Глибинні нейронні мережі в задачах розпізнавання мови

Науковий керівник: к.т.н., доцент Тимощук О. Л.

Виконав: студент 6 курсу група КА-41м Квітко М.В.

Постановка задачі

ОБ’ЄКТ ДОСЛІДЖЕННЯ:

- Проблеми та цілі голосової аналітики та задач розпізнавання мовлення

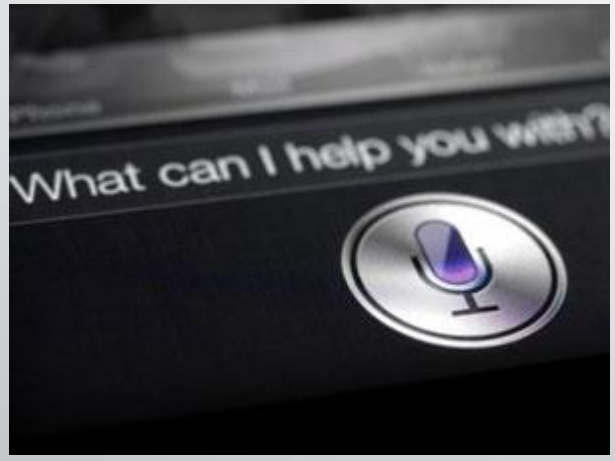
ПРЕДМЕТ ДОСЛІДЖЕННЯ:

- Застосування методів на основі глибинних нейронних мереж для розпізнавання мовлення

ПОСТАНОВКА ЗАДАЧІ:

- Дослідити нейронні мережі, як складову у задачах розпізнавання
- Провести порівняльний аналіз існуючих підходів до розпізнавання мовлення, визначити їх особливості
- Глибинні нейронні мережі як сучасний апарат для задач розпізнавання мовлення
- Проаналізувати алгоритм розпізнавання мовлення на основі рекурентних нейронних мереж.
- За допомогою програмного продукту проаналізувати роботу запропонованого алгоритму розпізнавання у порівнянні з існуючими алгоритмами

Актуальність



- Сучасний світ прямую до повної автоматизації більшості процесів своєї життєдіяльності. Проте однією з основних проблем є розпізнавання мовлення для передачі команд різним пристроям, відтворення промовленого у вигляді тексту тощо. Проте точність розпізнавання ще не досягла свого максимуму і кожного дня намагаються її збільшити.
- Глибинні нейронні мережі (ГНМ) в останні роки неодноразово показували суттєві результати у процесах прогнозування, класифікації, розпізнавання образів, рукописного тексту та мовлення. Тому використання ГНМ та їх модифікації у задачах розпізнавання мовлення – є актуальною задачею сьогодення.

