МРНТИ 28.23.17
УДК 004.82:16

https://doi.org/10.51889/9340.2022.21.68.023

Ж.М. Кожирбаев[1], Ж.А. Есенбаев[1]*

[1]National Laboratory Astana, г.Нур-Султан, Казахстан
[*]e-mail: zhanibek.kozhirbayev@nu.edu.kz

ИНТЕГРАЛЬНЫЙ (END-TO-END) СИНТЕЗ РЕЧИ ДЛЯ КАЗАХСКОГО ЯЗЫКА

Аннотация

Синтез речи, также называемый преобразованием текста в речь (TTS), считается одной из важных задач обработки речи наряду с распознаванием речи. Это способ преобразования данного текста в речь. Существует несколько подходов синтеза речи. В 20 веке была разработана первая компьютерная система синтеза голоса. Некоторыми из ранних методов компьютерного синтеза речи являются артикуляционный синтез, формантный синтез и конкатенативный синтез. Статистический параметрический синтез речи позже был предложен по мере развития машинного обучения. С 2010-х годов синтез речи на основе нейронных сетей постепенно становится все более популярным и улучшает качество голоса. Целью данной работы является обзор статистических параметрических и сквозных методов, которые можно рассматривать как линию эволюционного развития TTS. Кроме того, мы проведем эксперимент со сквозным методом на базе Tacotron2 и ParalleWavegan. Для экспериментов были собраны текстовые материалы произведений Ахмета Байтурсынулы. Всего из собранных материалов было записано 50 часов аудиозаписи. Из произведений Байтурсынулы было отобрано шесть книг, из которых были отобраны наиболее распространенные произведения и собраны в аудиотекстовые материалы. Один профессиональный диктор-мужчина озвучивал собранные текстовые данные.**Ключевые слова:** синтез речи, формантный синтез речи, конкатенативный синтез речи, статистический параметрический синтез речи, интегральный синтез речи.

Аңдатпа

Ж.М. Кожирбаев[1], Ж.А. Есенбаев[1]
National Laboratory Astana, Нұр-Сұлтан қ., Қазақстан

ҚАЗАҚ ТІЛІ ҮШІН ИНТЕГРАЛДЫҚ (END-TO-END) СӨЙЛЕУ СИНТЕЗІ

Сөйлеу синтезі, оны мәтіннен сөйлеуге (TTS) деп те атайды, сөйлеуді танумен қатар сөйлеуді өңдеудің маңызды міндеттерінің бірі болып саналады. Бұл берілген мәтінді сөйлеуге түрлендіру тәсілі. Сөйлеу синтезінің бірнеше тәсілдері бар. 20 ғасырда бірінші компьютерлік сөйлеу синтезі жүйесі жасалды. Компьютерлік сөйлеу синтезінің алғашқы әдістерінің кейбірі артикуляциялық синтез, формант синтезі және конкатенативті синтез болып табылады. Машиналық оқыту дамыған сайын статистикалық параметрлік сөйлеу синтезі ұсынылды. 2010 жылдардан бастап нейрондық желіге негізделген сөйлеу синтезі біртіндеп танымал бола бастады және сөйлеу сапасын жақсартады. Бұл жұмыстың мақсаты статистикалық параметрлік және түпкілікті әдістерді қарастыру болып табылады, оларды TTS эволюциялық даму желісі ретінде қарастыруға болады. Сонымен қатар, біз Tacotron2 және ParalleWavegan негізіндегі әдіспен тәжірибе жасаймыз. Эксперимент үшін Ахмет Байтұрсынұлының шығармаларынан мәтіндік материалдар жинақталды. Жиналған материалдардан барлығы 50 сағат аудиожазба жазылды. Байтұрсынұлының шығармаларынан алты кітап таңдалып, олардың ішінен ең көп таралған шығармалар таңдалып, аудиомәтіндік материалдарға жинақталды. Бір кәсіби ер диктор жиналған мәтіндік деректерді оқыды.

Түйін сөздер: сөйлеу синтезі, формантты сөйлеу синтезі, конкатенативті сөйлеу синтезі, статистикалық параметрлік сөйлеу синтезі, интегралды сөйлеу синтезі.

