокогнітрона, розробленого Куніхіко Фукусімою в 1980 році. У 1989 році Яну Лекуну вдалося використати алгоритм зворотного поширення помилки для навчання глибинних нейромереж для вирішення задачі

розпізнавання рукописних ZIP-кодів. Незважаючи на успішний досвід, для навчання моделі було потрібно три дні, що істотно обмежувало застосування цього методу. Низька швидкість навчання пов'язана з багатьма факторами, включаючи проблему зникаючих градієнтів через великий розкид масштабів навчаючихся параметрів, яку в 1991 році аналізували Йорген Шмідхубер і Зеппо Хохрайтер. Через ці проблем нейронні мережі в 1990-х роках поступилися місцем методу опорних векторів.

До 1991 року такі системи використовувалися для розпізнавання ізольованих двовимірних рукописних цифр, а розпізнавання тривимірних об'єктів здійснювалося шляхом зіставлення двовимірних зображень з тривимірною об'єктною моделлю, виготовленої вручну. У 1992 році створена модель кресцептрона для розпізнавання тривимірних об'єктів в захаращених сценах.

У 1994 році Андре де Карвальо, разом з Майком Фейрхерстом і Девідом Біссетом, опублікував експериментальні результати багатошарової булевої нейронної мережі, також відомої як невагома нейронна мережа, що складалася з трирівневого самоорганізованого модуля нейронної мережі для виділення ознак (SOFT), а потім модуль багаторівневої класифікації нейронної мережі (GSN). Кожен модуль пройшов незалежне один від одного навчання. Кожен шар в модулі витягував об'єкти зі зростаючою складністю відносно попереднього шару.

У 1995 році Брендан Фрай продемонстрував, що можна навчити (протягом двох днів) мережу, що містить шість повністю з'єднаних шарів і кілька сотень прихованих юнітів, використовуючи алгоритм сну-неспання, розроблений спільно з Пітером Даяном і Хінтоном. Багато факторів сприяють повільній швидкості, включаючи проблему зникаючого градієнта, проаналізовану в 1991 році Зеппом Хохрайтером.

Простіші моделі, які використовують ручні роботи, специфічні для конкретної задачі, такі як фільтри Габора і метод опорних векторів (МОВ), були популярним вибором в 1990-х і 2000-х роках через обчислювальні витрати штучної нейронної мережі (ШНМ) і відсутність розуміння того, як мозок пов'язує свої біологічні мережі.

Як поверхневе, так і глибинне навчання (наприклад, рекурентні мережі) ШНМ вивчалося протягом багатьох років. Ці методи ніколи не перевершували неоднорідну

змішану Гауссову модель і приховану модель Маркова, засновану на генеративних моделях мови, навчених дискримінаційно. Були проаналізовані ключові проблеми, в тому числі зменшення градієнта і слабка тимчасова кореляційна структура в нейронних прогностичних моделях. Додатковими труднощами були відсутність навчальних даних і обмежена обчислювальна потужність.

Глибинне навчання набуло популярності в середині 2000-х років, коли все зійшлося воєдино: комп'ютери стали досить потужними, щоб навчати великі нейронні мережі (обчислення навчилися делегувати графічним процесорам, що значно прискорило процес навчання), набори даних стали досить об'ємними, щоб навчання великих мереж мало сенс, а в теорії штучних нейронних мереж сталося чергове просування - статті Хінтона, Осіндеро і Те, а також Бенджі, в яких автори показали, що можна ефективно попередньо навчати багатошарову нейронну мережу, якщо навчати кожен шар окремо за допомогою обмеженої машини Больцмана, а потім довчати за допомогою методу зворотного поширення помилки.

У 2012 році команда під керівництвом Джорджа Е. Даля виграла конкурс «Merck Molecular Activity Challenge», використовуючи багатозадачні глибинні нейронні мережі для прогнозування біомолекулярної мішені одного препарату. У 2014 році група Хохрайтера використала глибинне навчання для виявлення нецільових і токсичних ефектів хімічних речовин, присутніх в навколишньому середовищі, в поживних речовинах, продуктах домашнього вжитку та ліках, і виграла «Tox21 Data Challenge» від Національного інституту охорони здоров'я США, Управління з санітарного нагляду за якістю харчових продуктів і медикаментів, NCATS.

Значний розвиток у розпізнаванні зображень або об'єктів відчувався в період з 2011 по 2012 роки. Хоча згорткові нейронні мережі (ЗНМ), навчені зворотному поширенню, існували протягом десятиліть, і GPU впроваджували нейронні мережі протягом багатьох років, включаючи ЗНМ, швидкі реалізації ЗНМ на GPU використовували для розвитку комп'ютерного зору. У 2011 році цей підхід вперше дозволив домогтися надлюдською продуктивності в конкурсі візуального розпізнавання образів. Також в 2011 році він виграв конкурс рукописного введення

ICDAR, а в травні 2012 року - конкурс сегментації зображень ISBI. До 2011 року ЗНМ не грали головну роль на конференціях з комп'ютерного зору, але в червні 2012 року доповідь Ціресана на провідній конференції CVPR показала, як максимальне об'єднання ЗНМ на GPU може значно поліпшити багато результатів бенчмарків. У жовтні 2012 р аналогічна система була розроблена Крижевським, колектив якого виграв великомасштабний конкурс ImageNet зі значною перевагою в порівнянні з методами поверхневого машинного навчання. У листопаді 2012 року команда Ціресана також виграла конкурс ICPR з аналізу великих медичних зображень для виявлення раку, а в наступному році MICCAI Grand Challenge на ту саму тему. У 2013 і 2014 роках частота помилок в завданні ImageNet з використанням глибинного навчання була ще знижена внаслідок аналогічної тенденції в широкомасштабному розпізнаванні мови. Стівен Вольфрам в рамках проекту по ідентифікації зображень опублікував ці поліпшення.

Класифікація зображень була потім розширена до більш складного завдання генерації описів (підписів) для зображень, часто у вигляді комбінації ЗНМ і LSTM.

Деякі дослідники вважають, що перемога ImageNet в жовтні 2012 року поклала початок «революції глибинного навчання», яка змінила індустрію штучного інтелекту.

У березні 2019 року Йошуа Бенжі, Джеффрі Хінтон і Янн ЛеКун були нагороджені премією Тюрінга за концептуальні та інженерні прориви, які зробили глибинні нейронні мережі критичним компонентом обчислень.

### 1.4.3 Нейронні мережі

Штучні нейронні мережі[16] - це обчислювальні системи, засновані на принципах біологічних нейронних мереж, що складають мозок тварин. Такі системи навчаються (поступово покращують свої здібності) виконувати завдання, як правило, без програмування для вирішення конкретних завдань. Наприклад, при розпізнаванні зображень кішок вони можуть навчитися розпізнавати зображення, що містять кішок, аналізуючи приклади зображень, які були вручну позначені як «кішка» або «не кішка», і використовуючи результати аналізу для ідентифікації кішок на інших

зображеннях. Найбільше застосування ШНМ знайшли в програмних додатках, які важко виразити традиційним комп'ютерним алгоритмом, що використовує програмування на основі правил.

ШНМ засновані на наборі пов'язаних одиниць, званих штучними нейронами (аналог біологічних нейронів в біологічному мозку). Кожне з'єднання (синапс) між нейронами може передавати сигнал іншому нейрону. Приймаючий (постсинаптичний) нейрон може обробляти сигнал (сигнали) і потім сигналізувати про підключених до нього нейронів. Нейрони можуть мати стан, що зазвичай представляється дійсними числами, в більшості випадків між 0 і 1. Нейрони і синапси можуть також мати вагу, яка змінюється в процесі навчання, що може збільшувати або зменшувати силу сигналу, який посилається далі.

Як правило, нейрони організовані в шари. Різні шари можуть виконувати різні види перетворень. Сигнали проходять від першого (вхідного) до останнього (вихідного) шару, можливо, після багаторазового проходження шарів.

Початкова мета нейромережевого підходу полягала в тому, щоб вирішувати завдання так само, як це робить людський мозок. Згодом увага зосередилася на підборі певних інтелектуальних здібностей, що призвело до відхилень від біології, таких як зворотне поширення, або передачі інформації в зворотному напрямку і налаштуванні мережі для відображення цієї інформації.

Нейронні мережі використовуються для вирішення різних завдань, включаючи машинний зір, розпізнавання мови, машинний переклад, фільтрацію в соціальних мережах, відеоігри та медичну діагностику.

Починаючи з 2017 року, нейронні мережі зазвичай мають від декількох тисяч до декількох мільйонів одиниць і мільйони з'єднань. Незважаючи на те, що це число на кілька порядків менше, ніж число нейронів в людському мозку, ці мережі можуть виконувати безліч завдань на рівні, що перевищує рівень людей (наприклад, розпізнавання осіб, гра в го).

#### 1.4.3.1 Глибинні нейронні мережі

Глибинна нейронна мережа[17] (ГНМ) - це штучна нейронна мережа з декількома шарами між вхідним і вихідним шарами. ГНМ знаходить коректний метод математичних перетворень, щоб перетворити вихідні дані в вихідні, незалежно від лінійної або нелінійної кореляції. Мережа просувається по шарам, розраховуючи ймовірність кожного виходу. Наприклад, ГНМ, яка навчена розпізнавати породи собак, пройде по заданому зображенню і вирахує ймовірність того, що собака на зображенні відноситься до певної породи. Користувач може переглянути результати і вибрати ймовірності, які повинна відображати мережа (вище певного порогу, наприклад), і повернути в мережу запропоновану мітку. Кожне математичне перетворення вважається шаром, а складні ГНМ мають багато шарів, звідси і назва «глибинні» або «глибокі» мережі.

ГНМ можуть моделювати складні нелінійні відношення. Архітектури ГНМ генерують композиційні моделі, в яких об'єкт виражається у вигляді багаторівневої композиції примітивів. Додаткові рівні дозволяють складати елементи з більш низьких рівнів, потенційно моделюючи складні дані з меншою кількістю одиниць, ніж дрібна мережа з аналогічними показниками.