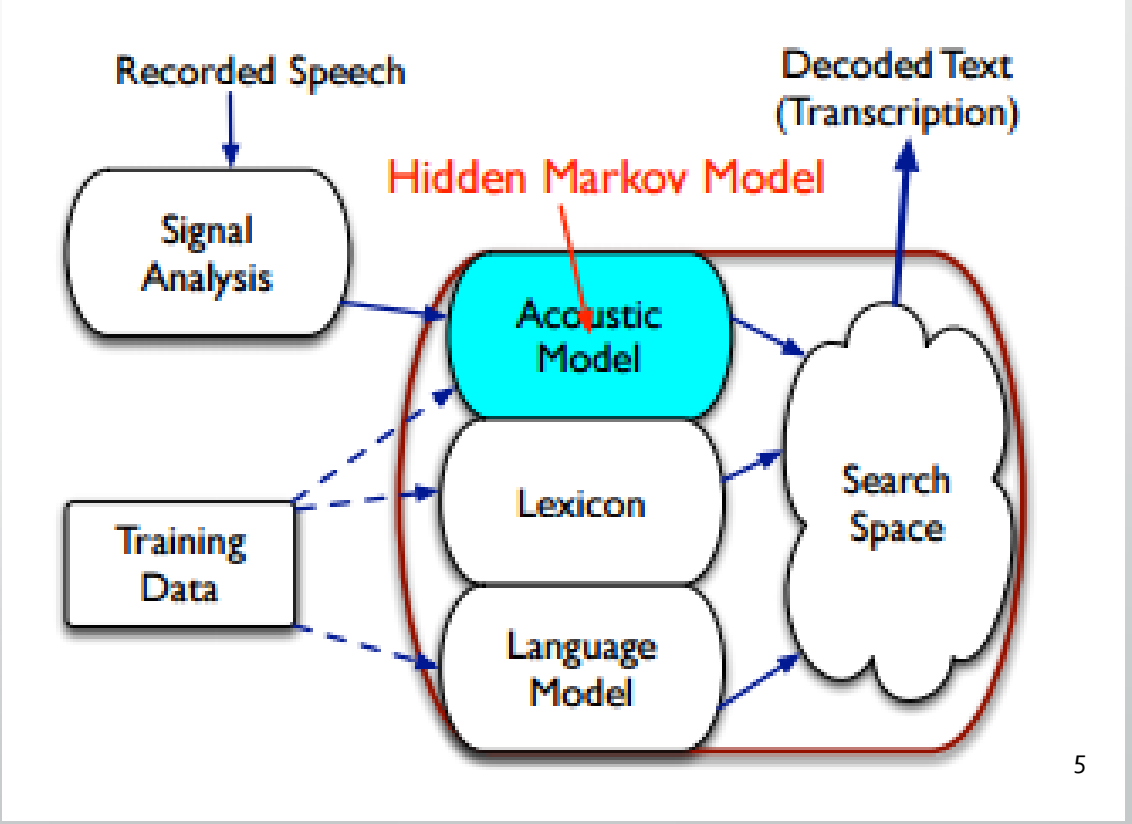
Розпізнавання мови та нейронні мережі

- На сьогоднішній день системи Speech-To-Text, керування голосом, голосова автентифікація використовуються усе частіше.
- В основі процесу виділення звуків, літер, побудови слів та речень лежать алгоритми класифікації та прогнозування. Починаючи з 1970-х років саме нейронні мережі з кожним роком показують усе кращі результати в подібних задачах.

Yandex SpeechKit

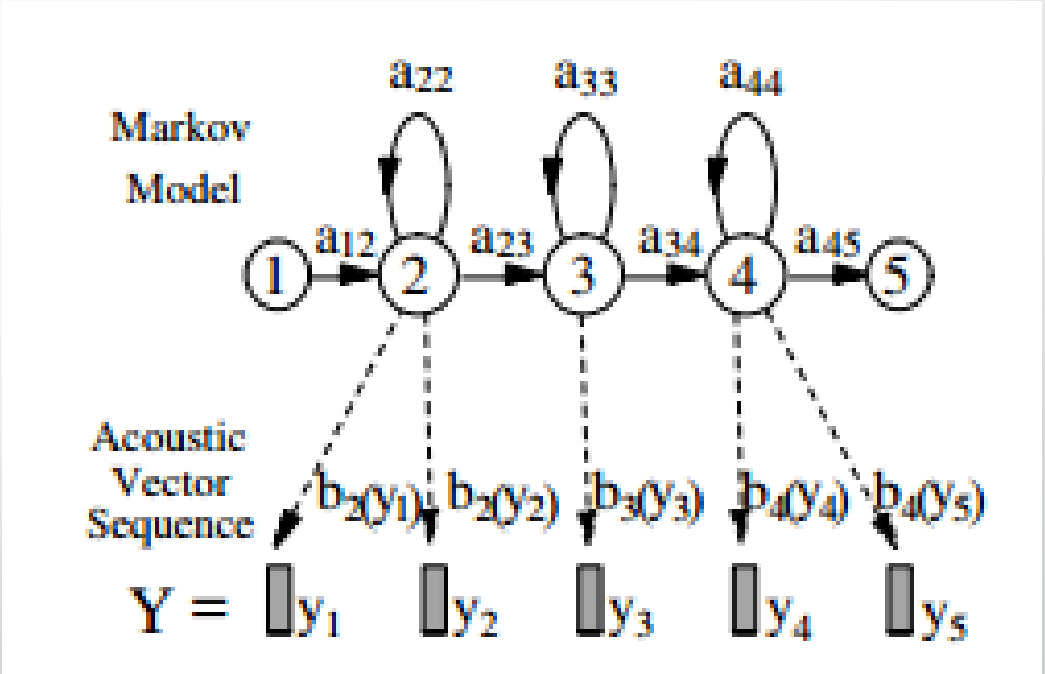
Існуючі підходи до розпізнавання мови: HMM

