# Тема 12. Обробка природної мови

Обробка природної мови (NLP - Natural Language Processing) - область, що знаходиться на перетині інформатики, штучного інтелекту та лінгвістики. З розвитком голосових інтерфейсів і діалогових систем, NLP стала однією з найважливіших технологій штучного інтелекту.Повне розуміння та відтворення змісту мовного фрагменту - надзвичайно складне завдання, оскільки людська мова має особливості:

* Це специфічна система усвідомленої передачі змісту задуманого або написаного.
* Це символьна система, для якої притаманна адаптивність, надійність та простота для опанування.
* Символи мови використовуються для спілкування по кількох каналах: звук, жести, текст, зображення і подібне.
* Природній голос не вимагає спеціальних знань і навичок, забезпечує всіх людей рівними можливостями.

Технології обробки природної мови давно вирішують вузькі завдання: пошукові системи обробляють запити з врахуванням семантики мови, сучасні перекладачі пристойно відображають загальний зміст фрази іншими мовами, голосові асистенти роблять все менше помилок, а спеціальні архітектури нейронних мереж дозволяють витягувати актуальну інформацію з довільних текстів.

Основним завданням в NLP на сьогодні постає створення універсальних мовних моделей і архітектур, які будуть оперувати з текстом і голосом за допомогою однієї системи “End-To-End”. Системи, які будуть «розуміти» інформацію і зможуть взаємодіяти з користувачем подібно до людини з певним багажем знань.

#### Основні напрямки досліджень NLP

* Мовні технологій:
  + Розпізнавання мови (STT, Speech-to-Text).
  + Синтез мови (TTS, Text-to-Speech).
  + Голосова біометрія (Voice Biometry).
* Розуміння природної мови (NLU, Natural Language Understanding).
  + Великі мовні моделі (Large Language Models)
  + Діалогові системи, голосові помічники.
  + Машинний переклад (Machine Translation).
  + Системи автоматичної генерації контенту (Content Generation Systems)
* Інші сфери взаємодії «людина-машина».

## Мовні технології

Мовні технології складаються з кількох складних напрямів.

* **Розпізнавання мови.** Перетворення мовного повідомлення у текст, бажано з максимальною точністю і якістю.
* **Синтез мови.** Завдання, зворотне до розпізнавання – озвучення тексту голосом так, щоб було зрозуміло для людини.
* **Голосова біометрія.** Визначення за голосом багатьох важливих характеристик людини: ідентифікація особи, стать, вік, емоційний стан.

На базі мовних технологій реалізуються алгоритми розуміння природної мови та складні системи: голосові інтерфейси, діалогові системи, інтелектуальні помічники, машинні перекладачі тощо.

## 13.1. Розпізнавання мови

Розпізнавання мови - це процес перетворення мовного сигналу в цифрову інформацію. Це є одним з найскладніших технічних завдань, що відноситься до систем штучного інтелекту. Машинне перетворення мови в текст і озвучення відповіді увійшли в повсякденне життя та поставлені на комерційний потік.

Стрибок у розвитку мовних технологій стався завдяки збільшенню потужності обчислювальних ресурсів та використанню нейронних мереж для вирішення багатьох завдань. У розпізнаванні мови використовуються останні досягнення в області машинного навчання, що сприяють високій точності розпізнавання.

#### Основні сфери застосування машинного розпізнавання мови:

* **Системи голосового обслуговування та інтерактивні автовідповідачі.** Найбільшого поширення набули в контакт-центрах, сервісах самообслуговування, онлайн-банкінгу: привітання, голосове меню, озвучення стану рахунків тощо.
* **Ідентифікація особи.** Використовується великими банками для підтвердження особи користувача за голосовим відбитком, для голосового підпису, а також в системах безпеки.
* **Аналітика дзвінків і переговорів.** Призначена для оцінки відгуків клієнтів, підвищення якості роботи операторів, виявлення трендів при зверненнях до служби підтримки або відділів з продажу.
* **Голосове управління.** Застосовується в багатьох сферах людської діяльності: в Інтернеті для пошуку інформації, в побуті для керування домашніми приладами, в системах «розумний будинок», навігаторах, управління автотранспортом, в мобільних додатках для виконання певних завдань.
* **Транскрипція та аудіоіндексація.** Перетворення мови на текст використовується для створення транскрипцій аудіо- та відеозаписів, інтерв'ю, конференцій та інших аудіоматеріалів. Це спрощує процес аналізу, пошуку та організації звукової інформації.

#### Методи розпізнавання мови

На сьогодні сформувалося багато методів розпізнавання мови, у загальному їх можна розподілити на наступні типи:

* **Розпізнавання голосових міток та окремих команд.** Побудовано на роздільному проголошенні і подальшому розпізнаванні слова або словосполучення з невеликого тематичного словника. Використовується у відносно простих системах, призначених для виконання заздалегідь записаних голосових команд.
* **Методи на основі словників.** Ці методи використовують попередньо підготовлені словники та мовні моделі для зіставлення аудіосигналу з відомими словами та фразами. Це часто використовується в системах розпізнавання мовлення, які працюють з обмеженим набором команд чи слів.
* **Розпізнавання лексичних елементів мови.** Ці методи ґрунтуються на аналізі фонем (звукових одиниць), з яких складається мова. Моделі фонем використовуються для розпізнавання та сегментації мовлення, не обмежуючись при цьому заданою граматикою. При його реалізації з потоку мови виділяються окремі звуки і з них складаються слова. Ситуацію ускладнює те, що одна і та ж фраза, вимовлена ​​різними людьми в різних умов, буде звучати по-різному і давати несхожі один на одного сигнали.
* **Статистичні методи.** Ці методи ґрунтуються на аналізі статистичних властивостей мови та використання ймовірнісних моделей для розпізнавання звуків та слів в аудіосигналі. Використовуються статистичні закономірності, отримані з навчальних даних, для ухвалення рішень, що було в промові.
* **Методи глибокого навчання.** Сучасні методи, такі як рекурентні нейронні мережі (RNN), згорткові нейронні мережі (CNN) та рекурентно-згорткові нейронні мережі (RCNN), демонструють високу ефективність у завданнях розпізнавання мови. Глибоке навчання дозволяє отримувати ознаки з аудіосигналу та будувати точніші моделі розпізнавання.
* **Змішані методи**. Деякі системи розпізнавання мови комбінують кілька методів досягнення більш високої точності. Наприклад, для ініціалізації моделей глибокого навчання можуть використовувати статистичні методи.

#### Основні терміни

Для подальшого розуміння матеріалу варто знати основні терміни, які закладено в алгоритмах розпізнавання мови.

Речення та повідомлення складаються зі слів, слова - зі складів і морфем, склади і морфеми - з фонем і алофонів.