Abstract

END-TO-END SPEECH SYNTHESIS FOR THE KAZAKH LANGUAGE

Kozhirbayev Zh.M.[1], Yessenbayev Zh.A.[1]
National Laboratory Astana, Nur-Sultan, Kazakhstan

Speech synthesis, also called text-to-speech (TTS), is considered one of the important tasks of speech processing along with speech recognition. It is a way of converting given text to speech. There are several approaches to speech synthesis. In the 20th century, the first computer voice synthesis system was developed. Some of the early computer speech synthesis methods are articulatory synthesis, formant synthesis, and concatenative synthesis. Statistical parametric speech synthesis was later proposed as machine learning developed. Since the 2010s, neural network-based speech synthesis has gradually

become more popular and improves voice quality. The purpose of this work is to review statistical parametric and end-to-end methods, which can be considered as a line of evolutionary development of TTS. In addition, we will experiment with an end-to-end method based on Tacotron2 and ParalleWavegan. For the experiments, textual materials from the works of Akhmet Baitursynuly were collected. In total, 50 hours of audio recording were recorded from the collected materials. From Baitursynuly's works, six books were selected, from which the most common works were selected and collected in audio text materials. One professional male announcer voiced the collected text data.

Keywords: speech synthesis, formant speech synthesis, concatenative speech synthesis, statistical parametric speech synthesis, integral speech synthesis.

Введение

Синтез речи направлен на создание естественной и человеческой речи из входного текста. Он имеет широкий спектр приложений для взаимодействия человека с машиной, включая голосовую навигацию, широковещательную передачу информации, интеллектуальных помощников и голосовых операторов в колл-центра, и принес значительные экономические выгоды. Кроме того, он используется в различных новых дисциплинах, включая аудиокниги, изучение новых языков и реабилитационное лечение. Приложения TTS стали неотъемлемой частью повседневной жизни людей. Разработка системы синтеза речи включает в себя несколько дисциплин, таких как лингвистика, акустика, цифровая обработка сигналов и машинное обучение.

Формантный, конкатенативный и статистический параметрический методы требуют большого профессионального опыта и значительного времени. Необходимость упростить системы синтеза речи и устранить потребность в ручном вмешательстве и лингвистическом опыте стимулирует интегральные (end-to-end) подходы. Интегральная модель может напрямую синтезировать сигнал из заданного текста после обучения достаточного набора данных. Самые современные интегральные модели TTS, основанные на глубоком обучении, продемонстрировали способность синтезировать речь, почти человеческую.

Методология

Формантный синтез речи. Синтез речи на основе формант, основанный на модели речи «источник-фильтр», вероятно, был наиболее часто используемым подходом к синтезу на первых этапах эволюции синтеза речи. Он генерирует речь, используя правила, которые регулируют простую модель исходного фильтра [1, 2]. Лингвисты часто создают эти правила, чтобы максимально точно соответствовать формантной структуре и другим спектральным характеристикам речи. Для генерации речи используется модуль аддитивного синтеза и акустическая модель с настраиваемыми характеристиками, такими как основная частота, тембр голоса и уровень шума. Формантный синтез может создавать очень понятную речь с небольшими вычислительными ресурсами, что делает его идеальным для встроенных устройств. Искусственная речь имеет сбои и звучит менее реалистично. Кроме того, определение правил синтеза является сложной задачей. Параметры формантного фильтра можно вычислить следующим образом:

$$y(n) = ax(n) - by(n-1) - cy(n-2)\,, \tag{1}$$

$$a = 1 + b + c\,, \tag{2}$$

$$b = -2exp(-\pi BT_s)cos(2\pi FT_s)\,, \tag{3}$$

$$c = exp(-2\pi BT_s)\,, \tag{4}$$

где B - пропускная способность,

F - резонансная частота фильтра,

T - частота дискретизации.