- Попередні 30 років лідером у задачах мовлення вважались моделі, побудовані на основі прихованих ланцюгів Маркова (HMM) та Gaussian mixture model (GMM).



Існуючі підходи до розпізнавання мови: HMM

- Записаний звук ділиться на короткі (10 мс) фрагменти, які аналізуються на вміст частот. Отриманий в результаті вектор характеристик пропускається через акустичну модель, яка видає набір імовірнісних розподілів серед всіх можливих фонем. HMM допомагає виявити послідовні структури в цьому наборі розподілів ймовірностей



Переваги моделей на основі HMM

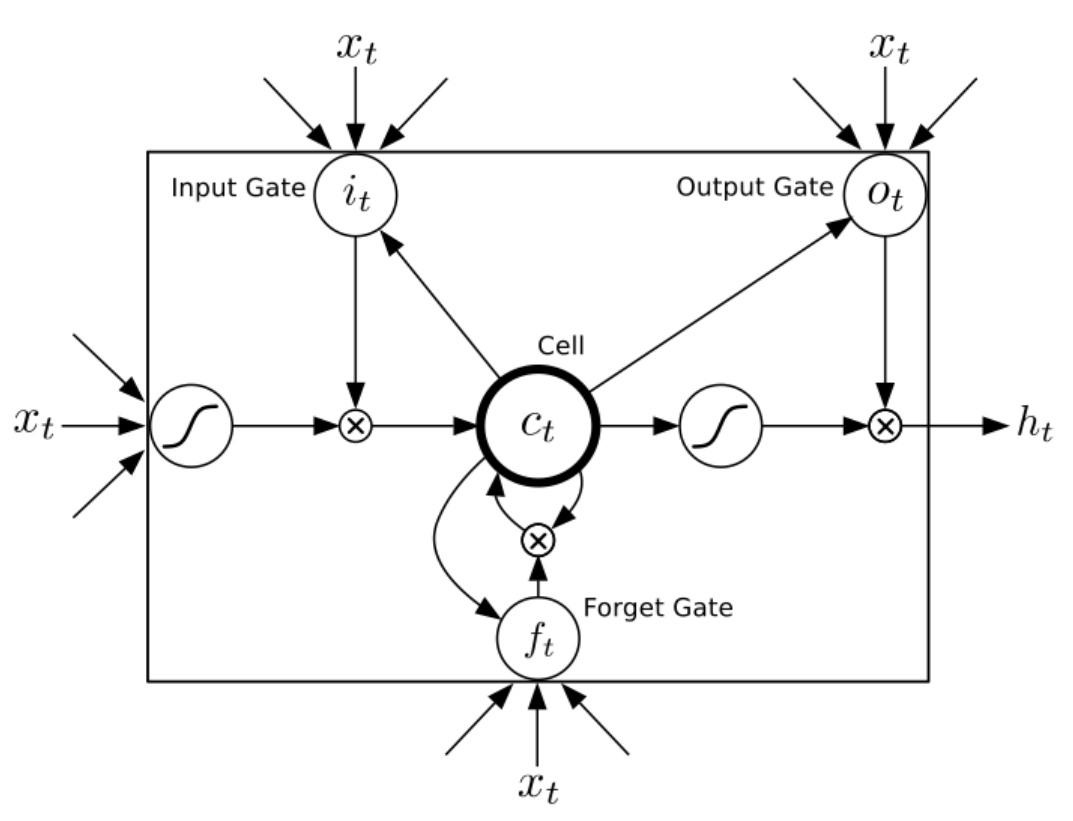
ПЕРЕВАГИ

- Аналітичне вирішення проблеми розпізнавання
- Можливість розпізнавати слова, що складаються з набору букв позбавлених конкретного смислового значення
- Прості в реалізації та навчанні