* **Морфема** — найменша частина слів, що має певне змістовне значення.
  + Морфеми – це цеглинки, з яких складаються слова: корінь, префікс, суфікс, постфікс, основа, закінчення, інтерфікс.
  + Приклад: [комп’ютер/н/ий]
* **Фонема** - це найменша звукова одиниця мови. Фонемою може бути як окремий звук, так і поєднання декількох звуків, що на слух сприймається як один звук. В слов’янських мовах є близько 50 фонем.
  + Приклад: [к] [о] [м] [п] [йу] [т] [е] [р] [н] [и] [й]
  + Характеристиками фонеми є сенони - початок, середина або кінець фонеми.
* **Алофон** – це конкретний мовний звук, що обумовлений конкретним фонетичним оточенням. Всі реально вимовлені звуки - це алофони, що об'єднуються у відносно невелике число фонем.
  + Приклад: [к] [о] [м] [п] [й] [у] [т] [е] [р] [н] [и] [й]

### 13.1.1. Попередня обробка звукових сигналів

Попередня обробка мовного сигналу призначена для видалення шумів і сторонніх сигналів, частотний спектр яких знаходиться поза спектром людської мови. Відфільтрований звуковий сигнал проходить через аналого-цифрове перетворення і оцифровується.

Зазвичай, системи розпізнавання мови не вимагають для своєї роботи спеціальних пристроїв. Попередню обробку звукового сигналу можна зробити за допомогою стандартного звукового адаптера, встановленого в пристрої чи комп'ютері. Додаткова цифрова обробка звукового сигналу (наприклад, частотна фільтрація) може виконуватися центральним процесором комп'ютера.

Важливим етапом попередньої обробки вхідного сигналу є нормалізація рівня сигналу. Це дозволяє зменшити похибки розпізнавання, пов'язані з тим, що диктор може вимовляти слова з різним рівнем гучності. Якщо вхідний звуковий сигнал має занадто малий рівень гучності, то після нормалізації може з'явитися шум. Для успішної роботи системи розпізнавання мови необхідно відрегулювати оптимальним чином чутливість мікрофона. Втім, надмірна чутливість може привести до спотворень сигналу і до збільшення похибки розпізнавання мови.

Пристрій (комп’ютер чи смартфон) записує вимовлений голосовий запит і надсилає до серверу (рис. 13.1).



Рис. 13.1. Передача вимовленого голосу до серверу

Там запис розділяється на багато маленьких фрагментів - фреймів. На кожну секунду мови доводиться сто фреймів. Вони тривають по 25 мілісекунд і слідують один за одним, перекриваючи попередній фрейм, як черепиця, щоб інформація на стиках не втрачалася.

Спектр сигналу масштабується і після ряду перетворень для кожного фрейму обчислюють коефіцієнти. Кожен фрейм піддається ряду перетворень, в результаті яких виходить близько 40 коефіцієнтів, що описують його частотні характеристики.

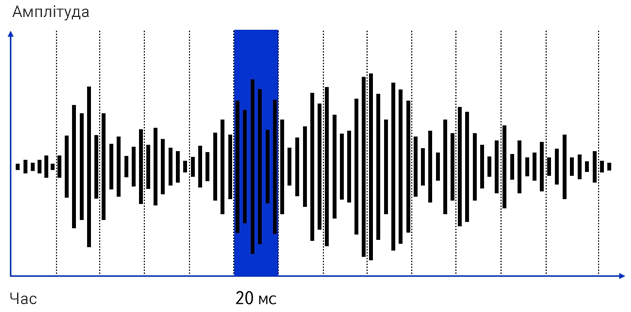


Рис.13.2. Формування фреймів з сигналу

Правильно інтерпретувати оброблені звуки допомагають акустична та мовна моделі.

### 13.1.2. Акустична модель

Фрейми надходять на вхід акустичної моделі, яка обчислює розподіл ймовірностей, яке сполучення фонем закладено в кожній частині цього звуку. Цьому модель вчиться на великому корпусі начитаних дикторами текстів та їх транскрипцій (сотні годин розміченої мови). Збільшення бази сприяє зростанню якості навчання моделі, але важливим є правильна підготовка даних і акуратний запис транскрипцій, що дозволяє з прийнятною якістю навчатися на відносно невеликій базі.

Як матеріал для прикладів беруться різноманітні аудіозаписи: офіційні промови на публічних заходах, діалоги з фільмів, аудіокниги, любительські відеоролики в Інтернеті та багато різних джерел. Спеціально підготовлені фахівці слухають записи, пишуть, що було сказано в цій промові, і позначають, де були шуми чи фонові розмови.

Найскладнішим завданням в розпізнаванні мови є відокремлення основного спікера від фонових. Це завдання називається «вечірка з коктейлем»: коли багато народу, всі говорять, чутно обривки фраз, гримить музика на фоні, і треба здогадатися, хто є основним спікером, чий голос потрібно розпізнати.

Для розпізнавання особливостей мови і, зокрема, акцентів створюють бази текстів від людей, що мають певні особливості у вимовлянні. Великої допомоги надає автоматичний збір аудіозаписів від різних сервісів: навігаторів, голосових пошукових запитів, ігрових та розважальних додатків. В результаті накопичуються максимально можливі варіанти звучання довільної природної мови. Ці дані використовуються в навчанні акустичної моделі і у підсумку вона вміє правильно розуміти будь-які акценти, логопедичні недоліки мови та особливості темпу вимовляння.

Алгоритм акустичної моделі зіставляє промову з відповідним текстом. Після ознайомлення з достатньою кількістю матеріалів і накопичивши достатню статистику збігів, модель визначає, якому звуку з такими характеристиками відповідає яка фонема, іншому звуку - інша, і так далі.

#### Рівні розпізнавання мовлення

Розпізнавання промови є багаторівневим процесом. На кожному з рівнів сигнал кодується з врахуванням попередніх рівнів:

1. Виділення звукових елементів мови (фонеми та алофони).
2. Виділення складів і морфем.
3. Виділення слів, речень та повідомлення.

При переході з рівня на рівень передаються і деякі додаткові ознаки, тимчасові залежності і відносини між сигналами. Збираючи сигнали з попередніх рівнів, вищі рівні містять більший об'єм інформації, і можуть здійснювати управління процесами на нижчих рівнях.

Звуки мови є різноманітними, оскільки звучання фонеми залежить від того, в якій частині слова вона знаходиться - на початку, середині або наприкінці, і які фонеми її оточують. Наприклад, [а] між двома голосними у поєднанні «на аудіозапису» відрізняється від [а] між приголосними в слові «бак». Тому, для якісного розпізнавання фонема є занадто грубою одиницею.

Щоб точніше змоделювати вимову фонеми вживають наступне

* Кожна фонема поділяється на три частини-сенони: умовні початок, середину і кінець.
* Використовується фонетичний алфавіт, який враховує позицію і контекст фонем. Він складається з тисяч елементарних одиниць різноманітних поєднань і комбінацій. З цим набором і працює технологія розпізнавання мови.