Несколько фильтров с F_1, F_2, F_3… развернуты в диапазоне формант. Вся система будет аналогична передаточной характеристике голосового тракта в виде резонансной частоты. Каскадная и параллельная – два основных типа архитектур. Выход каждого формантного резонатора в каскадном формантном синтезаторе речи (рис.1) подается на вход следующего за ним резонатора. Это устройство состоит из последовательно соединенных полосовых резонаторов. Для управляющей информации требуются только формантные частоты. Основным преимуществом каскадной архитектуры является

устранение потребности в индивидуальном контроле относительных амплитуд формант для гласных [3]. Было доказано, что каскадная структура более эффективна для не носовых звуков речи. К тому же его проще реализовать. Тем не менее, производство фрикативных звучания представляет собой трудность при каскадном подходе.

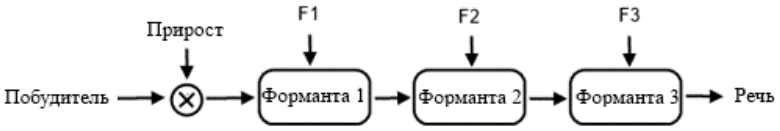

Рисунок 1. Каскадный формантный синтез речи

Резонаторы параллельного формантного синтезатора речи (рис. 2) соединены друг с другом. Дополнительные резонаторы иногда используются для носовых звучания. Все форманты получают сигнал возбуждения одновременно, а затем выходные сигналы объединяются. Параллельная конструкция позволяет отдельно регулировать полосу пропускания и усиление для каждой форманты. Следовательно, требуется больше управляющей информации. Хотя было обнаружено, что носовые, фрикативные и смычные согласные больше всего выигрывают от параллельной структуры, некоторые гласные не могут быть представлены с помощью синтезатора с параллельными формантами так же эффективно, как с каскадным.

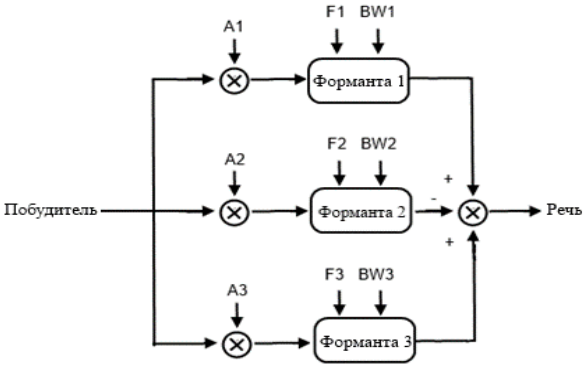

Рисунок 2. Параллельный формантный синтез речи

Конкатенативный синтез речи. Конкатенативный синтез речи, также известный как синтез речи с выбором единиц, [4, 5] основан на конкатенации речевых сегментов. Преимущество этого метода в том, что сгенерированная речь звучит естественно; пока система хорошо структурирована и для разработки доступны соответствующие речевые данные. Недостатком является то, что все используемые фрагменты речи должны быть предварительно записаны и сохранены в базе данных. Кроме того, на выбор мало голосов дикторов, и сложно внести изменения с точки зрения эмоций, таких как ударение, просодия и так далее. Коллекция базы данных обычно содержит фрагменты речи на разных уровнях от полных предложений до отдельных слогов, которые были записаны дикторами. Система последовательного синтеза речи ищет речевые единицы, которые соответствуют предоставленному входному тексту, а затем объединяет эти единицы для создания речевых выходных данных. Ядром конкатенационных систем синтеза речи являются компоненты подслов. В связи с тем, что в любом языке есть несколько фонем, их акустические свойства сильно зависят от контекста. Таким образом, используются дифоны и трифоны, поскольку они зависят от контекста. Ниже приведены основные этапы (рис. 3) последовательного синтеза речи [6]:

- *построение целевой спецификации из входного текста*: этот процесс состоит из синтезируемой фоннемной строки, а также дополнительных просодических признаков,
- *выбор единиц* для каждого сегмента фонемы,
- *постобработка* для смягчения влияния возможных процессов конкатенации.