НЕДОЛІКИ

- Досить низька точність
- Погана робота в умовах шуму

Існуючі підходи до розпізнавання мови: RNN



- В останні 3-4 роки широкого використання набувають методи на основі рекурентних нейронних мереж (RNN)
- Для побудови необхідної акустичної моделі з метою виділення фонем використовуються комірки довгої коротко-тривалої пам’яті (LSTM) у складі стандартних RNN

Переваги та недоліки моделей на основі RNN

ПЕРЕВАГИ

- Вища швидкість роботи
- Точність розпізнавання більше
- Краще працюють в умовах підвищеного шуму
- Добре працює в умовах неточності та незавершеності промовлених слів

НЕДОЛІКИ

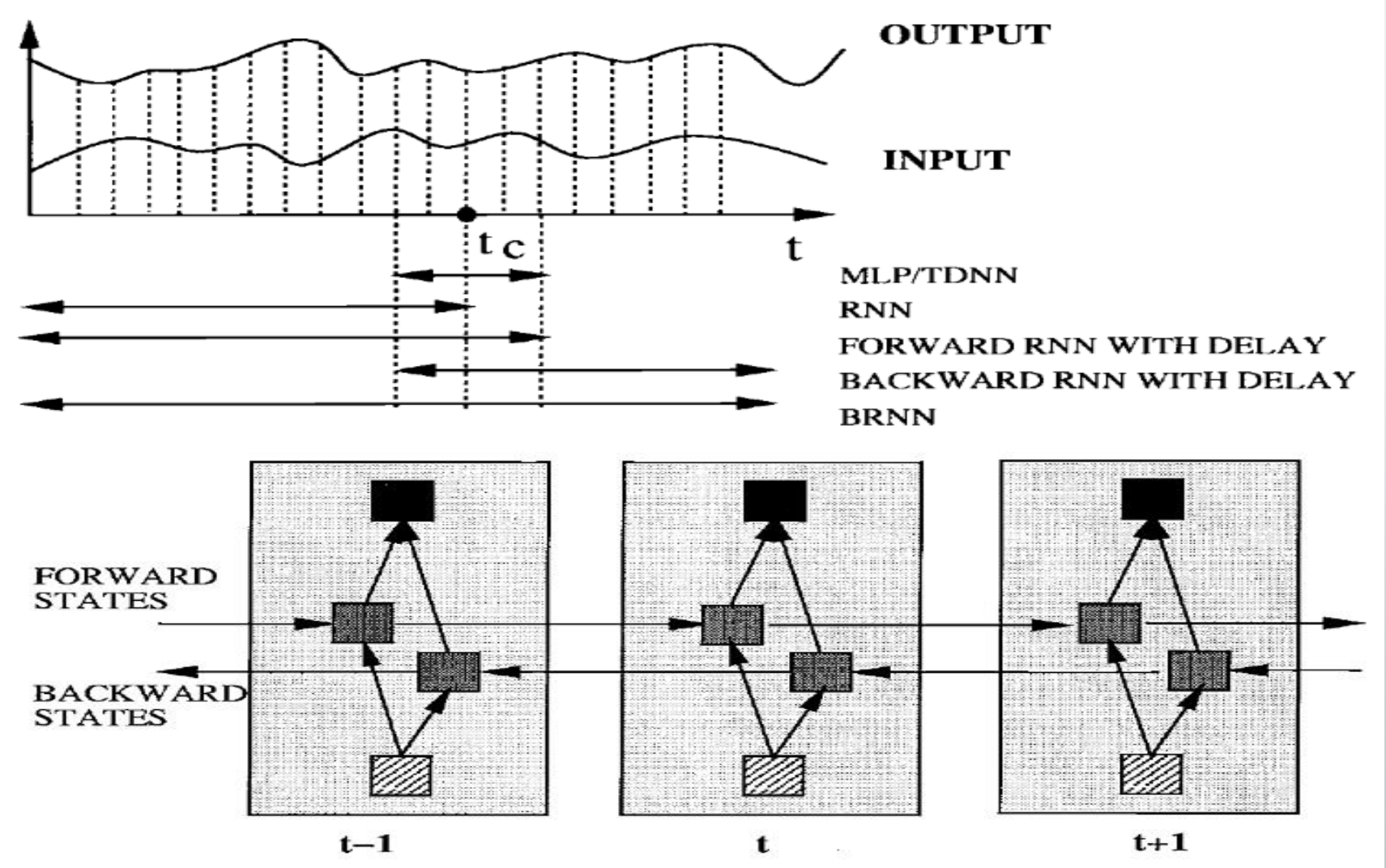
- Вимагає великих обчислювальні потужності
- Необхідна велика кількість прикладів для навчання
- Багато часу для навчання