Для розпізнавання фонем та алофонів застосовуються нейронні мережі та їх ансамблі. До ансамблів застосовують навчання без вчителя, при якому відбувається статистична обробка всіх сигналів, що надходять на вхід нейронної мережі. Ансамблі формуються відповідно до сигналів, що найчастіше зустрічаються. Запам'ятовування рідкісних сигналів відбувається пізніше і вимагає контролю вищого рівня.

В алгоритмах розпізнавання також закладається голосова біометрія - це класифікатор, який навчається на багатьох різних розмічених аудіозаписах з певною ймовірністю розпізнавати особу диктора: чоловік або жінка, приблизний віковий діапазон. Звичайно, іноді бувають помилки - голоси деяких жінок схожі на дитячі, а у літньої людини молодий, бадьорий голос.

### 13.1.3. Мовна модель

Якщо людина не може зрозуміти або не розчула все звуки, тоді вона добудовує слово виходячи з контексту. Для цього люди спираються на власний мовний досвід: якщо співрозмовник скаржиться на закладений ніс, можна зрозуміти, що у нього «нежить», а не «дежить». Аналогічно працює система розпізнавання, тільки замість мовного досвіду вона використовує мовну модель.

Мовна модель описує відповідну тематичну область та містить загальні відомості. Наприклад, для навігатора - це тексти на географічну тематику (адреси, назви організацій і так далі), для пошуку в Інтернеті - тематика пошукових запитів, все, що люди запитують в пошуковому рядку. Щоб знати все різноманіття слів та словоформ, наголоси, стійкі та поширені поєднання слів, контекст вживання мовна модель навчається на текстах, у великій кількості (терабайти тексту).

Під час навчання мовна модель звертає увагу не на відповідність звуків і фонем, а на те в які послідовності - тобто слова і фрази – зазвичай, складаються фонеми.

Мовна модель працює з ланцюжком ймовірних фонем. Акустична модель після обробки частотних ознак фрейму, видає не одну конкретну фонему, а кілька - і в кожної з них свій коефіцієнт ймовірності. Наприклад, у разі «нежиті» акустична модель надає дві приблизно однакових ймовірних фонеми на початку слова - [н] і [д]. Мовна модель призначена вирішити цю суперечність. Під час навчання слово «дежить» зустрічалося їй нечасто, а от слово «нежить» - багато разів. Тому, система зробить висновок, що на початку слова, найімовірніше, знаходиться фонема [н] (рис. 13.3).

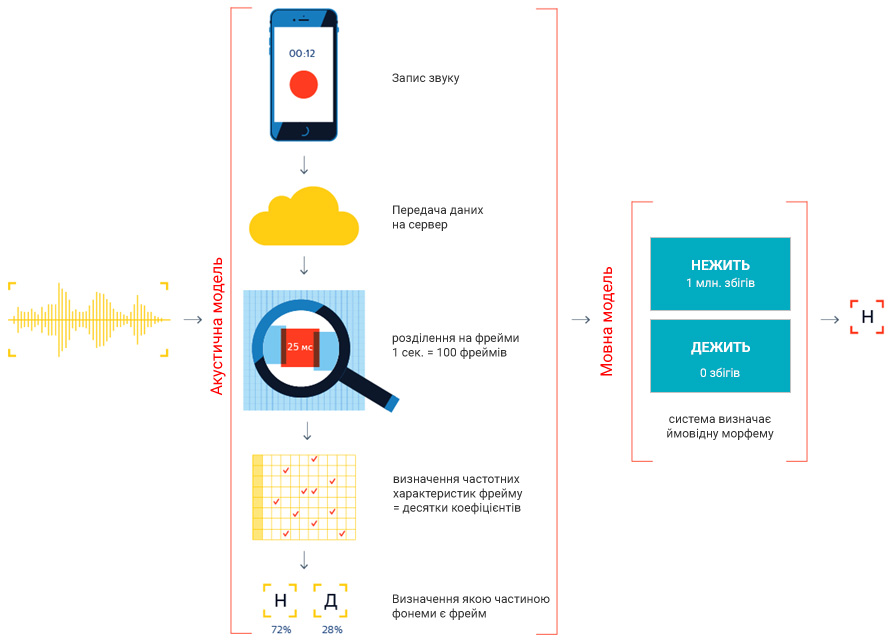


Рис.13.3. Формування фреймів з сигналу

Приблизно так само, виходячи з контексту, мовна модель визначає послідовності слів. Наприклад, вибираючи між «ніч яка місячна» і «нічья комісячна», вона надасть перевагу першому словосполученню, оскільки воно знайоме їй з навчального корпусу текстів.

## 13.2. Синтез мови

Синтез мови (перетворення тексту в голос) на сьогоднішній день застосовується в самих різних областях. Це голосові асистенти, IVR-системи (Interactive Voice Response, інтерактивне голосове меню), розумні будинки і багато інших сервісів. Написаний текст повинен вимовлятися природньо правдоподібно до людської промови.

В напрямку синтезу голосу пройдено довгий шлях і застосовано багато методів, які використовуються для створення штучного мовлення на основі текстової інформації.

### Конкатенативний синтез «Unit-Selection»

Цей метод ґрунтується на попередньому записі мовних фрагментів та їх подальшому об'єднанні для створення цілого речення чи тексту. У цьому методі записи людського мовлення поділяються на невеликі фрагменти (фони, склади, слова), які потім комбінуються, щоб згенерувати бажаний текст. Це забезпечує високу натуральність та чіткість вимови, але потребує великої бази даних зі звуковими записами. Фрагменти можуть бути записані людиною або витягнуті з великих баз даних мовних матеріалів. У процесі синтезу вибираються та об'єднуються відповідні фрагменти для створення безперервної та плавної мови.

Якщо набір текстів, які треба озвучити, є відносно невеликим і там зустрічаються повторні висловлювання - як, наприклад, в оголошеннях про відправлення чи прибуття поїздів на вокзалі, - достатньо запросити диктора, записати в студії потрібні слова і фрази, а потім збирати з них повідомлення.

Диктор в студії наговорює багато текстів, всі промовлені слова розділяються на елементарні елементи (наприклад, на фонеми). Складається велика база даних: як звучить кожна фонема в залежності від того слова, в якому вона знаходиться. І коли виникає завдання озвучити якийсь текст, з цієї бази збираються і склеюються звуки.

На коротких текстах це працює пристойно: стандартні фрази: «Вас вітає компанія така-то» або «Сьогодні погода така-то». Але це не синтез мови, а просто відтворені частинки реальної мови. Такий підхід не працює з довільними довшими текстами. Тоді у синтезі починає спотворюватися інтонація або голос, оскільки немає алгоритму, який відповідає за правильне об’єднання інтонації на стиках змістовних фрагментів фраз чи речень. Синтез довгого аудиотексту (наприклад, озвучити книгу) швидко починає дратувати, оскільки інтонація є монотонною і людина втомлюється це чути.

