ORES
Научно-практический электронный журнал
Оригинальные исследования (ОРИС), *ISSN 2222-0402*
www.ores.su
original-research.ru

УДК 004.89

АСПЕКТЫ ЧЕЛОВЕКО-МАШИННОГО ВЗАИМОДЕЙСТВИЯ В TTS-СИСТЕМАХ РЕАЛЬНОГО ВРЕМЕНИ

Белоножко Павел Евгеньевич
Калужский филиал федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»
Студент магистратуры
belonozhkope@student.bmstu.ru

Федоров Виктор Олегович
Калужский филиал федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»
Кандидат технических наук, доцент
fedorov_vo@bmstu.ru

Аннотация

Рассмотрено человеко-машинное взаимодействие в системах преобразования текста в речь (TTS) с использованием анализа мел-спектрограмм. Представлены основные концепции в TTS-системах, построенных на генеративных архитектурах WaveNet и Tacotron-2. Показан конвейерный процесс преобразования текста в речь, в котором ключевые функции выполняют WaveNet и Tacotron. Анализ мел-спектрограммы рассмотрен как важный метод для понимания звукового содержания, внесший изменения в представление спектральной информации. Подчеркнуто влияние распределения энергии по частотам и динамики изменений во времени на выделение звуковых элементов и контекста звука. Сделан вывод о уникальности платформы, позволяющей обучение моделей отдельно на различных наборах данных для повышения устойчивости к шуму.

Ключевые слова: Мел-спектрограмма, человеко-машинное взаимодействие, преобразование текста в речь, Tacotron, WaveNet

ASPECTS OF HUMAN-MACHINE INTERACTION IN REAL-TIME TEXT-TO-SPEECH (TTS) SYSTEMS

Pavel E. Belonozhko
Federal State Budgetary Educational Institution of Higher Education «Bauman Moscow State Technical University» (Kaluga Branch)
Master's degree student
belonozhkope@student.bmstu.ru

Viktor O. Fedorov
Federal State Budgetary Educational Institution of Higher Education «Bauman Moscow State Technical University» (Kaluga Branch)
Candidate of Technical Sciences, associate professor
fedorov_vo@bmstu.ru

ABSTRACT

The human-machine interaction in Text-to-Speech (TTS) systems utilizing mel-spectrogram analysis has been examined. The fundamental concepts in TTS systems based on generative architectures, namely WaveNet and Tacotron-2, have been presented. The pipeline for the text-to-speech conversion process has been illustrated, with WaveNet and Tacotron playing key roles. Mel-spectrogram analysis has been highlighted as a crucial method for understanding the acoustic content, bringing changes to the representation of spectral information. The impact of energy distribution across frequencies and the dynamics of temporal changes on the extraction of sound elements and sound context has been emphasized. The conclusion is drawn regarding the uniqueness of the platform, allowing the training of models separately on different datasets to enhance resilience to noise.

Keywords: Mel-spectrogram, human-machine interaction, text-to-speech conversion, Tacotron, WaveNet.

Введение. Системы синтеза речи играют важную роль в современной цифровой эпохе, где взаимодействие человека с компьютерами и устройствами становится все более естественным и интегрированным в повседневную жизнь. Одной из ключевых технологий, существенно влияющей на этот процесс, являются системы текст в речь (TTS), обеспечивающие возможность преобразования письменного текста в аудиоинформацию с естественным звучанием. Современные TTS-системы, работающие в реальном времени, представляют собой сложные технологии, включающие в себя передовые алгоритмы обработки языка, глубокого обучения и синтеза речи. Эти системы не только обеспечивают пользователей высококачественной аудиоинформацией, но и стремятся создать максимально естественный и понятный звук, сближаясь с интонацией и мелодичностью человеческой речи [1].

Особое внимание уделяется аспектам человеко-машинного взаимодействия, предоставляемого через анализ мел-спектрограмм в режиме реального времени. Системы преобразования текста в речь, основанные на анализе мел-спектрограмм, позволяют более глубоко воссоздавать тонкости человеческой речи, улавливая интонации, ритм и нюансы выражения. Этот подход позволяет создавать более реалистичные и естественные аудиоотклики, что важно для улучшения взаимодействия пользователей с TTS-системами.

Изучение человеко-машинного взаимодействия в TTS-системах через анализ мел-спектрограмм не только акцентирует внимание на технологических инновациях, но также выдвигает вопросы, связанные с адаптацией, этикой и безопасностью, открывая перспективы для более глубокого понимания и совершенствования этой важной области [2].