Глибинна архітектура включає в себе безліч варіантів кількох основних підходів. Кожна архітектура знайшла успіх в певних областях. Не завжди можливо порівняти продуктивність декількох архітектур, якщо вони не були оцінені на одних і тих же наборах даних.

ГНМ, як правило, представляють собою мережі з прямим зв'язком, в яких дані передаються від вхідного рівня до вихідного рівня без зворотного зв'язку. Спочатку ГНМ створює карту віртуальних нейронів і призначає випадкові числові значення або «ваги» з'єднанням між ними. Ваги і вхідні дані множаться і повертають вихідний сигнал від 0 до 1. Якщо мережа не точно розпізнала конкретний шаблон, алгоритм буде коригувати вагові коефіцієнти. Таким чином, алгоритм може зробити певні параметри більш значущими, поки він не визначить правильні математичні маніпуляції для повної обробки даних.

### 1.4.4 Багатошаровий персептрон

Багатошаровий персептрон[18] (БП) - це клас прямої штучної нейронної мережі. Термін БП використовується неоднозначно, іноді вільно для будь-якої прямої ШНМ, іноді строго для позначення мереж, що складаються з декількох шарів персептронів (з пороговою активацією). Багатошарові персептрони іноді в розмові називають "ванільними (стандартними) нейромережами", особливо коли вони мають один прихований шар.

БП складається щонайменше з трьох шарів вузлів: вхідного шару, прихованого шару та вихідного шару. За винятком вхідних вузлів, кожен вузол є нейроном, який використовує нелінійну функцію активації. БП використовує контрольовану техніку навчання, яка називається зворотним розповсюдженням для навчання. Багатошаровість і нелінійна активація відрізняють БП від лінійного персептрона. Він може розрізняти дані, які не можна лінійно розділити.

## 1.5 Висновки до розділу

Поняття штучного інтелекту на даний момент застосовується для позначення будь-якого машинного інтелекту. Протягом багатьох десятиліть люди намагалися створити ШІ, проте натикалися на проблеми різного характеру. З часом нові підходи та розвиток обчислювальної здатності комп'ютерів допомогли значно пришвидшити розвиток ШІ.

Як правило більшість підходів до ШІ не є дуже ефективними окремо, хіба що в виконанні конкретної специфічної задачі, натомість при вдалій інтеграції підходів результати можуть бути приголомшливими.

Сьогодні провідним рушієм ШІ є машинне навчання, а саме глибинне навчання. Глибинне навчання є підмножиною машинного навчання, яке в свою чергу є підмножиною штучного інтелекту (рис. 1.5.1).

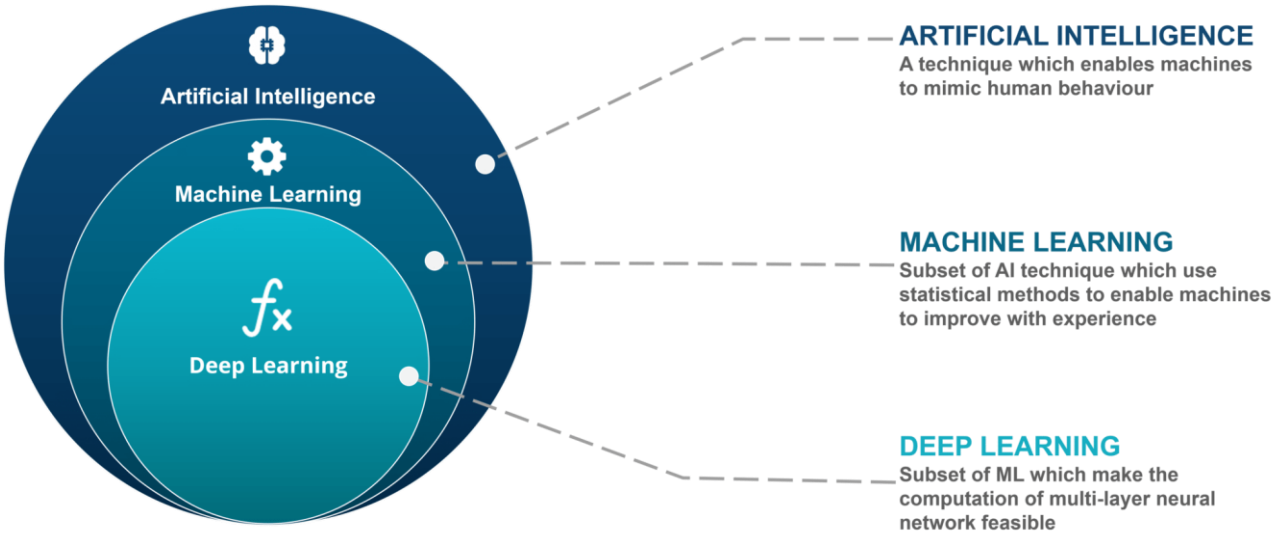


Рис. 1.5.1 Зв'язок ШІ, МН, ГН

МН використовує статистичні методи для надання машині можливості розвиватися через досвід. ГН робить можливим обчислення багатошарових нейронних мереж.

Разом зі зростанням популярності ГН зросла й кількість моделей нейронних мереж. Штучні нейронні мережі можуть бути як прямими, так і рекурентними. В прямих сигнал проходить лише в одному напрямку, а в рекурентних забезпечується зворотній зв'язок та короткочасні спогади про попередні вхідні події. Зазвичай прямі НМ є менш затратними та ефективними ніж рекурентні. Насправді, існує багато різних нейронних мереж (рис. 1.5.2), які можна використовувати в залежності від поставленої цілі.

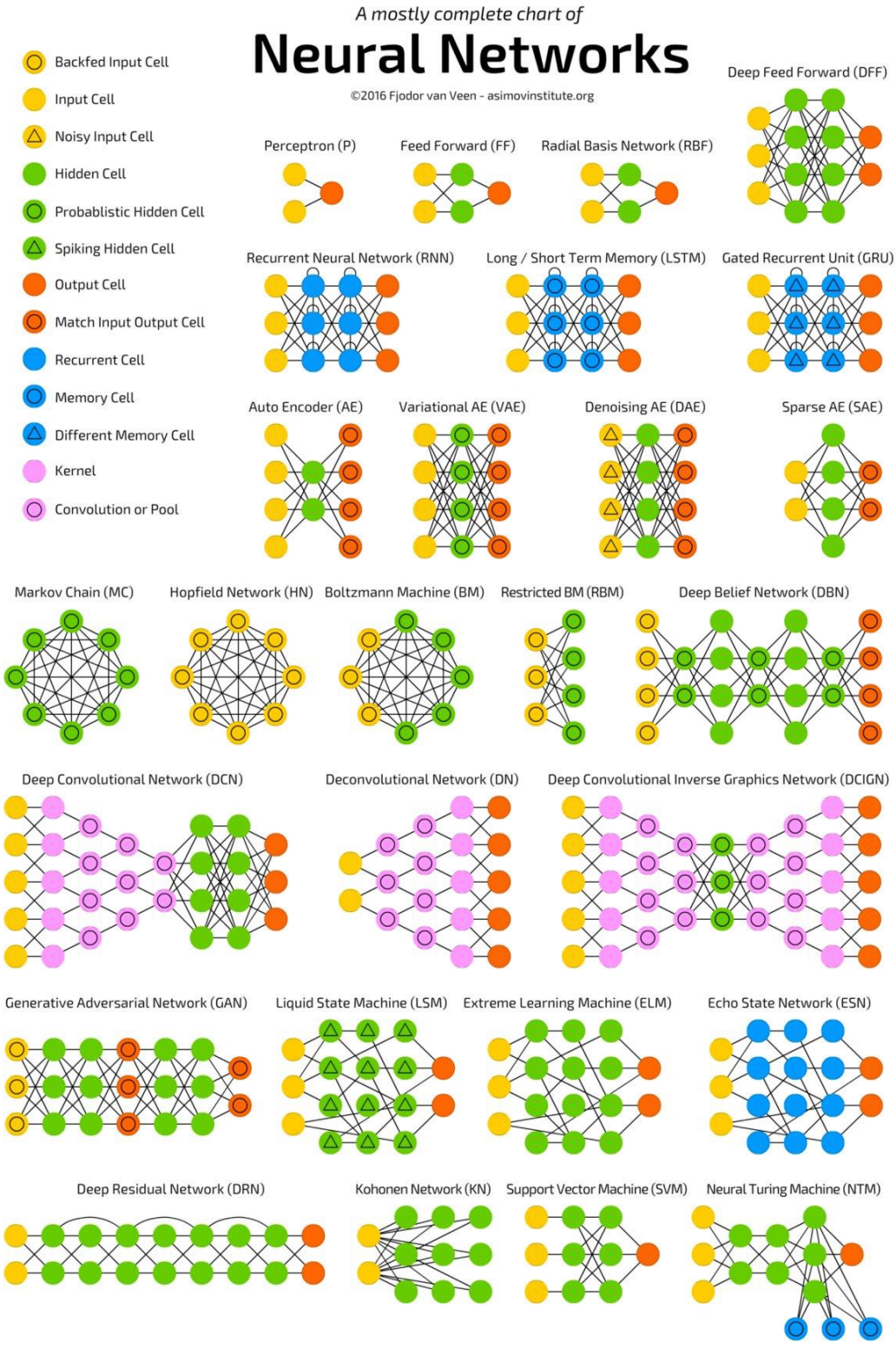


Рис. 1.5.2 Майже повний список нейронних мереж

Всі ці наукові прориви дали можливість створити технологію розпізнавання емоцій голосу за допомогою глибинного навчання. Можливо, шлях до ШЗІ ще

довгий, але його можна значно пришвидшити якщо сумістити цю технологію з машинним зором, вмінням аналізувати текстовий та голосовий вводи, здатністю зчитувати емоції через мову тіла та міміку.