### Синтез на основі формантного моделювання

Цей метод моделює штучну мову на основі моделювання формантних характеристик звуків. Форманти – це резонансні піки, які визначають звуковий спектр мовлення. Модель формантних характеристик може бути використана для генерації звуків з певними формантними параметрами, що дозволяє створювати різні мовні звуки.

Синтез на основі формантного моделювання починається зі створення моделі формантних характеристик різних звуків. Ця модель може бути створена шляхом аналізу та вивчення звуків на основі мовних даних. У процесі моделювання формантних характеристик враховуються параметри, такі як частота та амплітуда формантних піків.

Коли текст подається на вхід до системи синтезу, вона використовує моделі формантних характеристик для генерації відповідних звуків. Звуки можуть бути згенеровані шляхом комбінування та зміни формантних піків відповідно до заданих параметрів, таких як тривалість та інтенсивність звуку.

Синтез на основі формантного моделювання зазвичай добре працює для створення мови з чіткими та вираженими звуками. Однак, він може стикатися з обмеженнями при синтезі більш природного і виразного мовлення, оскільки форманти не враховують інші аспекти, такі як інтонація, мелодійні характеристики тощо. Також потрібний великий обсяг даних та точна модель формантних характеристик для досягнення якісного синтезу.

### Статистичний (параметричний) метод

Статистичний (параметричний) синтез мови аналогічний до розпізнавання, тільки відбувається в зворотному напрямку. Є акустична модель, яка навчається на даних - записи мови з текстовими транскрипціями. Для збору даних диктор читає текст, але в значно меншому обсязі. Наприклад, відбувається запис 600 речень, що промовлені непрофесійним диктором за один день в студії, здійснюється синтез і отримання голосу.

Створюються статистичні моделі, які описують зв'язок між текстом та акустичними характеристиками мови. Ці моделі можуть бути навчені великому обсязі мовних даних, щоб захопити різні аспекти мовлення, такі як звуки, інтонація і ритм.

Один із популярних підходів у статистичному синтезі мови – це приховані марківські моделі (ПMM). ПMM моделює статистичні залежності між послідовністю звуків та відповідними текстовими даними і складається з кількох компонентів.

* Прихований стан, який є абстрактним станом, пов'язаним з певною фонемою.
* Спостереження, яке є набором акустичних характеристик, таких як спектральні ознаки звуку.
* Перехідні можливості, які визначають можливості переходу між прихованими станами.

Наприклад, є текст, який потрібно перетворити на мову. Синтез за методом ПMM розділяє цей текст на окремі звуки або фонеми, які є основними звуковими одиницями мови. Потім модель ПMM будує статистичну модель, яка описує залежності між фонемами та відповідними акустичними характеристиками цих звуків. Синтезована мова виходить шляхом вибору найімовірнішої послідовності звуків для заданого тексту.

Синтез на основі ПMM дозволяє створювати мову з врахуванням статистичних залежностей між фонемами та акустичними характеристиками. Він може бути досить гнучким і здатний моделювати різні стилі та інтонації мови. Однак він має обмеження у створенні дуже природного та виразного мовлення, оскільки потрібно вловлювати складні акустичні особливості.

Моделі навчаються на великих обсягах даних і налаштовуватись для різних голосів та стилів мови. Вони вимагають значних обчислювальних ресурсів та часу для навчання моделі та генерації мови.

### Глибоке навчання

У синтезі мови також застосовується глибоке навчання. Глибокі нейронні мережі, такі як рекурентні нейронні мережі (RNN) або згорткові нейронні мережі (CNN), можуть бути навчені на великому обсязі мовних даних для генерації мовних сигналів. Цей метод дозволяє моделювати складніші залежності та створювати більш природну та виразну мову.

### Гібридні підходи

Деякі системи синтезу мовлення можуть комбінувати кілька методів для досягнення найкращих результатів. Наприклад, можна використовувати глибоке навчання для моделювання тривалих залежностей та конкатенативний синтез для створення безперервної мови.

Кожен із цих методів має свої переваги та обмеження, і вибір конкретного методу залежить від необхідної якості мови, доступних ресурсів та контексту застосування. В останні роки глибоке навчання стало основним підходом у синтезі мови, завдяки своїй здатності моделювати складні залежності та створювати високоякісну та природну мову.

### 13.2.1. Етапи створення системи синтезу

Для створення системи синтезу мовлення, потрібно виконати багато завдань, що потребує команди фахівців з різних напрямків. По кожному етапу реалізації існує ціла маса алгоритмів і підходів.

**Нормалізація тексту.** Спочатку спеціальний алгоритм готує текст, щоб роботу було зручно його читати: розгорнути всі скорочення, перевести числа і дати в текст. «50-ті роки XX століття» має перетворитися в «п'ятдесяті роки двадцятого століття», а «м.Київ, просп. Т.Шевченка» в «місто Київ, проспект Тараса Шевченка». Це повинно відбуватися природно, як якщо б людину попросили прочитати написане.

Далі текст поділяється на фрази, тобто на словосполучення з безперервною інтонацією - для цього комп'ютер орієнтується на знаки пунктуації та стійкі конструкції. Для всіх слів складається фонетична транскрипція.

**Підготовка словника наголосів.** Розставляння наголосів може проводитися за правилами мови. В англійській наголос часто ставиться на перший склад, а в іспанській - на передостанній. При цьому обов'язково потрібно враховувати цілу низку виключень, які не слідують загальним правилам. Для української мови в загальному сенсі правил щодо наголосів взагалі не існує, тому, наявність словника з розставленими наголосами є обов’язковою.

Щоб зрозуміти, як читати слово і де поставити в ньому наголос, синтезатор спочатку звертається до класичних, складених вручну словників, які вбудовані в систему. Якщо потрібного слова в словнику немає, синтезатор будує транскрипцію самостійно - спираючись на правила, які запозичено з академічних довідників. Якщо звичайних правил виявляється недостатньо - а таке трапляється, адже будь-яка жива мова постійно змінюється, - він використовує статистичні правила. Якщо слово зустрічалося в корпусі навчальних текстів, система запам'ятовує, на якому складі в ньому зазвичай наголошували диктори.

**Зняття омографів.** Омографи - це слова, які збігаються в написанні, але різняться в вимові. Носій мови легко розставить наголоси: «дверний замок» і «замок на горі». А «ключ від замка» є значно складнішим завданням. Повністю зняти омографів без врахування контексту неможливо.