Алгоритм розпізнавання мови за допомогою RNN

- За основу беруться рекурентні нейронні мережі, але їх стандартна структурна не підходить до задач розпізнання мови, оскільки при розрахунках відхилень та ваг малі помилки дуже швидко зникають, а великі дуже швидко зростають через велику кількість множень.
- Використовуються LSTM комірки. Комірка LSTM описується наступним рівнянням: $c_t = f_t c_{t-1} + i_t \tanh(W_{hc} h_{t-1} + W_{xc} x_t + b_c)$
- Таким чином коли $f_t = 0$ комірка повністю забуває старе значення і запам'ятовує нове. Коли $i_t = 1$ встановлюється значення на основі вхідних даних та попереднього прихованого шару, якщо інакше то або залишається незмінним($f_t = 1$), або стирається ($f_t = 0$)

$$
\begin{aligned}
f_t &= \sigma(W_{xf} x_t + W_{hf} h_{t-1} + W_{cf} c_{t-1} + b_f) \\
i_t &= \sigma(W_{xi} x_t + W_{hi} h_{t-1} + W_{ci} c_{t-1} + b_i) \\
o_t &= \sigma(W_{xo} x_t + W_{ho} h_{t-1} + W_{co} c_{t-1} + b_o)
\end{aligned}
$$