Основная цель последовательного синтеза речи состоит в том, чтобы выполнить выбор единиц таким образом, чтобы выходная речь точно соответствовала спецификации, но при этом звучала очень естественно. Основная цель последовательного синтеза речи состоит в том, чтобы выполнить выбор единиц таким образом, чтобы выходная речь точно соответствовала спецификации, но при этом звучала очень естественно.

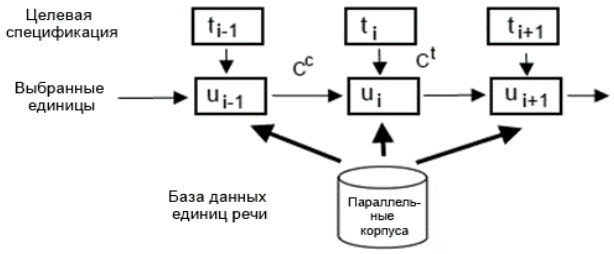

Рисунок 3. Конкатенативный синтез речи

Целевая стоимость $C^t(u_i, t_i)$, которая представляет несоответствие между спецификацией целевой единицы речи t_i и единицей-кандидатом базы данных u_i, и стоимость конкатенации $C^c(u_{i-1}, u_i)$, которая представляет несоответствие между единицей-кандидатом u_i и предшествующуй единицей u_{i-1}, представляют собой две функции затрат, которые используются при минимизации затрат для достижения этой основной цели. Целевая стоимость может быть рассчитана следующим образом [6]:

$$C^t(t_i, u_i) = \sum_{j=1}^{P} w_j^t \, C_j^t(t_i, u_i) \tag{5}$$

где $\mathbf{w}^t = [w^t_1, w^t_2, \ldots, w^t_P]$ – сумма весов внутренних целевых затрат. Стоимость конкатенации $C^c(u_{i-1}, u_i)$ может быть рассчитана следующим образом:

$$C^c(u_{i-1}, u_i) = \sum_{j=1}^{Q} w_j^c \, C_j^c(u_{i-1}, u_i) \tag{6}$$

где $\mathbf{w}^c = [w^c_1, w^c_2, \ldots, w^c_P]$ – сумма весов внутренних затрат на конкатенацию.

Статистический параметрический синтез речи. Применяя статистические модели речи, а не завися от предварительно записанных единиц, подход статистического параметрического синтеза речи преодолевает основные недостатки предыдущих систем. Основная идея состоит в том, чтобы сгенерировать акустические параметры [7], необходимые для формирования речи, а затем применить различные методы [8] для восстановления речи из сгенерированных акустических параметров. В основном для построения этих статистических моделей из набора речевых данных используются методы машинного обучения. Таким образом, системы статистического параметрического синтеза речи похожи на системы автоматического распознавания речи: в то время как системы распознавания речи используют модели машинного обучения для преобразования речи в строку, а системы синтеза речи используют те же модели для преобразования строки в речь.

В стандартной статистической параметрической системе синтеза речи параметрические представления речи получаются из речевого корпуса. Эти представления включают спектральные характеристики и характеристики возбуждения. После этого модель будет построена с использованием набора генеративных моделей. Критерий максимального правдоподобия, используемый для расчета параметров модели, может быть рассчитан следующим образом [9]:

$$\hat{\lambda} = \arg\max_{\lambda}\{p\,(O \mid W, \lambda)\}, \tag{7}$$

где λ - набор параметров модели, O - набор обучающих данных, W - набор последовательностей слов, соответствующих O. Затем мы создаем параметры речи из набора оценочных моделей, чтобы максимизировать их выходные вероятности для сгенерированного входного текста:

$$\hat{o} = \arg\max_{o}\{p\,(o \mid w, \hat{\lambda})\}, \tag{8}$$

где λ - набор расчетных моделей, o - параметры речи, w - последовательности слов.

В конце, речевой сигнал создается из параметрических речевых представлений. Можно использовать любую генеративную модель, как HMM. Статистический параметрический подход состоит из трех компонентов: модуля анализа текста, акустической модели и вокодера (рис. 4). Традиционный метод использования этих методов основан на сочетании вокодера для синтеза

сигналов и архитектуры смешанной гауссовой модели со скрытой марковской моделью (HMM-GMM) для генерации признаков.

