УДК 004.934:004.85

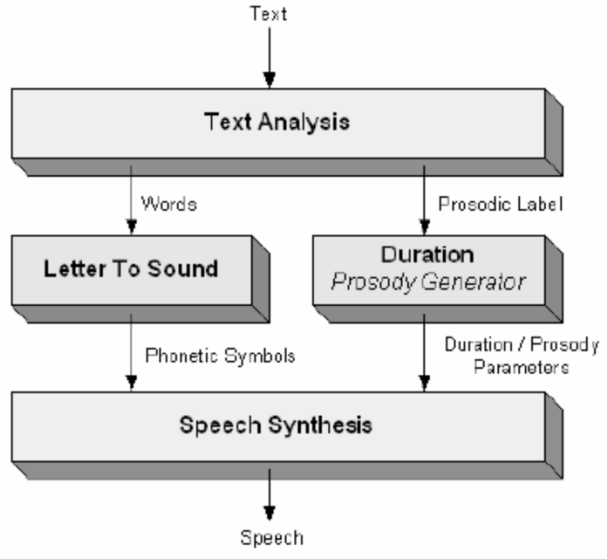
СИНТЕЗ МОВИ З ВИКОРИСТАННЯМ ГЛИБОКОГО НАВЧАННЯ У MATLAB МОДЕЛЮВАННЯ СИСТЕМИ TEXT-TO-SPEECH

Ігнатюк І.В.
e-mail: ivan.ihnatiuk@nure.ua
Науковий керівник – ст. викладач Чумак В. С.
Харківський національний університет радіоелектроніки, каф. МТС
м. Харків, Україна

This work examines the implementation of a thought synthesis system based on deep knowledge, using MATLAB. The main idea is to model the Text-to-Speech (TTS) system using a variety of neural boundary methods to ensure high accuracy of the generated speech.

Нейронні мережі на стають все більш популярними завдяки їхнім унікальним можливостям прискорення обробки даних та виконання складних алгоритмів в реальному часі. Ця тенденція викликана потребою у військових додатках у здатності швидко адаптуватися до змінних ситуацій на полі бою, а також в усуненні потреби у постійному поновленні апаратного забезпечення для підтримки нових функціональних можливостей.

Система TTS складається з двох основних компонентів: акустичної моделі та вокодера. Акустична модель перетворює вхідний текст у набір акустичних ознак, а вокодер здійснює генерацію звукового сигналу. У MATLAB реалізовано підтримку різних архітектур нейронних мереж, зокрема, згорткових (CNN) і рекурентних (RNN) мереж, які можуть бути використані для створення акустичної моделі. Архітектура системи:



Для навчання акустичної моделі використовуються глибокі нейронні мережі, що навчаються на великому корпусі мовних даних. MATLAB забезпечує зручні інструменти для роботи з такими мережами через Deep Learning Toolbox, що містить готові архітектури для аналізу та генерації мовлення.

Основні етапи реалізації:

1. Предобробка тексту: токенізація, нормалізація та транслітерація.
2. Екстракція акустичних ознак: використання Mel-спектрограм та інших параметрів.
3. Навчання нейромережевої моделі: використання LSTM або Transformer-моделей.
4. Синтез мовлення: генерація аудіофайлу через вокодер (WaveNet, Griffin-Lim).

Використання LSTM або Transformer-мереж дозволяє ефективно моделювати залежності в послідовностях, що критично для відтворення інтонацій. Для покращення якості синтезу використовуються вокодери, такі як WaveNet і Griffin-Lim, які зменшують артефакти та забезпечують більш природний звук. Формально процес синтезу можна описати як:

$$\gamma = V(f(T))$$

де T – вхідний текст, f – функція перетворення тексту в акустичні ознаки, V – вокодер, що відтворює мовний сигнал.

Реалізація TTS у MATLAB дозволяє ефективно розробляти та тестувати нейромережеві моделі для синтезу мовлення. Використання глибокого навчання значно підвищує якість синтезованого звуку, забезпечуючи природність та зрозумілість мовлення.

Список використаних джерел:

1. Чумак В.С. Використання нейронних мереж в адаптивних системах онлайн-медичної освіти на базі мікроконтролерів STM32 в умовах воєнних криз // Автоматизація, електроніка та робототехніка. Стратегії розвитку та інноваційні технології (AERT-2023): матеріали V форуму, 29–30 листопада 2023 р. – Харків: ХНУРЕ, 2023. – С. 134-135.

2. Чумак В. С. Інтеграція нейронних мереж у медичні пристрої на основі STM32 для автоматичної діагностики та моніторингу пацієнтів / В. С. Чумак // Автоматизація, електроніка та робототехніка. Стратегії розвитку та інноваційні технології (AERT-2023): матеріали V форуму, 29–30 листопада 2023 р. – Харків : ХНУРЕ, 2023. – С. 132-133.

3. Луценко О. В. Використання FPGA для реалізації штучної нейронної мережі / О. В. Луценко, В. С. Чумак // Автоматизація, електроніка та робототехніка. Стратегії розвитку та інноваційні технології : матеріали IV форуму, 24–25 листопада 2022 р. – Харків : ХНУРЕ, 2022. – С. 26-27.

4. Малахова О. Ю. Електроміограф на FPGA / О. Ю. Малахова, І.О. Шевцов // Авіація, промисловість, суспільство : матеріали III Міжнар. наук.-практ. конф. (м. Кременчук, 12 трав. 2022 р.) / МВС України, Харків. нац. ун-т внутр. справ, Кременчуц. льотний коледж., Наук.парк «Наука та безпека». – Харків : ХНУВС, 2022. – С. 117 -120.