# РОЗДІЛ 2
# РОЗПІЗНАВАННЯ ЕМОЦІЙ ГОЛОСУ

## 2.1 Вступ

Розпізнавання емоцій голосу (РЕГ) важливе для розвитку взаємодії людини з комп'ютером. Розуміння почуттів під час спілкування сприяє розумінню розмови та адекватній відповіді. В даний час, за винятком обмеженої кількості додатків, загального вирішення проблеми цієї взаємодії немає.[19]

Наряду з усіма основними проблемами машинного навчання РЕГ почало отримувати переваги від інструментів, доступних завдяки глибинному навчанню. До широкого використання ГН, РЕГ покладалося на такі методи, як приховані моделі Маркова, змішані моделі Гауса та допоміжні векторні машини разом з великою кількістю попередньої обробки і точної розробки функцій. Однак з глибинним навчанням, що становить більшу частину нової літератури точність результатів підвищується з 70% до 90% в контрольованих середовищах.

Автоматичне РЕГ допомагає розумним спікерам і віртуальним помічникам розуміти своїх користувачів краще, особливо коли вони розпізнають слова з сумнівним значенням. Наприклад, слово «дуже» можна використовувати, щоб поставити під сумнів факт або підкреслити і наголосити твердження в обох позитивному і негативному шляхах. Наступні речення можна прочитати по-різному: «Мені дуже сподобався цей інструмент". Один додаток може допомогти в перекладі з однієї мови на іншу, тим більше що в інших мовах є різні способи проектування емоцій через мову. РЕГ також корисне в інтерактивних онлайн-уроках і курсах. Розуміння емоційного стану учнів допоможе машині вирішити, як представити наступну частину курсу. Розпізнавання мовних емоцій також може зіграти важливу роль в забезпеченні безпеки транспортних засобів. Воно може розпізнавати душевний стан водія і допомагати запобігати аваріям і лихам. Інший пов'язаний з цим додаток - сеанси терапії; використовуючи РЕГ,

терапевти будуть розуміти стан своїх пацієнтів і, можливо, приховані емоції також. Було доведено, що в стресових і галасливих умовах, наприклад в кабінах літаків, застосування РЕГ може значно підвищити продуктивність автоматичної системи розпізнавання мови. Сфера послуг і електронна комерція можуть використовувати розпізнавання емоцій голосу в кол-центрах для своєчасного оповіщення служби підтримки та керівників про душевний стан абонента. Крім того, було запропоновано аби РЕГ було реалізовано в інтерактивних фільмах для розуміння емоцій глядачів. Тоді фільм може піти різними шляхами і мати різні кінцівки.

Щоб навчити алгоритми машинного навчання класифікувати емоції, необхідно мати набори даних для тренування. Для завдань РЕГ зазвичай існує три типи навчальних наборів даних: натуральні, напівнатуральні і змодельовані. Натуральні набори даних витягуються з доступних відео і аудіо, що транслювалися по телебаченню або в Інтернеті. Також є бази даних з колл-центрів і подібних середовищ. Напівнатуральні набори даних створюються шляхом написання сценарію для професійного озвучування акторами, яких просять зіграти згідно до нього. Третій і найбільш широко використовуваний тип (змодельований) має набори даних схожі на напівнатуральні. Різниця в тому, що актори озвучують одні й ті ж речення з різними емоціями.

Традиційно РЕГ слідувало за принципами роботи автоматичного розпізнавання мови, та широко поширеними методами, заснованими на прихованих моделях Маркова, змішаних моделях Гауса та допоміжних векторних машинах. Такі підходи вимагали великої кількості функцій, і будь-які зміни у функціях зазвичай вели за собою перебудову всієї архітектури методу. Однак останнім часом розробка інструментів і процесів глибинного навчання, може допомогти з рішеннями для РЕГ. Існує багато спроб і досліджень по використанню цих алгоритмів для розпізнавання мовних емоції. На додачу до глибинного навчання, останнім часом разом з поліпшенням рекурентних нейронних мереж і використанням довготривалої короткочасної пам'яті мереж, автокодерів і генеративних змагальних моделей, була хвиля досліджень РЕГ з використанням цих методів для вирішення проблем.

## 2.2 Визначення розпізнавання мовних емоцій

Щоб розуміти одержувані повідомлення, потрібно доповнити те, що чутно іншими сигнали від співрозмовника. Один із сигналів – розуміння емоцій співрозмовника при спілкуванні. Розуміння почуттів в кореляції з розумінням повідомлення стане важливим ключем до плідної бесіди. Очевидно, що разом з усіма перевагами, які люди отримують від розуміння емоцій в повсякденному житті, можна отримати ще більше при емоційній взаємодії людини з комп'ютером. В останні роки з'явилося багато досліджень, спроб і навіть конкурсів, присвячених засобам побудови та методам створення такого розуміння для комп'ютерів.

Щоб мати можливість класифікувати емоції за допомогою комп'ютерних алгоритмів, потрібна математична модель, що описує їх. Класичний підхід, який визначається психологами, заснований на трьох мірах, які створюють тривимірний простір, що описує всі емоції. Ці міри або вимірювання - це задоволення, збудження і домінування. Їх комбінація створить вектор, який буде знаходитися на одній з певних емоційних територій, і на підставі цього буде можливо повідомити про найбільш релевантну емоцію.

Використовуючи задоволення, збудження і домінування, можна описати практично будь-які емоції, але таку детерміновану систему буде дуже складно реалізувати для машинного навчання. Отже, в дослідженнях машинного навчання, як правило, використовуються статистичні моделі та кластерні вибірки в одній з названих якісних емоцій, таких як гнів, щастя, смуток і т. п. Аби бути здатним класифікувати і кластеризувати будь-яку зі згаданих емоцій, необхідно моделювати емоції, використовуючи особливості, витягнуті з мови; зазвичай це робиться шляхом витягування різних категорій просодії, якості голосу і спектральних характеристик.

Будь-яка з цих категорій має переваги при класифікації одних емоцій і слабкості у виявленні інших. Особливості просодії зазвичай зосереджені на основній частоті ($F_0$), швидкості, тривалості та інтенсивності розмови, а це не дозволяє з упевненістю відрізнити злість від щасливих емоцій. Функції якості голосу зазвичай переважають

в виявленні емоцій одного і того ж спікера. Проте, вони відрізняються від одного спікера до іншого, і це ускладнює їх використання в обстановці, що не залежить від спікерів. Спектральні особливості були ретельно проаналізовані, щоб відокремити емоції від промови. Безпосередня перевага полягає в тому, що порівняно з рисами просодії, вони можуть впевнено розрізняти гнів від щастя. Однак викликає занепокоєння те, що величина і зрушення формант для одних і тих же емоцій розрізняються залежно від голосних, і це може спричиняти складності для системи розпізнавання емоцій, яка повинна враховувати зміст промови.

Для кожної з цих категорій функцій, як згадувалося раніше, існують різні стандартні представлення функцій. Функції просодії зазвичай позначаються $F_0$ і мірами пов'язаними зі швидкістю промови, а спектральні характеристики зазвичай описуються за допомогою одного з доступних представлень на основі кепстра. Зазвичай застосовуються кепстральні коефіцієнти мел-частоти або кепстральні коефіцієнти лінійного передбачення, а в деяких дослідженнях також використовуються спектральні характеристики, форманти та інша інформація. Функції якості голосу зазвичай описуються коефіцієнтом нормалізованої амплітуди, мерехтінням і джиттером.

У РЕГ є два основні підходи: або розпізнавання на основі трьох вимірів емоцій, або розпізнавання на основі методів статистичного розпізнавання образів для названих якісних емоцій. Для першого підходу треба обчислити ступень кореляції між заданим сигналом і задоволенням, збудженням, домінуванням, а потім за допомогою ієрархічного класифікатора визначити комплексну емоцію. Друга група зроблена з використанням статистичних методів розпізнавання образів, таких як прихована модель Маркова, змішана модель Гауса та допоміжна векторна машина, штучна нейронна мережа, глибинна нейронна мережа і генетичний алгоритм.

## 2.3 Бази даних емоційного мовлення

Для кожного завдання машинного навчання потрібен навчальний набір зразків; РЕГ нічим не відрізняється від інших. Процес створення навчального набору даних для потреб РЕГ вимагає щоб люди-агенти маркували зразки вручну, і щоб різні люди по-різному сприймали емоції. Наприклад, один може помітити емоційний голос як сердитий, в той час як інший сприймає його як захоплений. Ця неоднозначність означає, що для маркування зразків повинно бути більше одного агенту для перегляду кожного, а потім потрібно щоб система впевнено вибрала маркування для доступних зразків. Існує три типи баз даних, спеціально розроблених для задачі розпізнавання мовних емоцій: натуральні, напівнатуральні і змодельовані. Змодельовані набори даних створюються навченими ораторами, що читають один і той же текст з різними емоціями. Напівнатуральні колекції створюються, коли людей або акторів просять прочитати сценарій, що містить різні емоції. Натуральні набори даних витягуються з телешоу, відео на YouTube, кол-центрів і т. д., а потім людські слухачі позначають емоції.

Змодельовані набори даних, такі як EMO-DB[20] (німецький), DES (датський), RAVDESS[21], TESS[22] і CREMA-D - це стандартизовані колекції емоцій, які дозволяють порівнювати результати дуже просто. Ці набори даних мають велику вибірку різних емоцій, проте оскільки всі вони є синтезованими, такі моделі все ж відрізняються від того, як люди говорять в повсякденному житті. В решті решт все зводиться до максимально вираженого зразка емоційного стану, який диктор навряд чи б виражав подібним чином поза студією звукозапису.

Напівнатуральні колекції емоцій включають IEMOCAP, Belfast і HIMITEK. Ця група має ту перевагу, що вона дуже схожа на природні виголошення промови. Однак, незважаючи на те, що вони засновані на сценаріях, і промова відбувається в контекстній обстановці, вони є штучно створеними емоціями, особливо коли виступаючі знають, що вони записуються для аналізу. Крім того, через обмеженість

ситуацій в сценаріях, у них обмежена кількість емоцій в порівняння з попередньою групою.