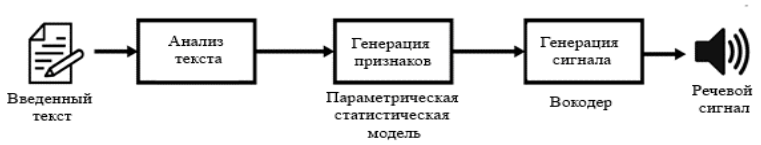

Рисунок 4. Статистический параметрический синтез речи

Цель генерации признаков состоит в том, чтобы перевести лингвистические признаки из данного текста в сопоставимое описание акустического сигнала. Стандартный подход к вероятностному отображению заключался в использовании HMM-GMM в качестве статистической параметрической модели. Подобно системе автоматического распознавания речи на основе HMM-GMM, состояния HMM соответствуют частям единиц подслов. Вероятности перехода между состояниями описывают, как речь развивается через каждую единицу подслова и от одной единицы к другой. Акустические характеристики, связанные с каждым состоянием, моделируются с помощью GMM, где GMM описывает распределение вероятностей по возможным векторам акустических признаков в этом состоянии (рис.5).

Рисунок 5. Обучение системы статистического параметрического синтеза речи

Обычный сигнал для надежной речи состоит из «непрерывных» значений амплитуды. Кроме того, на форму сигнала влияет ряд переменных, таких как усиление сигнала и амплитудные характеристики цепочки записи и передачи. Представление сигнала формы волны имеет значительную степень размерности, непредсказуемости и избыточности. В небольших окнах речевой сигнал можно рассматривать как квазистационарный. Набор спектральных свойств и свойств источника может использоваться для определения речевого содержимого сигнала в пределах этих коротких окон. Общая структура сигнала может быть представлена в значительно более низком размерном и менее изменчивом представлении, чем реальная форма волны, путем выделения признаков в скользящем окне с небольшими шагами окна. Следовательно, вокодер – это алгоритм, который может:

- преобразовать речевой сигнал в более сжатый набор функций,
- реконструировать речь по признакам,
- использовать функции для анализа и обработки речевых сигналов.

Развитие подходов к синтезу речи с течением времени значительно повысило естественность синтезированной речи. В последние годы область искусственного интеллекта приобрела новое направление исследований, называемое глубоким обучением. По сравнению с предыдущими подходами этот метод обладает более сильными возможностями моделирования и может эффективно получать скрытую информацию в данных [10, 11]. Основная цель глубокой нейронной сети (DNN) в статистическом параметрическом подходе состоит в том, чтобы смоделировать функцию отображения, которая переводит языковые характеристики в звуковые характеристики. Сложные отношения между языком и акустическими характеристиками могут быть эффективно распределены с использованием модели DNN. Тем не менее, у этой модели есть недостаток, заключающийся в игнорировании непрерывности речи. Рекуррентная нейронная сеть (RNN) предлагает эффективный способ моделирования корреляции между последующими речевыми рамками. В результате некоторые исследователи используют RNN, а не DNN, чтобы зафиксировать долгосрочную надежность речевых рамок [12-14].

Интегральный синтез речи. Типичный статистический параметрический синтез речи или синтез речи на основе глубокого обучения представляет собой сложный конвейер, состоящий из нескольких модулей, включая сеть преобразования текста в фонему, сеть сегментации звука, сеть прогнозирования длительности фонемы, сеть прогнозирования основной частоты и вокодер [15]. Создание этих модулей требует большого профессионального опыта и значительного времени. Кроме того, возможные ошибки в каждом модуле могут затруднить обучение модели. Необходимость упростить системы синтеза речи и устранить потребность в ручном вмешательстве и лингвистическом опыте стимулирует интегральные (end-to-end) подходы. Интегральная модель может напрямую синтезировать сигнал из заданного текста после обучения достаточного набора данных.

