ОЗВУЧУВАННЯ ТЕКСТІВ З КОПІЮВАННЯМ ГОЛОСУ У РЕАЛЬНОМУ ЧАСІ ДЛЯ МОВ З МАЛОЮ КІЛЬКІСТЮ РЕСУРСІВ

Максименко Д.В.
Науковий керівник - доцент Турута О.П.
Харківський національний університет радіоелектроніки
(61166, Харків, просп. Науки, 14, каф. Програмної інженерії,
тел. (057) 702-14-46), e-mail: daniil.maksymenko@nure.ua

With the growing demand on audio content people need some technology to make its production faster and for it to still be human-like even if generated by a computer. Also, voicing algorithms have to learn to talk with different voices without a long preparation. Finally their main problem is the high difficulty of learning new languages, especially if the language has low resources. So it is necessary to create a technique to voice text with passed voice parameters and ability to study new phonemes with small datasets. This thesis provides a possible solution for this problem.

Останнім часом кількість контенту, що потребує озвучування постійно зростає. Голосові асистенти мають розмовляти більш подібно до людини, комп'ютерні ігри ростуть у масштабах і потребують великої кількості озвучених реплік. Дублювання фільмів, ігор, аудіо книг на нові мови також потребує часу та грошей і винаймання цілої команди може не окупитись для певних продуктів. Медіа також намагаються освоїти нову площину і вийти в аудіо формат, тож деякі додають машинну озвучку найбільш цікавих статей. Люди з вадами зору також мають отримувати контент на рівні з іншими, а отже всі ці потреби вимагають від алгоритмів автоматичного озвучування давати кращу якість і більшу різноманітність голосів. Головні проблеми з нинішніми TTS системами - це їх роботизованість та підтримка лише одного чи декількох голосів. Старі рішення без використання машинного навчання не здатні надати рішення наведених задач. При цьому алгоритми машинного навчання важко натренувати розмовляти на мові з малою кількістю ресурсів, а таких у світі більшість. Авжеж можна знайти великі масиви даних для англійської чи китайської, але навіть для української мови датасетів уже значно менше, що ускладнює розробку рішень на базі нейронних мереж.

Архітектур для копіювання параметрів голосу та озвучування текстів з їх використанням існує не так багато. Перша - MelGanVC. Ця архітектура призначена для генерації спектрограм з копіюванням стилю іншого аудіо. Її можна використовувати як для генерації музики, так і для генерації озвученого тексту з переносом голосу з певного спікера. Проблемою цього рішення є потреба у великій кількості аудіо записаного певним голосом для подальшого копіювання заданих особливостей. Друга архітектура дозволяє копіювати голос з п'яти секунд запису і дає майже таку саму якість озвучування як попереднє рішення. Це називається SV2TTS або

Speaker Verification to Text to Speech. Ця архітектура складається з 3 нейронних мереж, де перша визначає параметри голосів, намагаючись верифікувати різних людей. Друга - синтезатор спектрогам без даних про фазу, а третя - утворює звук із згенерованих спектрограм та очищує вихідний сигнал.

Архітектура SV2TTS добре вирішує проблеми різноманіття спікерів та зачитування тексту взагалі, проте вона потребує чималої кількості даних для адаптації під нову мову. Якщо кодувальник голосу та вихідну мережу можна перенести без великих змін, то синтезатор має отримати уявлення про символи та фонеми обраної мови. Є декілька підходів: навчати синтезатор з нуля на більш ніж 1000 годинах аудіо та більше 500 спікерів як робили для китайської моделі або донавчити англомовну модель для мови, яка схожа з англійською. Перший підхід не працює для мов з малою кількістю ресурсів, якою є українська. Тому задачу навчання SV2TTS під українську було вирішено розв'язувати завдяки перекладу українських фонем на англійські символи. Таким чином можна перенести досвід англомовної моделі і використовуючи невеликий датасет, як для цієї задачі, навчити мережу генерувати україномовне аудіо.

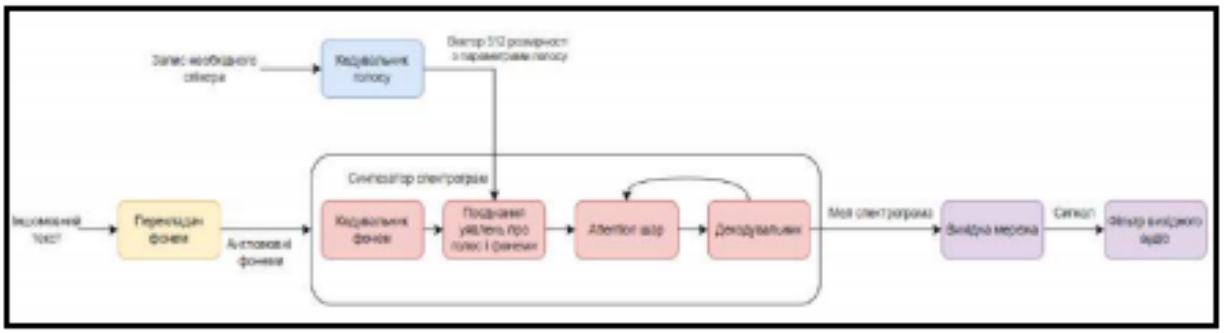

Рисунок 1 - Схема роботи алгоритму

Тож, на виході має бути отриманий алгоритм для спрощеного навчання SV2TTS архітектури для нових мов. Завдяки перекладу українських текстів на англійські фонеми можна отримати україномовну модель. На її основі далі можна навчати моделі для інших мов східної Європи з ще меншою кількістю ресурсів. Наприклад, створити білоруську модель, бо схожість цих мов дозволяє компенсувати нестачу даних для більш масштабного тренування з нуля. Також якщо розподілити аудіо для навчання синтезатора по мовцям, то можна отримати датасет для тренування кодувальника голосу під акценти та особливості мовлення певної національності. На виході архітектури варто також додати фільтр сигналу, щоб прибрати певні артефакти генерації озвучування. Для того, щоб позбутись їх повністю, варто збільшити кількість даних для тренування синтезатора, адже чимала кількість поміх та дефектів з'являється саме на момент створення спектрограми.