Алгоритм розпізнавання мови за допомогою RNN

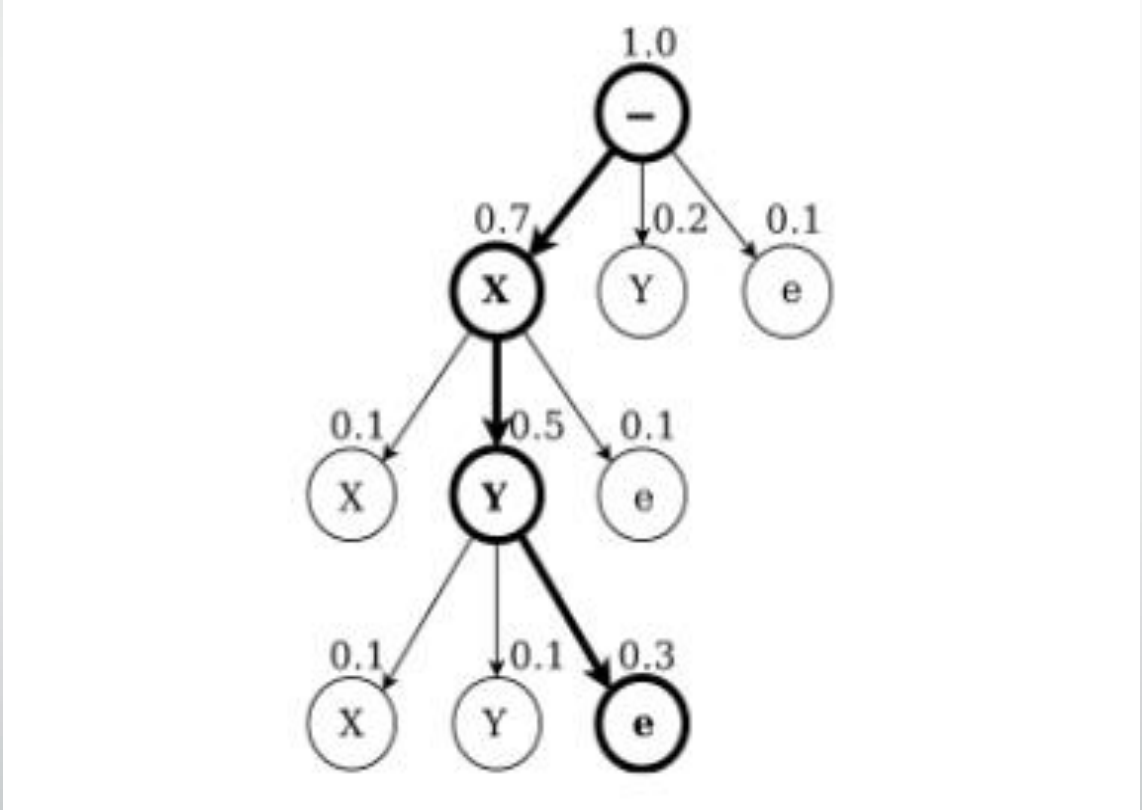


Алгоритм розпізнавання мови за допомогою RNN

- Наступним кроком ми маємо навчити акустичну модель, для чого ми використовуємо **Connectionist Temporal Classification (CTC)**
- Ключовою ідеєю CTC є те, що замість того, щоб якимось чином генерувати стрічку як виведення з нейронної мережі, ми замість того генеруємо розподіл імовірності на кожному часовому кроці (від t = 1 до t = T). Потім ми можемо розшифрувати цей розподіл ймовірностей за принципом максимальної правдоподібності і знайти шукану стрічку. І, нарешті, ми навчаємо нашу мережу, створивши цільову функцію, яка шукає максимальну ймовірність декодування для заданої послідовності x, щоб відповідати нашій бажаній ℓ мітці.
- Після виведення розподілу ймовірності ми декодуємо отриманні розподіли за певним алгоритмом (а саме алгоритм пошуку найкращого префіксу)
- Навчаємо мережу CTC мінімізуючи наступну функцію : $\mathcal{O}(S) = -\sum_{(x,\ell)\in S} \ln P(\ell|x)$

  Для максимізації кожного навчального прикладу $S = \{(x, \ell)\}$

Алгоритм декодування за пошуком найкращого префіксу



Параметри моделі

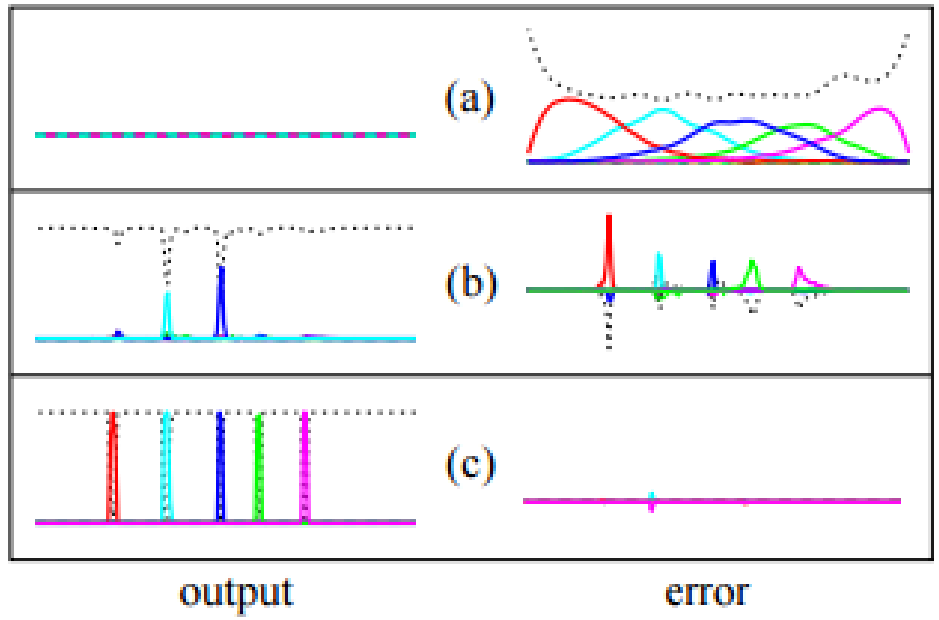
- 100 блоків LSTM в прихованих шарах прямого і зворотного поширення
- Активаційна функція – гіперболічний тангенс для вхідних і вихідних комірок
- Активаційна функція – сигмоїд для воріт комірок LSTM
- Приховані шари пов’язані самі з собою, з виходами і входами
- Вхідний шар складається з 26 нейронів, вихідний SoftMax шар складається з 62 юнітів (61 фонема і 1 вихід для пробілу)
- Загальна кількість ваг - 114, 662.