Самые современные интегральные модели TTS, основанные на глубоком обучении, продемонстрировали способность синтезировать речь, почти человеческую [16, 17]. Он в основном состоит из трех компонентов: анализа текста, акустической модели и вокодера. Первый компонент преобразует заданный текст в стандартную структуру. Промежуточные акустические характеристики, созданные вторым компонентом на выходе предыдущего этапа, затем используются для моделирования долговременной структуры речи. Затем третий компонент используется для создания образцов формы волны из акустических характеристик. Все эти компоненты обычно тренируются индивидуально, и в качестве альтернативы они могут быть отрегулированы совместно. Методы синтеза речи на основе глубокого обучения, включая WaveNet [18], Tacotron [17] и SampleRNN [19], имеют ряд недостаток. Обучение этих моделей требует много времени. Синтезированная речь обычно лишена эмоциональности и ритмичности. Кроме того, размер набора обучающих речевых данных оказывает значительное влияние на обучение таких моделей.

Результаты и их обсуждение

Разработана интеллектуальная система «Ахметтану», включающая в себя оцифрованные материалы, составленные из новой системы знаний со всеми структурными слоями на основе научного наследия Ахмета Байтурсынулы и его исследований. Собираются аудиотекстовые материалы произведений А. Байтурсынулы. Всего было собрано 50 часов данных. Всего из произведений Байтурсынулы было отобрано шесть книг, из которых были отобраны наиболее распространенные произведения и собраны в аудиотекстовые материалы. Один профессиональный диктор-мужчина озвучивал собранные текстовые данные. Диктору и оператору был дан набор инструкций, которым необходимо следовать, чтобы создать высококачественный набор данных. Записи были сохранены с использованием 16 bit/sample и дискретизированы на частоте 48 кГц. Модели E2E-TTS на основе архитектур Tacotron2 и ParalleWavegan были построены с использованием инструмента coqui-tts [20]. Рецепт LJ Speech, предоставляемый этим инструментом, был адаптирован для обучения модели с такими изменениями, как список символов. На рисунке 6 показаны графики, сгенерированные в процессе обучения.

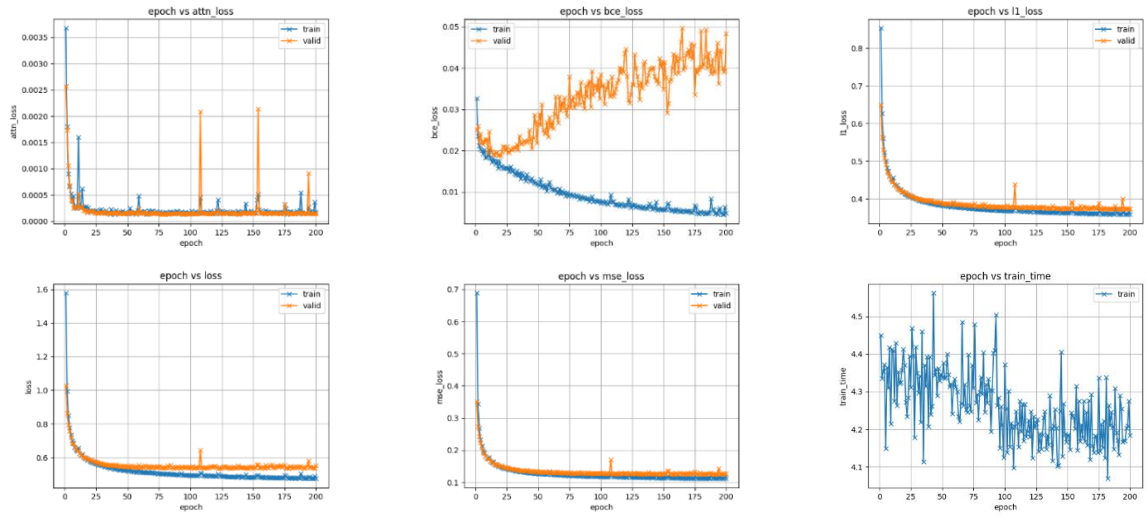

Рисунок 6. Графики, сгенерированные в процессе обучения

На рисунке 7 показано сравнение оригинальной речи и синтезированной речи.