**Виділення синтагм і розстановка пауз.** Синтагма – це слово або група слів, що представляють цілісний відрізок мовлення, який несе завершену за змістом думку. Коли людина говорить, вона зазвичай вставляє паузи між фразами. Потрібно навчитися розділяти текст на такі синтагми.

**Автопунктуація.** Роль розділових знаків у мові грають інтонаційні паузи, тому створюються алгоритми для їх розпізнавання. Паузи між словами отримують одну з міток: пробіл, точка, кома, знак питання, знак оклику, двокрапка. Щоб передбачити, яка мітка відповідає конкретної паузі, використовується контекст та враховуються три попередніх і два наступні слова. Ці нескладні правила дозволяють з досить високою точністю розставляти знаки.

Кожному знаку пунктуації призначається фонема, і з точки зору системи це є нові «слова», що складаються з «пунктуаційних» фонем - там, де виникали паузи або певним чином змінювалася інтонація.

**Визначення типу інтонації.** Вираз завершеності, питання і вигуки - найпростіші інтонації. А висловити іронію, сумнів або наснагу є складнішим завданням. Дослідження голосових характеристик надають можливість використовувати ці особливості для синтезу голосу. Синтезатор через характеристики звукових хвиль спроможний розпізнати емоційне забарвлення: сум, розпач, радість, подив тощо. Для цього тексти позначають мітками з відповідною емоцією.

Щоб модель голосу змогла побудувати алгоритми, які відповідають різним емоціям, треба правильним чином її навчити. Під час запису диктор вимовляє свої репліки по черзі нейтральним голосом, радісним та роздратованим. В ході навчання система виділяє і описує параметри і характеристики голосу, що відповідають кожному з цих станів.

**Отримання транскрипції.** Щоб прочитати підготовлений текст, використовується акустична модель, що відрізняється від акустичної моделі, яка застосовується при розпізнаванні мови. Тут важливішим є промовляння, а не написання, то очевидно замість букв (графем), логічно використовувати звуки (фонеми).

* У випадку з розпізнаванням моделі потрібно встановити відповідність між звуками з певними характеристиками і фонемами.
* У випадку з синтезом мови акустична модель, повинна, навпаки, за описами фонем скласти описи звуків.

Перетворення графемного запису у фонемний - окреме завдання, що складається з багатьох правил і винятків. Оригінальний текст передається в модуль G2P (Grapheme-To-Phoneme), де перетворюється в послідовність фонем.



Рис.13.4. Перетворення графемного запису сигналу у фонемний

Коли транскрипція готова, комп'ютер розраховує, як довго буде звучати кожна фонема, тобто скільки в ній фреймів. Потім кожен фрейм описується за багатьма параметрами:

* частиною якої фонеми він є і яке місце в ній займає;
* в який склад входить ця фонема;
* якщо це голосна, то чи на ній є наголос;
* яке місце вона займає в складі, склад у слові, слово у фразі, фраза в реченні;
* які розділові знаки є до і після цієї фрази;
* який знак стоїть наприкінці речення і яка його головна інтонація.

Для синтезу кожних 25 мілісекунд мови використовується багато даних. Інформація про найближче оточення забезпечує плавний перехід від фрейму до фрейму і від складу до слова, а дані про фразу і речення в цілому потрібні для створення правильної інтонації синтезованої мови. Модель намагається відтворювати унікальні характеристики голосу людини для впізнаваності синтезованого голосу.

Впізнати голоси, в першу чергу, можна за тембром, який залежить від особливостей будови органів мовного апарату в кожної людини. Тембр голосу можна змоделювати, тобто описати його характеристики - для цього достатньо начитати в студії невеликий корпус текстів. Після цього дані про тембр можна використовувати при синтезі мови, навіть такої, що невідома диктору.

Акустична модель видає числа, що задають параметри звукової хвилі. Для фізичного промовляння тексту використовується генератор звукових хвиль - вокодер. В нього завантажується інформація про частотні характеристики фрази, що отримана від акустичної моделі, а також дані про тембр, який надає голосу впізнаване забарвлення.



Рис.13.5. Генератор звукових хвиль - вокодер

**Обчислення параметрів інтонації.** Тембр такого голосу кілька «комп'ютерний», зате у нього природні і плавні інтонації. В параметрах налаштування можна задати ключ, і модуль синтезу буде продукувати голос з відповідним емоційним забарвленням.

Також, можна налаштовувати висоту основного тону і швидкість промовляння в залежності від розставлених пауз, підібраної послідовності фонем і типу вираженої інтонації. Чим більше даних, на яких навчається модель, тим краще її вимова і інтонація.

**Модифікація і звукові ефекти.** Для надання більшої природності застосовують пост обробку, спеціальні фільтри, що наближають синтезовану мову до людської та виправляють виявлені дефекти.

### 13.2.2. Застосування глибокого навчання нейронних мереж

На сьогодні в області синтезу мови широко застосовується глибоке навчання. Якщо цілий ряд компонентів системи синтезу можна замінити на нейронні мережі, що дозволить не просто наблизитися за якістю до існуючих алгоритмів, а навіть значно їх перевершити.

* **Генеративні моделі на основі нейронних мереж.** Глибокі нейронні мережі, такі як рекурентні нейронні мережі (RNN) та згорткові нейронні мережі (CNN), можуть бути використані для створення генеративних моделей мови. Ці моделі можуть перетворювати текстову інформацію в послідовності акустичних ознак, які можна синтезувати в мовний сигнал. Наприклад, рекурентні генеративні моделі (RNN-GAN) можуть бути використані для генерації мови, яка звучить більш природно та плавно.
* **Автоенкодери (VAE).** VAE – це клас моделей глибокого навчання, які можуть використовуватись для моделювання ймовірнісних розподілів прихованих змінних. У контексті синтезу мови VAE може використовуватися для моделювання залежностей між текстом і акустичними характеристиками мови. Вони дозволяють генерувати мову з врахуванням невизначеності та різноманітності варіантів звучання. Більш детальну інформацію можна отримати з відео урока «Автоенкодери. Що це та як вони працюють» <https://www.youtube.com/watch?v=iLoCVz3RJTc>
* **Трансформери.** Трансформери - це тип архітектури нейронних мереж, який широко використовується в галузі обробки природної мови. Вони можуть бути застосовані до синтезу мови для моделювання залежностей між текстом та акустичними ознаками. Трансформери дозволяють вловлювати контекстуальні залежності в тексті та створювати більш якісну та виразну мову.
* **Попередньо навчені моделі.** Глибоке навчання також дозволяє використовувати попередньо навчені моделі, такі як Tacotron і WaveNet, які навчені на великих обсягах мовних даних. Ці моделі є ефективними алгоритмами синтезу мови, які можуть генерувати мова високої якості з врахуванням інтонації, емоцій та інших аспектів звучання.