Цель исследования. Цель исследования заключается в анализе человеко-машинного взаимодействия в системах преобразования текста в речь (TTS) через анализ мел-

спектрограмм, с фокусом на генеративных архитектурах, таких как WaveNet и Tacotron-2. Исследование направлено на понимание роли этих моделей в конвейерном процессе TTS и их влияния на восприятие звукового содержания. Дополнительные задачи включают анализ взаимодействия пользователей с системой на платформе SV2TTS и оценку успешности синтеза речи, как визуально, так и количественно с использованием метрик качества звука. Полученные результаты могут быть применены для дальнейшего улучшения технологий синтеза голоса и оптимизации TTS-систем.

Основные понятия в системах преобразования текста в речь. Основой любой современной системы преобразования текста в речь является генеративная архитектура, которая включает в себя несколько моделей глубокого обучения. В данной работе рассматривается система на основе моделей WaveNet и Tacotron-2. Важно подчеркнуть, что данные модели не представляют собой самостоятельные системы преобразования текста в речь, а являются лишь частью более обширного набора моделей и эвристик, работающих совместно для достижения эффективного синтеза речи.

Механизм преобразования текста в речь включает в себя программное обеспечение, разделенное на конвейер, где каждый этап представляет собой модель или набор моделей. Данный конвейер включает в себя следующие процессы: нормализацию, маркировку частей речи, преобразование фонем, высокоуровневый синтез звука и синтез формы сигнала (рисунок 1).

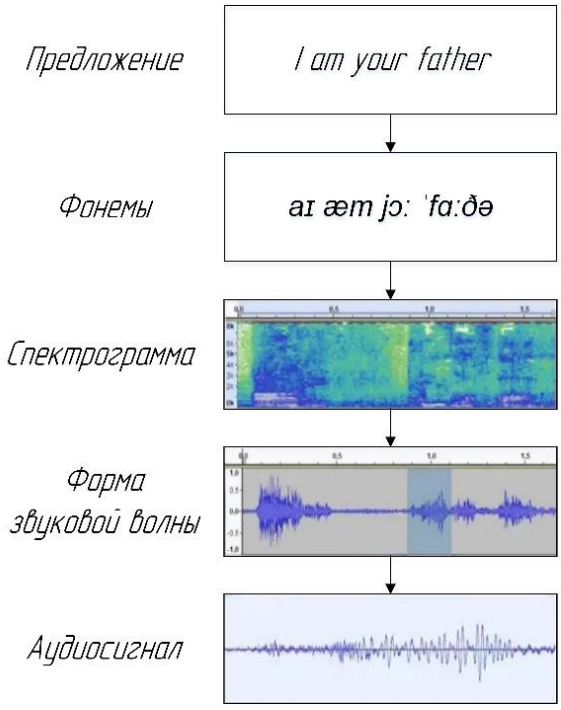

Рисунок 1. Конвейер преобразования текста в речь

WaveNet и Tacotron выполняют ключевые функции на более высоком уровне конвейера, отвечая за важные этапы в процессе синтеза речи. WaveNet, функционируя как нейронный вокодер, занимается созданием формы сигнала, а Tacotron представляет собой последовательную модель для синтеза спектрограмм, которые в свою очередь отвечают за высокоуровневый синтез звука [3].

Мел-спектрограммы и способы их обработки. Аудиоданные для моделей глубокого обучения обычно состоят из цифровых аудиофайлов. Эти файлы хранятся в различных форматах в зависимости от того, как сжимается звук. Далее эти аудиоданные проходят обработку путем дискретизации звуковой волны через равные промежутки времени и измерения интенсивности или амплитуды волны в каждой выборке. Метаданные для этого

аудио сообщают частоту дискретизации, которая представляет собой количество выборок в секунду.

В памяти звук представляется в виде временного ряда чисел, представляющих амплитуду на каждом временном шаге. Поскольку измерения проводятся через фиксированные промежутки времени, данные содержат только числа амплитуд, а не значения времени. Зная частоту дискретизации, можно выяснить, в какой момент времени было выполнено каждое измерение числа амплитуд [4].

Битовая глубина определяет, сколько возможных значений могут принимать измерения амплитуды для каждого образца. Например, битовая глубина 16 означает, что число амплитуд может быть от 0 до 65535. Битовая глубина влияет на разрешение измерения звука — чем выше битовая глубина, тем лучше качество звука.

Модели глубокого обучения редко принимают необработанный звук такого вида напрямую в качестве входных данных. Обычной практикой является преобразование аудио в спектрограмму. Спектрограмма представляет собой краткий «моментальный снимок» звуковой волны, и, поскольку это изображение, оно хорошо подходит для ввода в архитектуры на основе сверточных нейронных сетей, разработанных для обработки изображений.