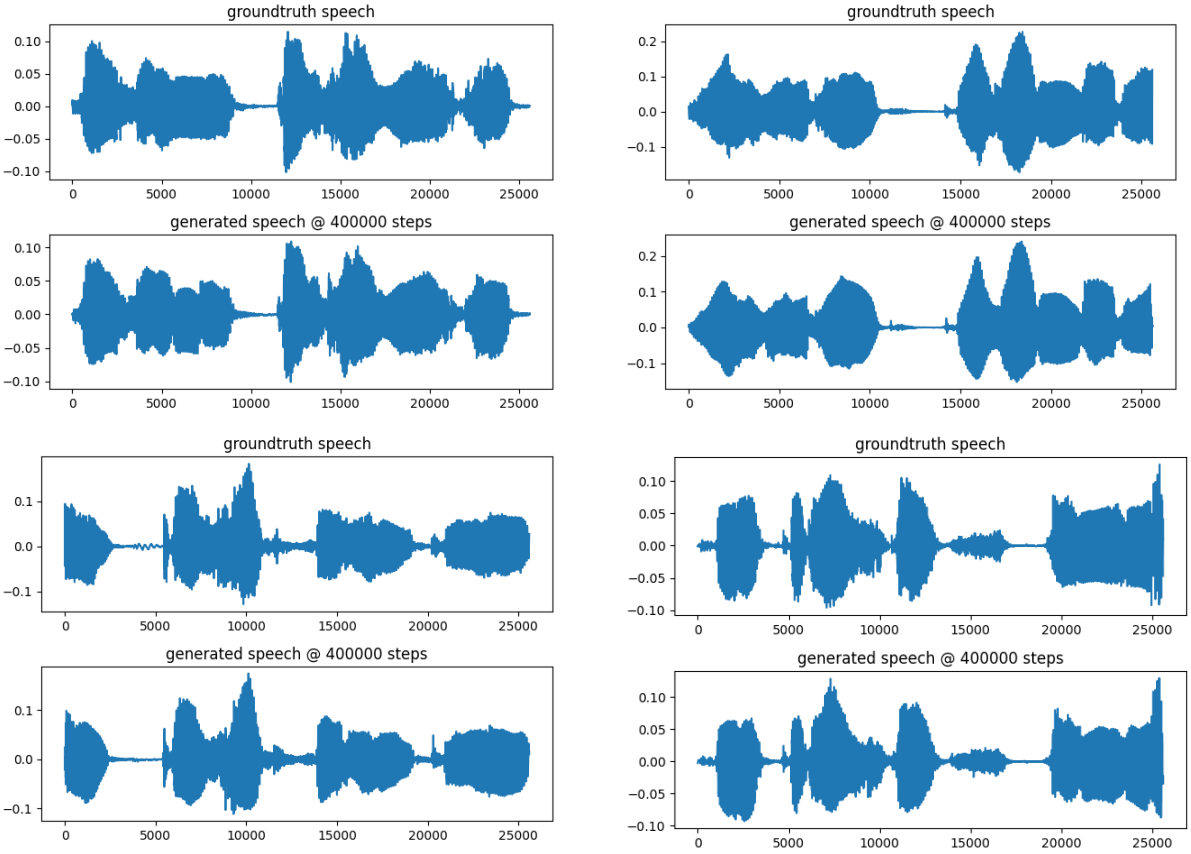

Рисунок 7. Сравнение оригинальной речи и синтезированной речи

Мы синтезируем четыре предложения для логического вывода, чтобы показать качество синтезированной речи. Мы выбираем четыре аудиозаписи случайным образом в качестве тестовых данных. Мы сравнили такие случайно выбранные аудиофайлы с эталонными аудиофайлами, чтобы получить достоверные сведения.