Глибоке навчання надає потужні інструменти для синтезу мовлення, що дозволяє моделювати складні залежності між текстом та акустичними характеристиками. Це дозволяє створювати більш природну, виразну та якісну штучну мову.

Популярними архітектурами для синтезу на сьогодні є рекурентні нейронні мережі (RNN). На них вирішують всі завдання, пов'язані з послідовностями - голосовими, текстовими або музичними і широко застосовують для моделювання залежностей між послідовностями вхідних та вихідних даних.

Більш детальну інформацію можна отримати з відео урока «Ознайомлення з рекурентними нейронними мережами» <https://www.youtube.com/watch?v=P6Zjr_dBCKE>

#### Використання RNN для синтезу мови містить наступні кроки:

**Попередня обробка тексту.** Збір та підготовка вхідних даних для навчання мережі. Вхідний текст спочатку піддається попередній обробці. Це є токенізація (розподіл тексту на окремі одиниці, такі як слова або символи), перетворення тексту в числові представлення (наприклад, індекси або вектори слів) та поповнення послідовностей для узгодження розмірності вхідних даних.

**Створення архітектури RNN.** Визначається архітектура RNN для моделювання залежностей між текстом та акустичними ознаками мови: довга короткострокова пам'ять LSTM (Long Short-Term Memory) або закрита рекурентна одиниця GRU (Gated Recurrent Unit). Визначається кількість прихованих шарів та розмірність прихованих станів.

**Побудова навчальних прикладів.** Формуються навчальні приклади, що складаються з вхідних послідовностей тексту та відповідних акустичних ознак. Ці приклади допоможуть моделі пов'язати текстову інформацію із відповідними акустичними характеристиками.

**Визначення функції втрат та оптимізатора**. Вибирається відповідна функція втрат, яка вимірює різницю між передбачуваними та фактичними акустичними ознаками. Зазвичай використовується середня квадратична похибка. Вибирається оптимізатор, який оновлюватиме ваги моделі на основі функції втрат. Поширеними оптимізаторами є алгоритм зворотного поширення похибки (Backpropagation Through Time, BPTT) та алгоритм адаптивного градієнта (Adaptive Gradient Algorithm, AdaGrad).

**Навчання моделі**. Процес навчання полягає у подачі вхідних текстових послідовностей з навчального набору даних у модель RNN та обчислення передбачуваних акустичних ознак. Обчислюється функція втрат та оновлюються ваги моделі за допомогою вибраного оптимізатора. Цей процес повторюється для всіх навчальних прикладів в епохах навчання.

**Тестування моделі**. Після завершення процесу навчання, модель можна протестувати на нових даних. На вхід моделі подається текст, на виході будуть передбачені акустичні ознаки. Для перетворення акустичних ознак на фактичний мовний сигнал використовують алгоритм синтезу, наприклад, голосову синтезуючу модель - вокодер.

**Оптимізація та налаштування.** Після тестування моделі можна провести оптимізацію та налаштування параметрів моделі для досягнення кращої продуктивності. Це може бути зміна архітектури моделі, експерименти з глобальними параметрами або застосування додаткових методів, таких як аугментація даних.

**Аугментація аудіо даних.** Це процес створення нових варіацій аудіофайлів шляхом застосування різних перетворень до вихідних аудіофайлів. Це техніка, яка дозволяє збільшити різноманітність навчального набору аудіо даних та покращити здатність моделі до узагальнення.

* Зсув у часі. Аудіофайли зсуваються у часі, що призводить до зміни часової структури звуку.
* Масштабування. Аудіофайли масштабуються в часі, що призводить до зміни тривалості звуку.
* Зміна швидкості відтворення. Швидкість відтворення звуку змінюється, що може призвести до прискорення або уповільнення звуку.
* Зміна тональності. Аудіофайли можуть транспонуватися на певну кількість півтонів вгору або вниз, змінюючи висоту звуку.
* Додавання шуму. До аудіофайлів додається шум, щоб зробити їх більш реалістичними та різноманітними.
* Застосування фільтрів. До аудіофайлів можуть бути застосовані різні аудіофільтри, щоб змінити звукові характеристики.
* Еластичні спотворення. Аудіофайли піддаються еластичному спотворенню, що призводить до невеликих змін швидкості та висоти звуку.

Аугментація аудіо даних допомагає моделям виявляти різні аспекти звуку, такі як акценти, інтонації, шуми та інші особливості. Це дозволяє моделі краще узагальнювати та робити більш точні прогнози на нових аудіо даних.

**Генерація мови** Після навчання моделі RNN може бути використана для генерації мови на основі вхідного тексту. Вхідний текст пропускається через RNN, і кожен крок моделі передбачає наступну акустичну ознаку. Цей процес повторюється доти, доки не буде згенерована повна послідовність акустичних ознак. Потім акустичні ознаки можуть бути перетворені на мовний сигнал з використанням алгоритмів синтезу, таких як голосова мікшерна модель (vocoder).

### Функціонування рекурентної мережі LSTM

Рекурентні нейронні мережі (RNN) є типом нейронних мереж, які обробляють послідовності даних, зберігаючи інформацію про попередні кроки. Одним з найпопулярніших варіантів RNN є довга короткострокова пам'ять (LSTM, Long Short-Term Memory).

LSTM - це особливий тип RNN, який був розроблений для більш ефективної роботи з довгостроковими залежностями у послідовних даних. Він має здатність запам'ятовувати інформацію протягом тривалих періодів часу, що робить його особливо корисним для обробки текстів, мовлення та інших послідовностей із тривалими залежностями.

Більш детальну інформацію можна отримати з відео урока «LSTM – довга короткострокова пам'ять» <https://www.youtube.com/watch?v=wDwnC0hBJ_A>

Звичайні RNN дуже добре запам’ятовують контексти та включають їх у прогнози. Наприклад, це дозволяє RNN розпізнати, що в реченні «Хмари скупчилися на \_\_\_» слово «небі» необхідне для правильного завершення речення в цьому контексті. З іншого боку, у довгому реченні контекст складніше підтримувати. У дещо зміненому реченні «Хмари, які частково перетікають одна в одну і низько висять, знаходяться на \_\_\_ », рекурентній нейронній мережі стає значно складніше вивести слово «небо».

Проблема рекурентних нейронних мереж полягає в тому, що вони просто зберігають попередні дані у своїй «короткочасній пам’яті». Коли пам’ять у ньому вичерпується, він просто видаляє найдовше збережену інформацію та замінює її новими даними. Модель LSTM намагається уникнути цієї проблеми, зберігаючи вибрану інформацію в довготривалій пам’яті.