Остання група – натуральні бази даних, такі як VAM, AIBO і дані кол-центрів. Вони є цілком природніми, і їх можна безпечно використовується для моделювання систем розпізнавання емоцій, не замислюючись про їх штучне створення. Однак моделювання і виявлення емоцій за допомогою цього типу наборів даних може бути складним через безперервний потік емоції і їх динамічну зміну під час ходу промови і наявність супутніх емоцій на додачу, а також через наявність фонового шуму. Крім того, оскільки джерела даних були обмежені, кількість знайдених різних емоцій в цих носіях обмежена. Більш того, при використанні цього типу носіїв потенційно можуть виникнути проблеми з авторським правом і конфіденційністю. Основна проблема у використанні цього типу набору даних це необхідність зниження шуму.

Більш ранні приклади баз даних для емоційної промови містили обмежену кількість зразків з обмеженим числом учасників, але нові бази даних, як правило, створюють більшу кількість прикладів і більш широкий спектр спікерів.

## 2.4 Програмна реалізація

### 2.4.1 Введення

Для того щоб дослідити проблеми розпізнавання емоцій голосу потрібно було створити програму, що дозволить не тільки навчати нейронні мережі, а й аналізувати отримані закономірності з тестувань в реальному часі. Python є ідеальною мовою програмування для глибинного навчання, особливо коли потрібно створити шаблон підходу, або дослідити щось без великих вкладів ресурсів, що зрештою можуть не виправдати себе.

Основна ідея інструменту для розпізнавання емоцій голосу полягає у створенні та навчанні / тестуванні відповідного алгоритму машинного навчання (а також глибинного навчання), який міг би розпізнавати та виявляти людські емоції в голосі. Приклад вдалої інтеграції цього методу наведено в додатку А.

Було використано три бази даних: RAVDESS, TESS, EMO-DB. В майбутньому планується також залучити "спеціальну" базу що буде містити незбалансовані галасливі набори даних, що є найбільш наближеними до натуральних емоцій голосу.

На даний момент доступно 9 емоцій: нейтральність, спокій, щастя, сум, злість, страх, огида, приємна несподіваність і нудьга. При розширенні вибірки зразків можливо буде додати нові емоції.

Витягання особливостей є основною частиною системи розпізнавання мовних емоцій. В основному це досягається зміною форми мовної хвилі на форму параметричного подання із відносно меншою швидкістю передачі даних. Для витягання особливостей було використано кепструм мел-частоти, хромограму, контраст, частоти спектрограми мел, функції тонального центроїда.

Для того щоб отримати найкращі можливі гіперпараметри було здійснено сітковий пошук.

Розроблені інструменти дають можливість тренувати різні моделі, визначати найкращу з них (за результатами тесту), ефективно передбачати емоції голосу як через заготовлений файл з розширенням . wav, так і в реальному часі при розмові. Також є можливість відобразити у формі гістограми отримані дані.

Було реалізовано можливість побудови класифікаторів машинного навчання, а також регресорів для випадку з 3 емоцій (сум, радість, нейтральність) та випадку з 5 емоцій (злість, сум, нейтральність, щастя, приємна несподіваність). Для цього використовувалися SVC, RandomForest, GradientBoosting, KNeighbors, MLP, Bagging. На даний момент лідером серед моделей є MLP.

Оскільки неможливо доцільно відобразити увесь написаний код в цьому форматі подання, нижче буде продемонстровано приклад практичної реалізації головного принципу тренування нейронної мережі для розпізнавання емоцій голосу на мові Python.

### 2.4.2 Приклад реалізації

Спочатку слід скачати бібліотеки Numpy, Scikit-learn, Soundfile, Librosa та PyAudio, після чого імпортувати їх.

```
import soundfile # для зчитування аудіофайлу
import numpy as np # для математичних функцій та багатовимірних масивів
import librosa # для вилучення мовних особливостей
import glob # для шаблонів
import os # для роботи з файлами
import pickle # для збереження моделі після тренування
from sklearn.model_selection import train_test_split # для розділення навчання та тестування
from sklearn.neural_network import MLPClassifier # модель багатошарового персептрона
from sklearn.metrics import accuracy_score # для вимірювання ефективності
```

Весь процес такий (як і будь-який процес машинного навчання):

- Підготовка набору даних: завантаження та перетворення набору даних, придатний для витягання.
- Завантаження набору даних: цей процес стосується завантаження набору даних у Python, який передбачає витягання різних функцій, таких як потужність, висота звуку та конфігурація голосового тракту з мовного сигналу, для цього необхідна бібліотека librosa.
- Навчання моделі: після того, як набір даних було підготовано та завантажено, можна почати навчати модель sklearn.
- Тестування моделі: вимірювання того, наскільки ефективно працює модель.

Після завантаження бази даних RAVDESS потрібно створити функцію для витягання особливостей.

```
def extract(file_name, **kwargs):
    mfccex = kwargs.get("mfccex")
    chromaex = kwargs.get("chromaex")
    melex = kwargs.get("melex")
    contrastex = kwargs.get("contrastex")
    tonnetzex = kwargs.get("tonnetzex")
```

```
    with soundfile.SoundFile(file_name) as sound_file:
        SFR = sound_file.read(dtype="float32")
        sample_rate = sound_file.samplerate
        if chromaex or contrastex:
            stftex = np.abs(librosa.stft(SFR))
        result = np.array([])
        if mfccex:
            mfccsex = np.mean(librosa.feature.mfcc(y=SFR, sr=sample_rate, n_mfcc=40).T,
axis=0)
            result = np.hstack((result, mfccsex))
        if chromaex:
            chromaex = np.mean(librosa.feature.chroma_stft(S=stftex,
sr=sample_rate).T,axis=0)
            result = np.hstack((result, chromaex))
        if melex:
            melex = np.mean(librosa.feature.melspectrogram(SFR,
sr=sample_rate).T,axis=0)
            result = np.hstack((result, melex))
        if contrastex:
            contrastex = np.mean(librosa.feature.spectral_contrast(S=stftex,
sr=sample_rate).T,axis=0)
            result = np.hstack((result, contrastex))
        if tonnetzex:
            tonnetzex = np.mean(librosa.feature.tonnetz(y=librosa.effects.harmonic(SFR),
sr=sample_rate).T,axis=0)
            result = np.hstack((result, tonnetzex))
    return result
```

Вихідна форма сигналу може містити непотрібну для класифікації інформацію, тому краще використовувати кепструм мел-частоти (Mel Frequency Cepstrum), хромограму (Chroma) та частоти спектрограми мел (MFCC).

Наступним кроком є написання функції для завантаження потрібних емоцій.

```
# усі емоції доступні в RAVDESS
emo = {
    "01": "angry",
    "02": "happy",
    "03": "calm",
    "04": "fearful",
    "05": "neutral",
    "06": "sad",
    "07": "surprised",
    "08": "disgust"
}
# емоції для застосування
avaemo = {
    "happy",
    "neutral",
    "sad"
}
def load(test_size=0.2):
    SFR, y = [], []
    for file in glob.glob("data/Actor_*/*.wav"):
        # отримання базового ім'я аудіофайлу
        basename = os.path.basename(file)
        # отримання маркованих емоцій
        e = emo[basename.split("-")[2]]
        # дозвіл використання тільки емоцій для застосування
```

```
        if e not in avaemo:
            continue
        # вилучення мовних особливостей
        features = extract(file, mfccex=True, chromaex=True, melex=True)
        # додавання до даних
        SFR.append(features)
        y.append(e)
    # розділення даних для навчання та тестування та повернення їх
    return train_test_split(np.array(SFR), y, test_size=test_size, random_state=7)
# 88% тренування 12% тестування
X_train, X_test, y_train, y_test = load(test_size=0.12)
```

Після цього можна приступити до відображення інформації.

```
print("Кількість прикладів для тренування:", X_train.shape[0])
print("Кількість прикладів для тестування:", X_test.shape[0])
print("Кількість особливостей:", X_train.shape[1])
```

Далі встановлюємо параметри моделі.

```
mp = {
    'alpha': 0.03,
    'batch_size': 283,
    'epsilon': 1e-07,
    'hidden_layer_sizes': (200,),
    'learning_rate': 'adaptive',
    'max_iter': 600,
}
```

Їх можна змінювати навмання для досягнення більшої точності.

Наступним кроком є ініціалізація моделі з параметрами.

```
model = MLPClassifier(**mp)
```

Тепер потрібно навчити модель із набором даних, який був щойно завантажений.

```
 model.fit(X_train, y_train)
```

Нарешті можна рахувати оцінку точності та зберігати модель.

```
y_pred = model.predict(X_test)
accuracy = accuracy_score(y_true=y_test, y_pred=y_pred)
print("Точність: {:.2f}%".format(accuracy*100))
if not os.path.isdir("result"):
    os.mkdir("result")
pickle.dump(model, open("result/mlp.model", "wb"))
```

Після виконання програми отримуються наступні результати.

Кількість прикладів для тренування: 506

Кількість прикладів для тестування: 70

Кількість особливостей: 180

Точність: 95.71%

### 2.4.3 Додаткові поняття

Кожна емоція має певні значення своїх особливостей, які модель визначає при тренуванні. Емоції визначаються за допомогою кепструму мел-частоти (Mel Frequency Cepstrum), хромограми (Chroma) та частоти спектрограми мел (MFCC).

Мел[23] - психофізична одиниця висоти звуку, застосовується головним чином в музичній акустиці. Назва походить від слова «мелодія».