Заключение

В данной работе мы провели обучение моделям E2E-TTS на базе архитектур Tacotron2 и ParalleWavegan. Был проведен анализ типов синтеза речи с целью выбора наилучшего варианта. Одним из самых значительных поворотных моментов в истории синтеза речи, несомненно, стало появление нейронной сети. Существует огромное количество общедоступных наборов инструментов E2E-TTS, и coqui-tts был выбран из-за его простоты. Как показали результаты эксперимента, размер параллельного набора данных и гиперпараметров играет существенную роль в обучении модели и ее производительности. В частности, в будущем могут быть проведены эксперименты по добавлению дополнительных наборов данных.

Благодарности

Работа выполнена при поддержке грантового финансирования проектов Комитета науки Министерства образования и науки Республики Казахстан (гранты № BR11765535, № AP13068635 и № AP08053085).

References:

1 Allen, Jonathan, Sharon Hunnicutt, Rolf Carlson, and Bjorn Granstrom. "MITalk-79: The 1979 MIT text-to-speech system." The Journal of the Acoustical Society of America 65, no. S1 (1979): 57−63.

2 Klatt, Dennis H. "Software for a cascade/parallel formant synthesizer." the Journal of the Acoustical Society of America 67, no. 3 (1980): 971-995.

3 Allen, Jonathan, M. Sharon Hunnicutt, Dennis H. Klatt, Robert C. Armstrong, and David B. Pisoni. From text to speech: The MITalk system. Cambridge University Press, 1987: 397

4 Olive, Joseph. "Rule synthesis of speech from dyadic units." In ICASSP'77. IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. 568-570. IEEE, 1977.

5 Hunt, Andrew J., and Alan W. Black. "Unit selection in a concatenative speech synthesis system using a large speech database." In 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, vol. 1, pp. 373-376. IEEE, 1996.

6 Schwarz, Diemo. "Concatenative sound synthesis: The early years." Journal of New Music Research 35, no. 1 (2006): 3-22.

7 Kawahara, Hideki, Ikuyo Masuda-Katsuse, and Alain De Cheveigne. "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds." Speech communication 27, no. 3-4 (1999): 187-207.

8 Kawahara, Hideki. "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds." Acoustical science and technology 27, no. 6 (2006): 349-353.

9 Zen, Heiga, Keiichi Tokuda, and Alan W. Black. "Statistical parametric speech synthesis." speech communication 51, no. 11 (2009): 1039-1064.

10 Fernandez, Raul, Asaf Rendel, Bhuvana Ramabhadran, and Ron Hoory. "F0 contour prediction with a deep belief network-Gaussian process hybrid model." In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6885-6889. IEEE, 2013.

11 Lu, Heng, Simon King, and Oliver Watts. "Combining a vector space representation of linguistic context with a deep neural network for text-to-speech synthesis." In Eighth ISCA workshop on speech synthesis. 2013.

12 Liu, Zhijun, Kuan Chen, and Kai Yu. "Neural Homomorphic Vocoder." In INTERSPEECH, pp. 240-244. 2020.

13 Zen, Heiga, and Haşim Sak. "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis." In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4470-4474. IEEE, 2015.

14 Zen, H., H. Sak, A. Graves, and A. Senior. "Statistical parametric speech synthesis based on recurrent neural networks." In Poster presentation given at UKSpeech Conference. 2014.

15 Gibiansky, Andrew, Sercan Arik, Gregory Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. "Deep voice 2: Multi-speaker neural text-to-speech." Advances in neural information processing systems 30 (2017).

16 Shen, Jonathan, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen et al. "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions." In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 4779-4783. IEEE, 2018.

17 Wang, Yuxuan, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang et al. "Tacotron: Towards End-to-End Speech Synthesis." Proc. Interspeech 2017 (2017): 4006-4010.

18 Oord, Aaron van den, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. "Wavenet: A generative model for raw audio." In 9th ISCA Speech Synthesis Workshop, pp. 125-145. 2016.

19 Mehri, Soroush, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. "SampleRNN: An unconditional end-to-end neural audio generation model." In Proceedings of ICLR, pp. 4006-4010. 2017.

20 Casanova, Edresson, Julian Weber, Christopher D. Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A. Ponti. "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone." In International Conference on Machine Learning, pp. 2709-2720. PMLR, 2022.