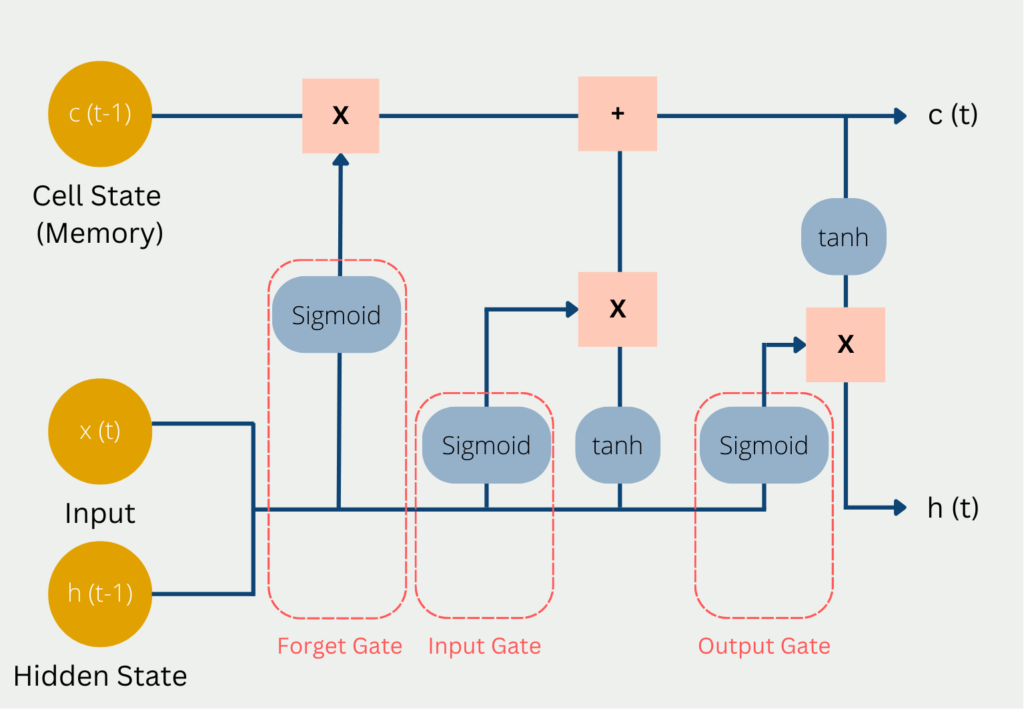


Рис. Архітектура LSTM

Ця довготривала пам'ять зберігається у "комірці пам'яті" (memory cell). Існує прихований стан, в якому зберігається короткочасна інформація з попередніх кроків обчислення. Прихований стан - це короткочасна пам'ять моделі. Це також пояснює назву довгострокових мереж.

На кожному кроці обчислення використовується поточний вхід x(t), попередній стан короткочасної пам’яті c(t-1) і попередній стан прихованого стану h(t-1).

Основними компонентами LSTM є три "гейти" (Gate): вхідний (Input Gate), забування (Forget Gate) та вихідний (Output Gate). Вхідний гейт вирішує, яка інформація буде додана в пам'ять, гейт, що забуває, контролює, яка інформація буде видалена з пам'яті, а вихідний гейт регулює, яка інформація буде передана на наступний крок.

Процес роботи LSTM:

* На кожному кроці послідовності вхідні дані подаються на вхід LSTM.
* LSTM обчислює значення гейтів на основі поточного входу та попереднього прихованого стану.
* Гейти визначають, яка інформація буде збережена в пам'яті комірки та яка буде забута.
* Пам'ять комірки оновлюється шляхом додавання нової інформації та видалення застарілої інформації.
* LSTM обчислює вихідне значення, яке передається на наступний крок і може бути використане для передбачень чи подальшої обробки.

Завдяки своїй архітектурі, LSTM здатна ефективно вловлювати довгострокові залежності даних. Вона може допомогти моделі розуміти контекст та взаємозв'язок між елементами послідовності на різних часових кроках. Це особливо корисно при роботі з текстом, де слова мають зв'язок із попередніми та наступними словами у реченні чи документі.

У рекурентній нейронній мережі проблема полягала в тому, що модель вже забула, що текст був про хмари до того часу, як вона прибула до розриву. У результаті не вдалося зробити правильний прогноз. У випадку LSTM інформаційна «хмара», швидше за все, просто опинилася б у стані комірки і, таким чином, зберігалася б протягом усіх обчислень. Дійшовши до прогалини, модель визнала б, що слово «небо» є важливим для правильного заповнення прогалини.

LSTM широко використовується в різних завданнях обробки природної мови, таких як машинний переклад, розпізнавання мови, створення тексту та аналіз настроїв. Вона також може бути використана в інших областях, де важлива обробка послідовних даних. LSTM не єдиний тип рекурентних нейронних мереж. Існують інші варіанти, такі як графічні LSTM (GRU) і прості RNN. Кожна з них має особливості та застосовується відповідно до конкретного завдання та вимог.

## 13.3. Голосова біометрія

Голосова біометрія - технологія, яка використовується для ідентифікації особи через індивідуальні фізичні, психологічні, поведінкові та деякі інші характеристики.

Основними поняттями голосової біометрії є:

* **Ідентифікація** - встановлення особи людини за голосом. Перевіряє збіг зразка голосу з багатьма відбитками з бази голосів. Як результат ідентифікації система показує список осіб зі схожими голосами в процентному відношенні. 100% збіг означає, що зразок голосу повністю збігається з голосом з бази даних і особу встановлено достовірно.
* **Верифікація** - підтвердження особи за голосом. Проводиться порівняння двох зразків голосу: голос людини, чию особу необхідно підтвердити, з голосом, який зберігається в базі даних системи і чия особа вже достовірно встановлена. Як результат верифікації система показує ступінь збігу одного голосу з іншим у процентному відношенні.

Для вирішення цих завдань технологія навчається на голосових відбитках. Маючи достовірний запис голосу користувача, з нього створюється унікальний відбиток. За спеціальним алгоритмом з характеристик його голосу обчислюється цифрова послідовність (відбиток), що в подальшому буде ідентифікатором особи. В подальшому, під час обробки голосового повідомлення, за тим же алгоритмом витягуються характеристики і обчислюється послідовність. Якщо ця послідовність певним чином збігається з голосовим відбитком, то з такою ймовірністю в цьому запису теж був голос даного користувача.

У частині аутентифікації голос можна використовувати як пароль. Людина, реєструючись в системі, читає стандартну фразу - наприклад, «Мій голос - мій пароль». Вона зачитує цю фразу тричі, і в системі зберігаються відбитки голосу людини із заздалегідь відомою фразою. І коли людина захоче залогуватися, вона не набирає пароль, а просто каже: «Мій голос - мій пароль», і система порівнює цей запис з відомими трьома.

Голос, звичайно, можна використовувати для авторизації, але це не дуже надійно. І там, де вона застосовується - в бізнесі, фінансових установах, службах безпеки - голосову біометрію використовують лише як додатковий фактор ідентифікації та авторизації. Коли запитали кодове слово, задали всі секретні питання, на всякий випадок перевіряють ще за голосом, чи це саме та людина, за кого вона себе видає.