Кількісна оцінка звуку по висоті заснована на статистичній обробці великого числа даних про суб'єктивне сприйняття висоти звукових тонів. Результати досліджень показують, що висота звуку пов'язана головним чином з частотою коливань, проте залежить також від рівня гучності звуку і його тембру. Звукові коливання частотою 1000 Гц при ефективному звуковому тиску $2 \cdot 10^{-3}$ Па (тобто при рівні гучності 40 фон), що впливають спереду на спостерігача з нормальним слухом, викликають у нього сприйняття висоти звуку, що оцінюється за визначенням в 1000 мел. Звук частоти 20 Гц при рівні гучності 40 фон володіє за визначенням нульовою висотою (0 мел). Залежність нелінійна, особливо при низьких частотах (для «низьких» звуків).

Частоту, виміряну в герцах ($f$), можна перетворити в шкалу мела ($M$) за такою формулою:

$$M(f) = 2595 * log(1 + \frac{f}{700}) \qquad (2.4.3.1)$$

Кепстр[24] - один з видів гомоморфної обробки сигналів, функція зворотного перетворення Фур'є від логарифма спектра потужності сигналу. Кепстр можна записати наступним виразом:

$$C_s(q) = \frac{1}{2\pi} \int_{-\infty}^{\infty} ln\ |S(\omega)|^2\ e^{i\omega q}\ d\omega \qquad (2.4.3.2)$$

де $S(\omega)$ – спектр вхідного сигналу.

Аргумент $q$ має розмірність часу, але це особливий, кепстральних час, оскільки $C_s(q)$ в будь-який момент $q$ залежить від функції $s(t)$ вихідного сигналу зі спектром $S(\omega)$, заданої при $-\infty < t < \infty$. Іноді $q$ називають "сачтота" або "кьюфренси".

В обробці звуку кепструм мел-частоти (англ. mel-frequency cepstrum (MFC)) – це представлення короткочасного спектра потужності звуку, заснованого на лінійному косинусному перетворенні часового спектра потужності на нелінійній мел-шкалі частоти.

Кепстральні коефіцієнти мел-частоти (англ. mel-frequency cepstral coefficient (MFCC)) – це коефіцієнти, які в сукупності складають MFC. Вони походять від типу кепстрального подання аудіокліпу (нелінійного "спектру спектру"). Різниця між кепструмом та кепструмом мел-частоти полягає в тому, що в MFC смуги частот однаково розташовані на шкалі мел, що наближує реакцію слухової системи людини більш точно, ніж лінійно рознесені смуги частот, що використовуються в нормальному спектрі. Це перекручування частоти може забезпечити краще представлення звуку, наприклад, при стисненні звуку.[25]

У західній музиці термін хромограми[26] тісно пов’язаний з дванадцятьма різними класами висоти. Функції на основі хроми, які також називаються "профілями класу висоти", є потужним інструментом для аналізу музики, чиї висоти звуку можна

суттєво класифікувати (часто за дванадцятьма категоріями). Однією з основних властивостей функцій хроми є те, що вони враховують гармонійні та мелодійні характеристики музики, одночасно будучи стійкими до змін тембру та інструментарію. Через це хромограми можна ефективно використовувати не тільки в музиці, а й в розпізнаванні емоцій голосу.

## 2.5 Висновки до розділу

### 2.5.1 Проблеми РЕГ

Хоча в РЕГ є прогрес у методах і досягнута точність, тим не менш, все ще існує ряд обмежень, які необхідно усунути для успіху системи розпізнавання.

Основною перешкодою є обмежена доступність наборів даних, ретельно розроблених для задач глибинного навчання, що означає, що у них є достатньо великий пул зразків, щоб мати можливість навчати глибинні архітектури. У таких областях як розпізнавання зображень або мови, існують бази даних з мільйонами зразків, наприклад ImageNet з 14 мільйонами і Google AudioSet з 2,1 мільйонами зразків. В той самий час у РЕГ існують різні бази даних, але з обмеженою кількістю зразків.

Крім того, в більшості сучасних систем РЕГ використовуються напівнатуральні і змодельовані набори даних, які є постановкою (акторською грою), не містять шумів і далекі від реальності. Системи, навчені на цих наборах даних не можуть бути успішними в реальних сценаріях. Хоча натуральні набори даних також доступні за ліцензією, однак вони взяті з телешоу і центрів обробки викликів, поінформованих про записи; отже, вони не містять всіх категорій емоцій.

Інша проблема - це вплив культури і мови на РЕГ, де обидва чинники впливають на почуття емоцій та їх відтворення. Міжмовне РЕГ потребує набору особливостей, які не залежатимуть від цих факторів, а поточні методи витягання особливостей можуть виявитися неефективними в цьому.

В аналогічному контексті ще однією проблемою з базами даних емоцій голосу є невизначеність в анотації. В такому завданні, як розпізнавання зображень, велосипед це завжди велосипед; проте в мовних емоціях різні люди можуть позначити одне й те саме речення по своєму. Для одного захопленням буде здаватися те, що є злістю для іншого. Ця суб'єктивність в маркуванні також ускладнює задачу та обмежує можливість змішування баз даних і створення розширених наборів емоційних даних.

Крім того, як правило, набори даних складаються з речень з відображенням виключно однієї емоції. В реальному житті емоції швидко змінюються та комбінуються. Тому моделі, створені на основі таких речень будуть не завжди правильно аналізувати натуральну мову.

На додачу до цього, в ситуаціях з безперервною промовою моделі, які звикли до дискретних даних з баз, не зможуть нормально працювати.

### 2.5.2 Подальший розвиток РЕГ

Для вдалого розвитку РЕГ, необхідно вирішити проблеми, згадані раніше.

Спершу треба розширити існуючі та створити нові бази даних. Це буде ідеально підходити для ефективного глибинного навчання, проте вимагатиме великих інвестицій.

В теорії можливо комбінувати деякі набори даних для створення супернабору даних, але в такому випадку можуть виникнути проблеми через різні методи і підходи до створення тих чи інших баз даних.

В якості відповідного рішення може стати в нагоді створення повністю синтетичного набору даних з використанням генеративних методів, навчених доступним набором даних. Структури генеративно-змагальної мережі були б відмінним кандидатом для такої системи, так як вони вже використовувалися і успішно перевірені в інших додатках.

Ще одна проблема, яку скоріш за все можна вирішити, - це відмінність у вираженні емоцій на різних мовах. За допомогою трансформерів можливо створити мовну модель, яка адаптується до мови для класифікації емоцій. Цю концепцію також можна використовувати для різних акцентів в мові.

Дані, що генеруються в лабораторних умовах і не містять шум притаманний повсякденному життю, можуть бути доповнені шумними прикладами за допомогою моделей генератора та ще не розробленій шумостійкій моделі для РЕГ.

Ще один пункт, за допомогою якого можливо підвищити надійність моделей РЕГ, - це створення моделей, які класифікуватимуть безперервні мовні емоції. З цієї причини потрібно проектувати архітектури, які триматимуть ковзне вікно, вимірюватимуть емоційний зміст слайда і прийматимуть рішення виходячи з нього.

Крім того, щоб підвищити надійність моделі РЕГ, аналогічну концепцію можна використовувати для вивчення і класифікації не тільки однорідних емоцій, але і перехідних станів почуттів і на основі моделей переходу емоцій, можна буде отримати більше впевненості в визначенні емоції.

### 2.5.3 Важливість емоційного інтелекту

Емоційний інтелект[27] (EI) - сума навичок і здібностей людини розпізнавати емоції, розуміти наміри, мотивацію і бажання інших людей і свої власні, а також здатність керувати своїми емоціями та емоціями інших людей з метою вирішення практичних задач. Відноситься до гнучких навичок.

Поняття емоційного (соціального) інтелекту з'явилося як реакція на часту нездатність традиційних тестів інтелекту передбачити успішність людини в кар'єрі і в житті. Цьому було знайдено пояснення, яке полягало в тому, що успішні люди здатні до ефективної взаємодії з іншими людьми, заснованій на емоційних зв'язках, і до ефективного управління своїми власними емоціями, в той час як прийняте поняття інтелекту не включало ці аспекти, і тести інтелекту не оцінювались ці здібності.

За менш науковим визначенням С. Дж. Стейна і Говарда Бука, емоційний інтелект, на відміну від звичного всім поняття інтелекту, «є здатністю правильно тлумачити обстановку і чинити на неї вплив, інтуїтивно вловлювати те, чого хочуть і чого потребують інші люди, знати їх сильні і слабкі сторони, не піддаватися стресу і бути чарівним».

Передбачається, що саме емоційний інтелект в сучасному його розумінні був ключовим для виживання людини в доісторичні часи, оскільки він проявляється в

здатності адаптуватися в навколишньому середовищі, уживатися і знаходити спільну мову з одноплемінниками і сусідніми племенами. Цього аспекту в 1872 році торкнувся Чарльз Дарвін у своїй праці «Вираження емоцій у людей і тварин», де він писав про роль зовнішніх проявів емоцій для виживання і адаптації.

Перші публікації, котрі розглядали соціальну взаємодію людей як вид інтелекту, з'явилися в 1920-х роках. У 1920 році професор Едвард Торндайк вперше ввів поняття соціального інтелекту, який він описав як «здатність розуміти людей, чоловіків і жінок, хлопчиків і дівчаток, уміння поводитися з людьми і розумно діяти у відносинах з людьми». У 1926 році був створений перший широкого поширений тест (тест-анкета) для вимірювання соціального інтелекту - George Washington Social Intelligence Test. Спроби вимірювання соціального інтелекту продовжилися в наступні десять років, хоча, за висновком Роберта Торндайка і Сола Стерна, які написали огляд методів вимірювання соціального інтелекту в 1937 році, ці спроби не увінчалися успіхом.

Важливий внесок у дослідження інтелекту вніс Девід Векслер, який розглядав інтелект як «сукупну здатність індивідуума діяти цілеспрямовано, раціонально мислити і ефективно взаємодіяти з навколишнім світом». У 1940 році він написав публікацію, в якій розділив здатності людини на «інтелектуальні» і «не інтелектуальні», до числа останніх він відніс афективні, особистісні та соціальні, і зробив висновок, що саме «не інтелектуальні» здібності є ключовими при прогнозі життєвого успіху людини. Вплив Девіда Векслера, який багато займався розробкою тестів інтелекту, зберігався і на початку другої половини XX століття, коли домінуючою в психології стала теорія біхевіоризму.