Параметри моделі

- Навчання проводилося методом зворотного поширення помилки і поточного градієнту (ваги оновлюються після кожного етапу тренування)
- Похибка навчання 10^{-4}.
- Значення активаційних функцій мережі скидаються на 0 для кожного етапу тренування.
- Для методу пошуку найкращого префіксу необхідна ймовірність пробілу встановлена на рівні 99,99%.
- Ваги ініціалізуються рівномірним розподілом з межами (-0,1 ; 0,1)
- Під час навчання додається Гаусівський шум із стандартним відхиленням 0,6 для поліпшення узагальненості навчання.

Результати роботи програмного продукту. Порівняння роботи існуючих методів із запропонованим

Система	Міра помилки, %
Контекстно незалежна ПММ	38.85
Контестно залежна ПММ	35.21
BLSTM/HMM	33.84
BLSTM/HMM із зваженими помилками	31.57
HTK (метод пошуку кращого шляху)	31.47
HTK (метод пошуку кращого префіксу)	30.51



Порівняння роботи існуючих методів із запропонованим

Система	Чистий звук (94 прикл.)	Шум (82 прикл.)	Комбіновано (176 прикл.)
Bing Speech	11.73	36.12	22.05
Google Speech	6.64	30.47	16.72
Wit.ai	7.94	35.06	19.41
Запропонована система	6.56	29.06	18.85

Запропонована модифікація алгоритму

$$h(t,u) = \exp(f_t + g_u)$$

- Заміна обрахунку експоненти суми двох векторів на простий персептрон із одним прихованим шаром

Шляхи розвитку

- Збільшення навчальної вибірки.
- Покращення первинної обробки сигналу для виділення саме частоти людської мови.
- Використання історії розпізнавання для підвищення швидкості розпізнавання та адаптації до диктора.
- Підбір оптимальних параметрів мережі та методу первинної ініціалізації ваг для збільшення швидкості та точності навчання мережі.

Висновки

- Під час виконання даної дисертаційної роботи було розглянуто глибинні нейронні мережі як основний інструмент для розпізнавання мовлення.
- Було проаналізовано переваги і недоліки старих моделей на основі HMM і сучасні моделі на основі РНН із модифікаціє структури LSTM комірками і використанням CTC моделі для класифікації звуків у фонеми. Результатом їх роботи є надзвичайно точні показники, а також гарні результати в умовах підвищеного шуму.
- За допомогою реалізованого програмного продукту було проведене порівняння існуючих систем розпізнавання мовлення із запропонованою. Також запропонована оптимізація моделі для зменшення похибки відкидання літер.

ДЯКУЮ ЗА УВАГУ

- Квитко М.В. Распознавание речи с помощью глубоких рекуррентных нейронных сетей [Електронний ресурс] / Квитко М.В. // IASA – 2016 р. – 223 стр. – Режим доступу: http://sait.kpi.ua/media/filer_public/73/32/7332a68e-e93b-4c57-a3c8-66f11ee074cd/sait2016ebook.pdf