Поки не варто цілком покладатися на голос як на ідеальний та єдиний засіб авторизації, але біометрія має всі шанси на активне застосування.

Голосова біометрія представляє відносно простий і економічний спосіб вирішення ряду практичних проблем.

* Підтвердження справжності особи за голосовим відбитком. Його паролем є його голос.
* Перевірка клієнта, чи це справді говорить людина чи транслюється запис голосу.
* В службах безпеки є загальні бази голосових відбитків зловмисників. Банки під час ідентифікації за голосом намагаються визначити особу і порівнюють записи голосів з базою пранкерів або людей, що помічені в шахрайстві чи в інших злодійських діях. Достатньо від десятьох секунд мови, щоб більш-менш точно виявити ідентичність особи. Якщо промовлено більше, то система однозначно визначиться щодо справжності.
* Голосова авторизація в додатках для мобільних телефонів спрощує доступ для власника і блокує будь які дії для сторонніх осіб.

Проблеми розпізнавання голосу за телефонним зв’язком ще залишаються. Якщо голосова команда призначена для навігатора, тоді смартфон використовується як диктофон. В нього хороший мікрофон і сигнал записується в хорошій якості.

Але якщо голос йде через телефонній зв’язок там виникають проблеми, оскільки мобільні оператори використовують алгоритми стиснення, GSM-кодеки, зрештою, якість зв'язку може бути не дуже хорошою. А якість сигналу в телефонній лінії безпосередньо впливає на якість розпізнавання голосу.

## 13.5. Загальні технологічні тренди NLP

Область обробки природної мови (Natural Language Processing, NLP) постійно розвивається, і є кілька основних технологічних трендів у цій галузі:

* Застосування глибоких нейронних мереж для вирішення завдань NLP. Моделі, такі як BERT, GPT (Generative Pre-trained Transformer) і Transformer, демонструють вражаючі результати в завданнях машинного перекладу, обробки текстів та діалогових системах.
* Одним із успішних підходів у NLP є попереднє навчання на великих корпусах текстів та донавчання на конкретних завданнях. Це дозволяє моделям вловлювати складніші мовні структури.
* Важливим завданням у NLP є розробка моделей, здатних працювати з кількома мовами та враховувати прагматичний контекст. Це допомагає моделям краще розуміти зміст та наміри користувачів.
* Аналіз тональності та емоцій у текстах стає ключовою технологією у маркетингу та зворотного зв'язку користувачів.
* Покращення алгоритмів для виявлення та зв'язування сутностей (Named Entity Recognition) та аналізу семантики тексту дозволяє глибше розуміти текстову інформацію.
* Розвиваються NLP-програми для голосових помічників, чат-ботів і розумних систем. Розробка систем, здатних підтримувати більш природні і контекстні діалоги з користувачами.
* Зростання NLP порушує питання про захист даних та етичність використання цих технологій, що призводить до розробки відповідних стандартів та регулювань.
* Зі збільшенням обсягу даних різними мовами та в різних форматах (текст, зображення, аудіо), технології NLP стають адаптивнішими до різноманітних джерел даних.
* Зростання розробки та впровадження індустріальних рішень на базі NLP для покращення бізнес-процесів та взаємодії з клієнтами.
* Розвиваються технології, що дозволяють створювати персоналізовані NLP-додатки, що враховують індивідуальні потреби та переваги користувачів.

Ці тренди свідчать про те, що NLP залишається областю, що активно розвивається, що призводить до створення більш потужних і універсальних рішень для аналізу і генерації тексту.

## Висновки

NLP (Natural Language Processing), або обробка природної мови, - це область обчислень, мета якої - допомогти комп'ютерам зрозуміти людську або «природну» мову. На даний момент хвиля інтересу до NLP тільки зростає, і у найближчому майбутньому голосова взаємодія отримає більшого поширення майже у всіх сферах діяльності.

Хоча природні мови не здаються чимось дивовижним, комп'ютерам дуже важко правильно їх інтерпретувати і використовувати. Жорсткий, обмежений правилами формат електронних таблиць і баз даних ідеальний для машин, а випадкова природа людських мов, що залежить від контексту і не завжди пов'язана певними правилами призводить штучний інтелект в ступор.

Звісно, NLP ще далекий від того, щоб змінити світ професій прямо зараз, але цей напрямок існує вже близько 30 років і роботи по його вдосконаленню ведуться постійно. Експерти вважають, що наступний прорив у розвитку NLP буде колосальний, зумовить перехід від структурованих (бази даних) до неструктурованих (текст) даних і значно поліпшить здатність машин «розуміти» людей в звичайній розмові. Пристрої, що здатні розпізнавати голос і генерувати його, стрімко дешевшають і вдосконалюються.

## Контрольні питання

1. Які особливості має людська мова?
2. Назвати основні напрямки досліджень в NLP. Які напрямки формують мовні технології?
3. Назвати основні сфери застосування машинного розпізнавання мови.
4. Перелічити поширені методи машинного розпізнавання мови.
5. Які існують підходи до синтезу мови.
6. Назвати основні етапи для статистичного методу синтезу мови
7. Яку користь привносить застосування нейронних мереж для розпізнавання та синтезу мови.
8. Перелічити функції, що виконує голосова біометрія.
9. Назвати основні складові голосових інтерфейсів.
10. Перелічити технологічні тренди розвитку NLP.

## Використані джерела

1. Історія обробки природньої мови <https://habr.com/ru/post/481228/>
2. NLP-технології розпізнавання людського мовлення <https://evergreens.com.ua/ua/articles/natural-language-processing.html>
3. Обробка природньої мови, розпізнавання та синтез мовлення <https://singapore-academy.org/libcdo/100.pdf>
4. Мовні технології <https://www.tadviser.ru/index.php/Речевые_технологии>
5. Мовні технології в бізнесі <https://hightech.fm/2020/05/21/chatbot-bussines>
6. Сучасний ринок мовних технологій <https://www.controlengrussia.com/innovatsii/rechevye-tehnologii/>
7. Розпізнавання мови <https://habr.com/ru/company/toshibarus/blog/490732/>
8. Машинне навчання це весело! Частина 6: Розпізнавання мови за допомогою глибокого навчання <https://medium.com/@ppleskov/машинное-обучение-это-весело-часть-6-aa3982f25d54>
9. 8 головних досягнень в нейромережному NLP <https://sysblok.ru/nlp/8-glavnyh-proryvov-v-nejrosetevom-nlp/>
10. 5 методів обробки природної мови <https://neurohive.io/ru/osnovy-data-science/5-metodov-v-nlp-kotorye-izmenjat-obshhenie-v-budushhem/>
11. <https://databasecamp.de/en/ml/lstms>