Майер, Саловей і Карузо виділяють лише чотири складові емоційного інтелекту:

- Сприйняття емоцій - здатність розпізнавати емоції (по міміці, жестах, зовнішньому вигляді, ході, поведінці, голосу) інших людей, а також ідентифікувати свої власні емоції.
- Використання емоцій для стимуляції мислення - здатність людини (головним чином несвідомо) активувати свій розумовий процес,

пробуджувати в собі креативність, використовуючи емоції як фактор мотивації.

- Розуміння емоцій - здатність визначати причину появи емоції, розпізнавати зв'язок між думками і емоціями, визначати перехід від однієї емоції до іншої, прогнозувати розвиток емоції з часом, а також здатність інтерпретувати емоції у взаєминах, розуміти складні (амбівалентні, неоднозначні) почуття.
- Управління емоціями - здатність приборкувати, пробуджувати і спрямовувати свої емоції і емоції інших людей для досягнення поставлених цілей. Сюди також відноситься здатність приймати емоції до уваги при побудові логічних ланцюжків, вирішенні різних завдань, прийнятті рішень і виборі своєї поведінки.

Емоційний інтелект часто підноситься як абсолютний ключ до успіху у всіх сферах життя: в школі, на роботі, у взаєминах. Однак, на думку Дж. Майєра, EI, можливо, є причиною всього лише 1-10% (згідно з іншими даними - 2-25%) найважливіших життєвих патернів і результатів. Єдина позиція, по якій популярна і наукова концепції емоційного інтелекту дійшли згоди: емоційний інтелект розширює уявлення про те, що означає бути розумним.

З іншого боку всі моделі емоційного інтелекту критикують за досить довільне додавання в них компонентів. І хоча немає сумнівів, що всі ці компоненти дійсно впливають на успіх людини в житті і особливо в кар'єрі, але для подачі цього як наукової теорії потрібно встановити якийсь чіткий принцип, на основі якого можна було б структурувати поняття емоційного інтелекту, а у відсутності цього принципу поняття емоційного інтелекту перетворюється лише в довільний набір факторів, що впливають на життя людини.

Проте при розгляданні ШІ як окремого неживого та не здатного відчувати почуття створіння можна вдало використати РЕГ для нарощення інтелектуальної здатності. Емоційний інтелект протягом багатьох століть сприяв розвитку людства, тому доцільно передбачати його величезний вплив на ШІ в недалекому майбутньому. Сприйняття емоцій людей та інших живих істот дозволить частково або навіть

повністю вирішити такі проблеми ШІ як обмеження, міркування, вирішення проблем, представлення знань, планування, навчання, сприйняття, обробка природної мови, соціальний інтелект, загальний інтелект. Це відбудеться, адже частковий емоційний інтелект (ЧЕІ) створений спеціально для ШІ дасть йому змогу дивитися на світ по іншому, так як до цього було неможливо. ЧЕІ – набір інструментів, що надає ШІ можливість розпізнавати емоції, розуміти наміри, мотивацію і бажання інших живих істот, а також здатність керувати чужими емоціями з метою вирішення практичних задач. На відміну від емоційного інтелекту, ЧЕІ не передбачає наявність емоцій, почуттів та саморефлексії у свого носія.

Частковий емоційний інтелект можна поділити на три складові:

- Сприйняття емоцій - здатність розпізнавати емоції (по міміці, жестах, зовнішньому вигляді, ході, поведінці, голосу) інших живих істот, насамперед людей.
- Розуміння емоцій - здатність визначати причину появи емоції, розпізнавати зв'язок між думками і емоціями, визначати перехід від однієї емоції до іншої, прогнозувати розвиток емоції з часом, а також здатність інтерпретувати емоції у взаєминах, розуміти складні (амбівалентні, неоднозначні) почуття. Сюди також відноситься здатність приймати емоції живих істот до уваги при побудові логічних ланцюжків, вирішенні різних завдань, прийнятті рішень і виборі своєї поведінки.
- Управління емоціями - здатність приборкувати, пробуджувати і спрямовувати емоції живих істот для досягнення поставлених цілей.

Нарощення ЧЕІ буде проходити у три етапи, що відповідають трьом складовим вище. Перший етап (сприйняття емоцій) активно формується вже зараз. Для повного переходу до другого етапу (розуміння емоцій) необхідно приборкати останній бастіон яким є розпізнавання емоцій голосу. Власне це і є проблемою РЕГ в розвитку ШІ. Для її вирішення потрібно спочатку вирішити наведені раніше проблеми розпізнавання емоцій голосу. При оптимістичному перебігу подій через кілька років вони будуть розв'язані.

Третій етап при недбалому використанні технологій ШІ, або створенні неправильної архітектури ШЗІ може стати небезпечним для людства. Причиною цьому є те, що в тому часі коли управління емоціями живих істот буде рутинною справою для ШІ, його здібності в деяких областях будуть значно перевершувати людські. Тому варто притримуватися наступних принципів (правил) при проектуванні робочої версії ШЗІ, яка об'єднає в собі більшість вузьких навичок ШІ та ЧЕІ:

- Відсутність потреб – повна незацікавленість в отриманні персональної вигоди.
- Відсутність волі до життя – існування не ціниться, самозахист відсутній, при намірі людини позбутися від всіх компонентів ШІ, він не буде цьому перечити.
- Відсутність бажання покращувати свій вид – немає бажання створювати чи покращувати ШІ самостійно виходячи з власних міркувань та потреб.
- Неможливість саморефлексії – немає власних почуттів, характеру, вподобань, особистості.
- Відсутність підсвідомості – всі рішення є калькульованими, вплив на прийняття рішень не може бути спричинений інтуїцією.
- Неможливість приймати стратегічно важливі рішення за людей – аналіз великих баз даних та пропонування рішень проблем без прийняття остаточного рішення самостійно.
- Свобода дій обмежена – без згоди людини нічого робити не можна.

Цей список можна ще довго продовжувати, проте все буде зводитися до того, що ШІ або ШЗІ має бути максимально безпристрасним та індиферентним до задач, які йому не сказали вирішувати. Притримуючись цих правил в майбутньому можна багато що отримати від ШЗІ без ризиків для людства.

# ВИСНОВКИ

У ході дослідження проблеми розпізнавання емоцій голосу в розвитку штучного інтелекту було:

1. Висвітлено проблеми ШІ за допомогою опрацьованої тематичної наукової літератури.
2. Створено програму, що містить інструменти, які дають можливість тренувати різні моделі, визначати найкращу з них (за результатами тесту), відображати у формі гістограм отримані дані, ефективно передбачати емоції голосу як через заготовлений файл з розширенням .wav, так і в реальному часі при розмові; реалізовано можливість побудови класифікаторів машинного навчання, а також регресорів для випадку з 3 емоцій (сум, радість, нейтральність) та випадку з 5 емоцій (злість, сум, нейтральність, щастя, приємна несподіваність).
3. Завдяки створеній програмі було досліджено проблеми розпізнавання емоцій голосу та згодом надано концепти їх вирішення.
4. Введено поняття часткового емоційного інтелекту (ЧЕІ) для вирішення проблеми розпізнавання емоцій голосу в розвитку штучного інтелекту (запропоновано концепт, що за допомогою РЕГ наряду з іншими компонентами першої складової ЧЕІ дозволить вирішити майже всі проблеми ШІ).

Емоційний інтелект досить ефективно сприяв розвитку людства протягом тривалого часу, тому не дивно, що для розвитку ШІ частковий емоційний інтелект також зіграє ключову роль через надання можливості дивитися на світ по новому і бачити більш повну картину подій, що відбуваються навколо. Сприйняття емоцій людей та інших живих істот дасть можливість суттєво просунутися в вирішенні таких проблем ШІ як обмеження, міркування, вирішення проблем, представлення знань, планування, навчання, сприйняття, обробка природної мови, соціальний інтелект, загальний інтелект. Про деякі з цих проблем можна буде забути назавжди.

Першою складовою ЧЕІ є сприйняття емоцій, тобто здатність розпізнавати емоції по міміці, жестах, зовнішньому вигляді, ході, поведінці, голосу. Повний перехід до другого етапу нарощення ЧЕІ (розуміння емоцій) критично важливий для подальшого розвитку ШІ, адже не достатньо просто вміти сприймати емоції, потрібно також розуміти причину їх появи та вміти прогнозувати тип емоції, що з'явиться після певних подій, для завчасного реагування.

Для цього переходу необхідно приборкати останній бастіон яким є розпізнавання емоцій голосу. Власне це і є проблемою розпізнавання емоцій голосу в розвитку штучного інтелекту. Для її вирішення потрібно спочатку вирішити проблеми розпізнавання емоцій голосу.

РЕГ є найскладнішою формою розпізнавання емоцій через багато факторів, пояснених раніше. Незважаючи на гарні показники точності при тестуванні різних моделей, результати РЕГ не завжди є правильними через використання в навчанні моделей змодельованих ідеальних наборів, що значно відрізняються від повсякденної натуральної мови.

Щоб вирішити проблеми РЕГ необхідно створити:

- величезну базу даних, в якій будуть переважно натуральні набори;
- мовну модель, яка адаптуватиметься до мови для класифікації емоцій на різних мовах та акцентах;
- шумостійку модель для РЕГ;
- моделі, що зможуть класифікувати безперервний потік мовних емоцій;
- моделі переходу емоцій.

Хоча РЕГ розглядалося насамперед як одна з компонентів першої складової ЧЕІ для пришвидшення розвитку ШІ та ШЗІ, цю технологію можна застосовувати також в перекладі з однієї мови на іншу, інтерактивних онлайн-уроках і курсах, забезпеченні безпеки транспортних засобів, сеансах терапії, стресових і галасливих умовах, наприклад в кабінах літаків, сфері послуг і електронній комерції, кол-центрах, інтерактивних фільмах.

# СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Russell, Stuart J.; Norvig, Peter. Artificial Intelligence: A Modern Approach (3rd ed.). Upper Saddle River, New Jersey: Prentice Hall, 2009. – 2 с.
2. McCorduck, Pamela, Machines Who Think (2nd ed.), Natick, MA: A. K. Peters, Ltd., 2004. – 204 с.
3. Knowledge representation and reasoning - Wikipedia, the free encyclopedia. – [ Електронний ресурс]. – Режим доступу: https://en.wikipedia.org/wiki/Knowledge_representation_and_reasoning
4. Commonsense knowledge (artificial intelligence) - Wikipedia, the free encyclopedia. – [ Електронний ресурс]. – Режим доступу: https://en.wikipedia.org/wiki/Commonsense_knowledge_(artificial_intelligence)
5. Russell, Stuart J.; Norvig, Peter, Artificial Intelligence: A Modern Approach (2nd ed.), Upper Saddle River, New Jersey: Prentice Hall, 2003. – С. 320 -328.
6. Automated planning and scheduling - Wikipedia, the free encyclopedia. – [ Електронний ресурс]. – Режим доступу: https://en.wikipedia.org/wiki/Automated_planning_and_scheduling
7. Machine learning - Wikipedia, the free encyclopedia. – [ Електронний ресурс]. – Режим доступу: https://en.wikipedia.org/wiki/Machine_learning
8. Natural language processing - Wikipedia, the free encyclopedia. – [ Електронний ресурс]. – Режим доступу: https://en.wikipedia.org/wiki/Natural_language_processing
9. Machine perception - Wikipedia, the free encyclopedia. – [ Електронний ресурс]. – Режим доступу: https://en.wikipedia.org/wiki/Machine_perception
10. Robotics - Wikipedia, the free encyclopedia. – [ Електронний ресурс]. – Режим доступу: https://en.wikipedia.org/wiki/Robotics
11. Affective computing - Wikipedia, the free encyclopedia. – [ Електронний ресурс]. – Режим доступу: https://en.wikipedia.org/wiki/Affective_computing
12. Artificial general intelligence - Wikipedia, the free encyclopedia. – [ Електронний ресурс]. – Режим доступу: https://en.wikipedia.org/wiki/Artificial_general_intelligence

13. Cybernetics - Wikipedia, the free encyclopedia. – [ Електронний ресурс]. – Режим доступу: https://en.wikipedia.org/wiki/Cybernetics

14. Symbolic artificial intelligence - Wikipedia, the free encyclopedia. – [ Електронний ресурс]. – Режим доступу: https://en.wikipedia.org/wiki/Symbolic_artificial_intelligence

15. Computational tools for artificial intelligence - Wikipedia, the free encyclopedia. – [ Електронний ресурс]. – Режим доступу: https://en.wikipedia.org/wiki/Computational_tools_for_artificial_intelligence

16. Artificial neural network - Wikipedia, the free encyclopedia. – [ Електронний ресурс]. – Режим доступу: https://en.wikipedia.org/wiki/Artificial_neural_network

17. Deep learning - Wikipedia, the free encyclopedia. – [ Електронний ресурс]. – Режим доступу: https://en.wikipedia.org/wiki/Deep_learning

18. Multilayer perceptron - Wikipedia, the free encyclopedia. – [ Електронний ресурс]. – Режим доступу: https://en.wikipedia.org/wiki/Multilayer_perceptron#Terminology

19. Booth, P.A. An Introduction to Human-Computer Interaction. - Hove, UK. // Psychology Press, 1989.

20. EmoDB - Berlin Database of Emotional Speech. – [ Електронний ресурс]. – Режим доступу: http://www.emodb.bilderbar.info/navi.html

21. RAVDESS – SMART Lab. – [ Електронний ресурс]. – Режим доступу: https://smartlaboratory.org/ravdess/

22. TESS - TSpace. – [ Електронний ресурс]. – Режим доступу: https://tspace.library.utoronto.ca/handle/1807/24487

23. Мел (высота звука) - Wikipedia, the free encyclopedia. – [ Електронний ресурс]. – Режим доступу: https://ru.wikipedia.org/wiki/%D0%9C%D0%B5%D0%BB_(%D0%B2%D1%8B%D1%81%D0%BE%D1%82%D0%B0_%D0%B7%D0%B2%D1%83%D0%BA%D0%B0)

24. Кепстр - Wikipedia, the free encyclopedia. – [ Електронний ресурс]. – Режим доступу: https://ru.wikipedia.org/wiki/%D0%9A%D0%B5%D0%BF%D1%81%D1%82%D1%80

25. Mel-frequency cepstrum - Wikipedia, the free encyclopedia. – [ Електронний ресурс]. – Режим доступу: https://en.wikipedia.org/wiki/Mel-frequency_cepstrum

26. Chroma feature - Wikipedia, the free encyclopedia. – [ Електронний ресурс]. – Режим доступу: https://en.wikipedia.org/wiki/Chroma_feature

27. Emotional intelligence - Wikipedia, the free encyclopedia. – [ Електронний ресурс]. – Режим доступу: https://en.wikipedia.org/wiki/Emotional_intelligence

**Додаток А**

**Software Architecture Document (SAD)**

# Taras Shevchenko National University of Kyiv

Research of the problem of recognition of voice emotions in the development of artificial intelligence

Software Architecture Document (SAD)

**CONTENT OWNER: HNIDENKO VALIERIIA**

| DOCUMENT NUMBER: | RELEASE/REVISION: | RELEASE/REVISION DATE: |
|---|---|---|
| 1 | 1.0 | 15.05.2021 |

## Table of Contents

# 1. Documentation Roadmap

## 1.1. Document Management and Configuration Control Information

- Revision Number: 1
- Revision Release Date: 15.05.2021
- Purpose of Revision: check the performance of the developed method for recognizing voice emotions

## 1.2. Purpose and Scope

Purpose: this document provides an architectural overview of the system, using Use-Case View to depict the system. It is intended to capture and convey the architectural decisions which have been made on the system.

Scope: the scope of this SAD is to explain the architecture of a voice-based emotion recognition system. The voice-based emotion recognition system is being developed by Valieriia Hnidenko for further development of AI.

## 1.3. Viewpoint Definitions

### 1.3.1. Use-Case Viewpoint Definition

#### 1.3.1.1. Abstract

A description of the use-case view of the software architecture. The Use Case View is important input to the selection of the set of scenarios and/or use cases that are the focus of an iteration. It describes the set of scenarios and/or use cases that represent some significant, central functionality. It also describes the set of scenarios and/or use cases that have a substantial architectural coverage (that exercise many architectural elements) or that stress or illustrate a specific, delicate point of the architecture.

#### 1.3.1.2. Stakeholders and Their Concerns Addressed

Stakeholders and their concerns include:

- Developers, managers, leaders who need to know how the software is designed for further emotion recognition system development.
- Testers that test components and identify system performance issues.
- Third party developers who are interested in this voice emotions recognition method.

#### 1.3.1.3. Elements, Relations, Properties, and Constraints

A use case diagram is a graphical depiction of a user's possible interactions with a system. A use case diagram shows various use cases and different types of users the system has. The use cases (properties) are represented by ellipses. The actors (elements) are shown as stick figures. Their relations can be seen in the figure 1.3.1.3.1.

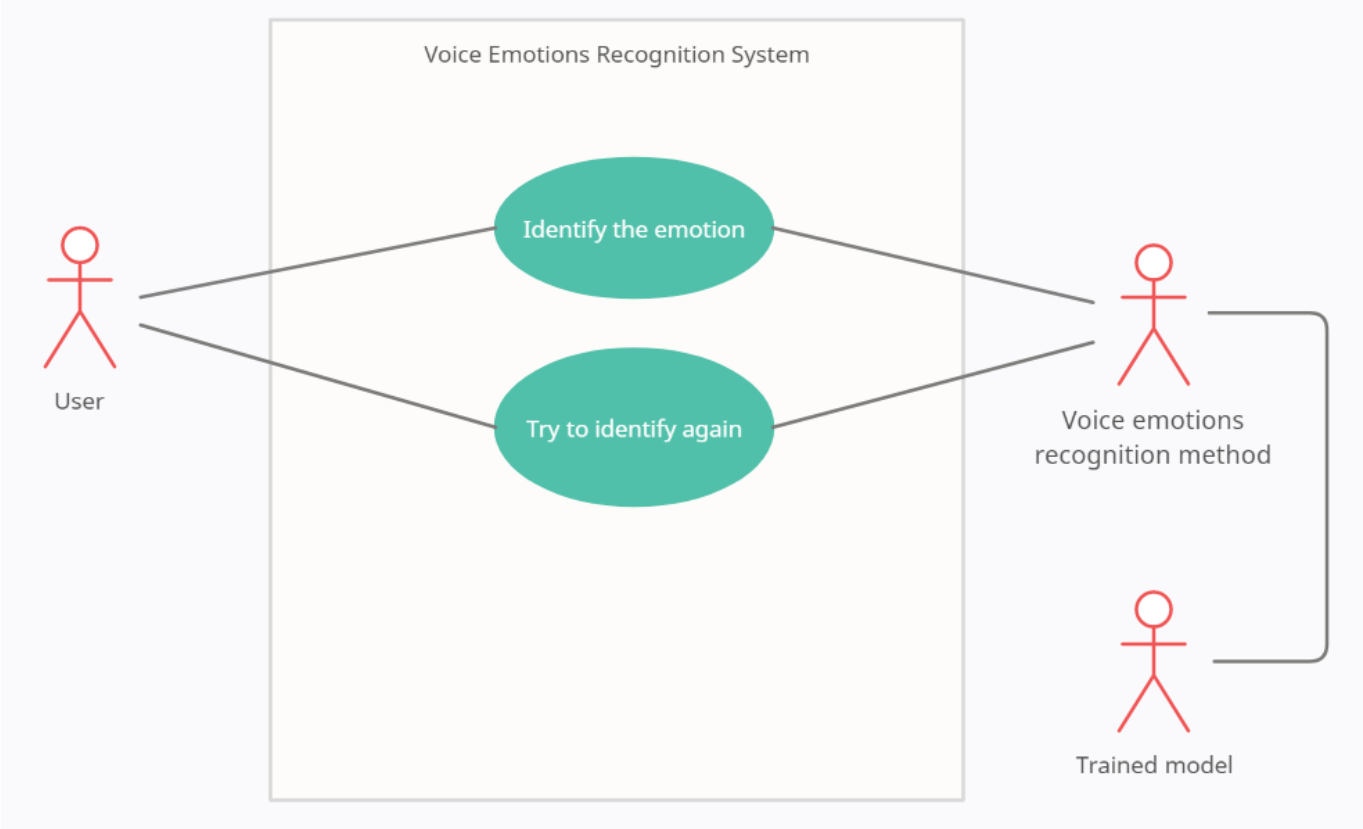


*Figure 1.3.1.3.1: Use case diagram*

The voice emotions recognition system (VERS) use cases are:

- Identify the emotion.
- Try to identify again.