Спектрограммы генерируются из звуковых сигналов с использованием преобразований Фурье. Преобразование Фурье разлагает сигнал на составляющие его частоты и отображает амплитуду каждой частоты, присутствующей в сигнале. Она делит продолжительность звукового сигнала на более мелкие временные сегменты, а затем применяет преобразование Фурье к каждому сегменту, чтобы определить частоты, содержащиеся в этом сегменте. Затем объединяет преобразования Фурье для всех этих сегментов в один график. Он отображает частоту (ось Y) в зависимости от времени (ось X) и использует разные цвета для обозначения амплитуды каждой частоты. Чем ярче цвет, тем выше мощность сигнала (рисунок 2).

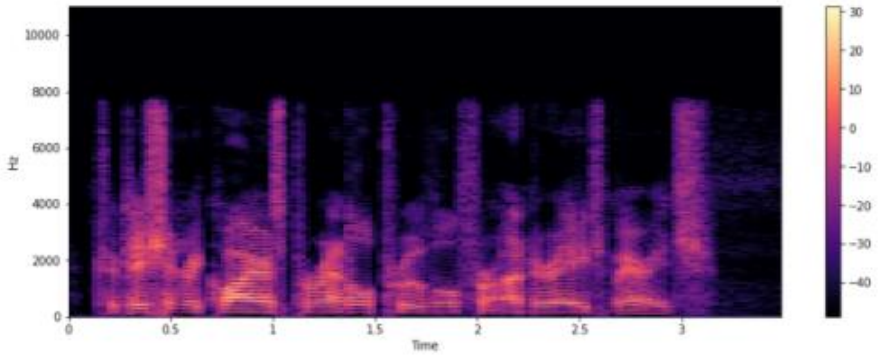

Рисунок 2. Простая спектрограмма, отображающая зависимость частоты от времени

Недостатком такой спектрограммы является отображение малого количества полезной информации. Это происходит из-за некорректного восприятия звука. Большая часть того, что можно услышать, сосредоточена в узком диапазоне частот и амплитуд. Чтобы учесть это, была разработана шкала Мела путем проведения экспериментов с большим количеством слушателей. Это шкала высоты тона, где каждая единица оценивается слушателями как равная по высоте тона от следующей. Для реалистичной работы со звуком важно использовать логарифмическую шкалу через шкалу Мела и шкалу децибел при работе с частотами и амплитудами в исходных данных. Именно для этого предназначена спектрограмма Мела [5].

Мел-спектрограмма (рисунок 3) вносит два важных изменения по сравнению с обычной спектрограммой, которая отображает зависимость частоты от времени. Она использует шкалу Мела вместо частоты по оси Y и шкалу децибел вместо амплитуды для обозначения цветов. Таким образом, мел-спектрограмма представляет собой визуализацию энергии в различных частотных диапазонах в зависимости от времени, что позволяет выделить важные аспекты аудиосигнала.

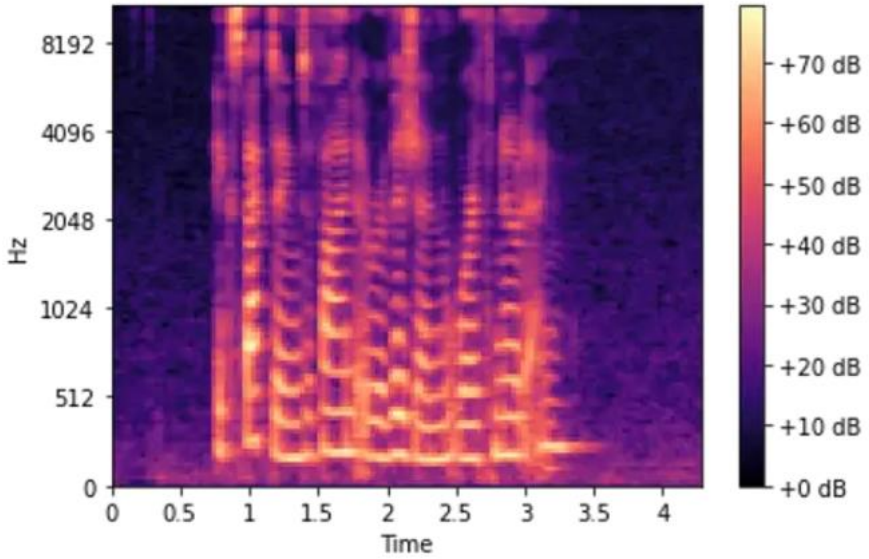

Рисунок 3. Мел-спектрограмма, используемая в генеративных моделях преобразования текста в речь

Анализ мел-спектрограмм играет ключевую роль в понимании звукового содержания аудиофайлов и может быть весьма полезным в различных областях, включая распознавание речи, обработку звука и музыкальный анализ..

Важной характеристикой мел-спектрограммы является распределение энергии по частотам. Высокая энергия в конкретных частотных диапазонах может свидетельствовать о наличии определенных звуковых элементов или особенностей. Например, в речи высокая энергия в определенных частотах может указывать на наличие сильных гармоник, что характерно для звуковых элементов, таких как голосовые гармоники.