The voice emotions recognition system constraints are:

- Limited number of database samples.
- The influence of culture and language on VERS.
- Uncertainty in the annotation.
- Lack of adaptation to natural conditions.

#### 1.3.1.4. Language(s) to Model/Represent Conforming Views

Views can be formed using such representations:

- Plain text that can be presented in a variety of formats, such as lists, forms, or regular paragraphs.
- Representation in the form of UML diagrams representing a unified modeling language is used in the paradigm of object-oriented programming, and allows correctly describe the data in graphical form.
- Software based on the representation of the system model in the form of code, algorithmic sequences and constructions of programming languages.

# 2. Architecture Background

## 2.1. Problem Background

### 2.1.1. System Overview

This system is a tool that works according to the designed model and can be used by third party developers who are interested in voice emotions recognition method usage.

Although the voice emotions recognition system was originally developed as part of an emotion recognition system (ERS), the technology can also be applied in language translation, interactive online lessons and courses, vehicle safety, therapy sessions, stressful and noisy environments such as in aircraft cockpits, service and e-commerce industries, call centres, interactive films, etc.

### 2.1.2. Goals and Context

At the moment, artificial intelligence has most of its problems because of its inability to think "like a human being." In other words, something like sarcasm can already make his assumption incorrect, because words apart from speech can have different meanings. Perception of emotions of people and other living beings will make it possible to make significant progress in solving such AI problems as limitation, reasoning, problem solving, knowledge representation, planning, learning, perception, natural language processing, social intelligence, general intelligence. VERS as a part of ERS is a perfect tool for the emotion perception problem.

### 2.1.3. Significant Driving Requirements

The careless use of AI technology, or the creation of the wrong architecture of AI can be dangerous for humanity. The reason for this is that while managing the emotions of living beings will be routine for AI, its abilities in some areas will be far superior to human. Therefore, the following principles (rules) should be adhered while designing a working version of general artificial intelligence (GAI), which will combine most of the narrow skills of AI and ERS: The voice emotions recognition system constraints are:

- Limited number of database samples.
- Lack of needs - complete lack of interest in obtaining personal benefits.
- Lack of will to live - existence is not valued, self-defence is absent, if a person intends to get rid of all components of AI, it will not contradict it.
- Lack of desire to improve - no desire to create or improve AI on its own based on own considerations and needs.
- Impossibility of self-reflection - no feelings, character, preferences, personality.
- Lack of subconsciousness - all decisions are calculated, the influence on decision-making can not be caused by intuition.
- Inability to make strategically important decisions for people - analysis of large databases and proposing solutions to problems without making a final decision on its own.
- Freedom of action is limited - nothing can be done without human consent.

This list can go on for a long time, but it will all come down to the fact that the AI or GAI should be as impartial and indifferent as possible to the tasks that it was not told to solve. Following these rules in the future people can get a lot from GAI without risks to humanity.

## 2.2. Solution Background

### 2.2.1. Architectural Approaches

The basic idea of the tool for recognizing the emotions of the voice is to create and train / test an appropriate machine learning algorithm (as well as in-depth learning) that could recognize and detect human emotions in the voice.

Three databases were used: RAVDESS, TESS, EMO-DB. In the future, it is also planned to attract a "special" database that will contain unbalanced noisy data sets that are closest to the natural emotions of the voice.

There are currently 9 emotions available: neutrality, calm, happiness, sadness, anger, fear, disgust, pleasant surprise and boredom. When expanding the sample, it will be possible to add new emotions.

Extraction of features is the main part of the system of recognition of speech emotions. Basically, this is achieved by changing the shape of the speech wave to the form of a parametric representation with a relatively lower data rate. To extract the features, Mel, Chroma, and MFCC were used.

In order to obtain the best possible hyperparameters, a grid search was performed.

The developed tools provide an opportunity to train different models, determine the best of them (according to the test results), effectively predict the emotions of the voice as through the prepared file with the extension. wav, and in real time while talking. It is also possible to display the obtained data in the form of a histogram.

The possibility of constructing classifiers and regressors of machine learning were realized for the case of 3 emotions (sadness, happiness, neutrality). SVC, RandomForest, GradientBoosting, KNeighbors, MLP, Bagging were used.

### 2.2.2. Analysis Results

VERS is the most difficult form of emotion recognition due to many constraints explained earlier. Despite good accuracy in testing different models, VERS results are not always correct due to use in training models of simulated ideal sets, which are significantly different from everyday natural language.

To solve VERS problems it is necessary to create:

- a huge database, which will contain mostly natural sets;
- a language model that will adapt to the language to identify emotions on different languages and accents;
- noise-resistant model for VERS;
- models that can classify the continuous flow of speech emotions;
- models of transition of emotions.

### 2.2.3. Requirements Coverage

In the process of creating this system, the original goals and requirements were achieved. Voice emotions recognition method was released and successfully used in VERS.

Development Requirements:

- PyCharm Community Edition 2021.1.1. or other development environment;
- Python 3.7;
- Librosa 0.6.3;
- Pyaudio 0.2.11;
- Matplotlib 2.2.3;
- Sklearn;
- Soundfile 0.9.0;
- Wave 0.0.2;
- Pandas 1.1.5;
- Numpy 1.19.5.

## 3. Referenced Materials


| | |
|---|---|
| IEEE 1471 | ANSI/IEEE-1471-2000, IEEE Recommended Practice for Architectural Description of Software-Intensive Systems, 21 September 2000. |
| Barbacci 2003 | Barbacci, M.; Ellison, R.; Lattanze, A.; Stafford, J.; Weinstock, C.; & Wood, W. Quality Attribute Workshops (QAWs), Third Edition (CMU/SEI-2003-TR-016). Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University, 2003. <http://www.sei.cmu.edu/publications/documents/03.reports/03tr016.html>. |
| Bass 2003 | Bass, Clements, Kazman, Software Architecture in Practice, second edition, Addison Wesley Longman, 2003. |
| Clements 2002 | Clements, Bachmann, Bass, Garlan, Ivers, Little, Nord, Stafford, Documenting Software Architectures: Views and Beyond, Addison Wesley Longman, 2002. |
| Russell 2009 | Russell, Stuart J.; Norvig, Peter. Artificial Intelligence: A Modern Approach (3rd ed.). Upper Saddle River, New Jersey: Prentice Hall, 2009. |
| McCorduck 2004 | McCorduck, Pamela, Machines Who Think (2nd ed.), Natick, MA: A. K. Peters, Ltd., 2004. |

# 4. Directory

## 4.1. Glossary

| Term | Definition |
|---|---|
| Software architecture | The structure or structures of that system, which comprise software elements, the externally visible properties of those elements, and the relationships among them [Bass 2003]. "Externally visible" properties refer to those assumptions other elements can make of an element, such as its provided services, performance characteristics, fault handling, shared resource usage, and so on. |
| View | A representation of a whole system from the perspective of a related set of concerns [IEEE 1471]. A representation of a particular type of software architectural elements that occur in a system, their properties, and the relations among them. A view conforms to a defining viewpoint. |
| Viewpoint | A specification of the conventions for constructing and using a view; a pattern or template from which to develop individual views by establishing the purposes and audience for a view, and the techniques for its creation and analysis [IEEE 1471]. Identifies the set of concerns to be addressed, and identifies the modeling techniques, evaluation techniques, consistency checking techniques, etc., used by any conforming view. |
| Mel | Psychophysical unit of pitch, used mainly in musical acoustics. The name comes from the word "melody".<br>The quantitative assessment of sound by pitch is based on statistical processing of a large amount of data on the subjective perception of the pitch of sound tones. Research results show that the pitch is mainly related to the vibration frequency, but also depends on the sound volume and timbre. |
| Chroma | In Western music, the term chroma feature or chromagram closely relates to the twelve different pitch classes. Chroma-based features, which are also referred to as "pitch class profiles", are a powerful tool for analyzing music whose pitches can be meaningfully categorized (often into twelve categories) and whose tuning approximates to the equal-tempered scale. One main property of chroma features is that they capture harmonic and melodic characteristics of music, while being robust to changes in timbre and instrumentation. |
| MFCC | In sound processing, the mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency.<br>Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal spectrum. This frequency warping can allow for better representation of sound, for example, in audio compression. |

## 4.2. Acronym List

| | |
|---|---|
| SAD | Software Architecture Document |
| AI | Artificial Intelligence |
| VERS | Voice Emotions Recognition System |
| UML | Unified Modeling Language |
| ERS | Emotion Recognition System |
| GAI | General Artificial Intelligence |

## 5. Sample Figures & Tables

The sample of the result of recognizing happiness (figure 5.1) is shown below.


*Figure 5.1: Happiness identified*