Второй аспект анализа мел-спектрограммы - изменения во времени. Динамика изменений в спектрограмме может указывать на темп, интонацию или другие временные характеристики аудиосигнала. Например, скачкообразные изменения могут указывать на быстрое переключение между звуковыми элементами, что может быть важным при анализе речи или музыки.

Структура и форма мел-спектрограммы могут предоставить информацию о контексте звука. Различные шаблоны и формы могут указывать на различные звуковые события, от фонового шума до выделенных аудио-сигналов. Путем внимательного анализа этих характеристик и их соотнесения с конкретными аудио-событиями можно делать выводы о содержании аудиофайла.

Взаимодействие с системой. Взаимодействие пользователей с прототипом системы осуществляется при помощи платформы SV2TTS. В качестве вокодера используется WaveNet, эффективно применяя структуру Tacotron 2 для вывода звуковой волны из спектрограммы. Однако, уникальность платформы SV2TTS заключается в том, что все модели могут быть обучены отдельно и на разных наборах данных. Кодировщик обучается

на модели, устойчивой к шуму, с использованием большого количества различных динамиков без жестких требований к уровню шума аудио.

Интерфейс системы представлен на рисунке 4, где пользователь начинает с выбора звукового файла высказывания из различных наборов данных. Система поддерживает обработку популярных речевых данных и может быть настроена для добавления новых. Пользователь также имеет возможность записывать высказывания для клонирования своего голоса.

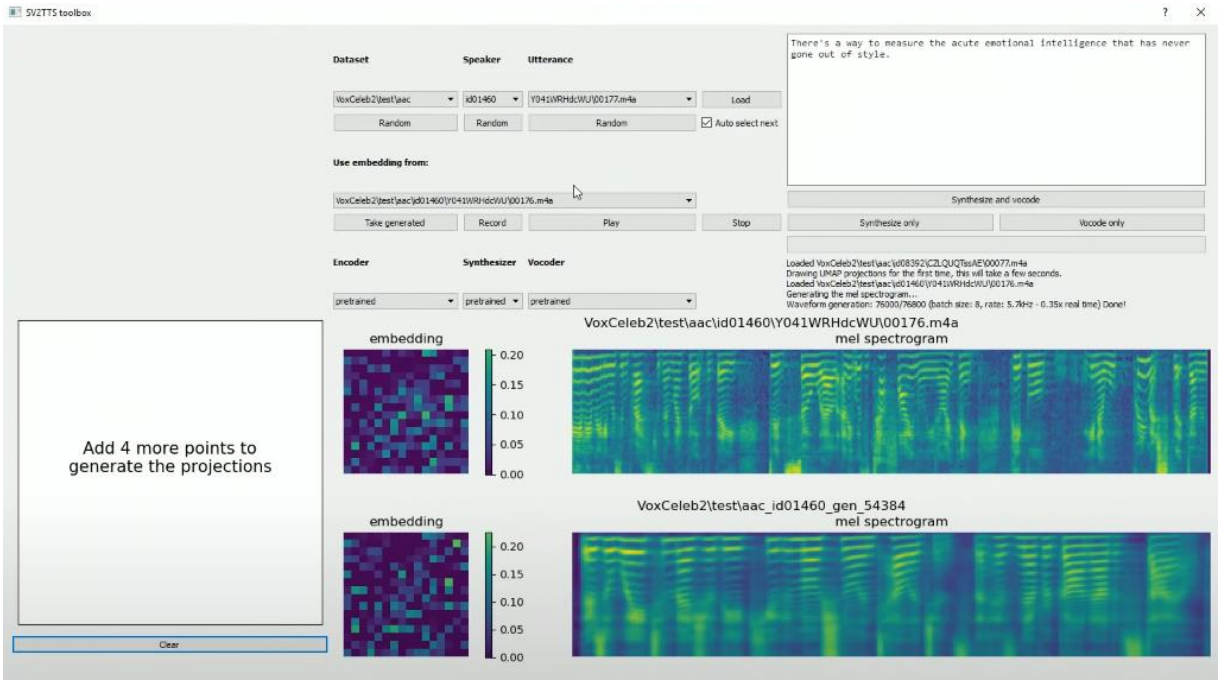

Рисунок 4. Интерфейс прототипа системы

После загрузки высказывания система вычисляет его частотные характеристики, обновляет прогнозы UMAP и отображает мел-спектрограмму. Вложение представлено вектором с графиком тепловой карты, что обеспечивает визуальные подсказки о различиях между вложениями.

Пользователь может вводить произвольный текст для синтеза, управлять просодией с помощью разрывов строк и генерировать спектрограмму. Синтезированная спектрограмма и вложение отображаются на интерфейсе, а затем сегмент спектрограммы можно сгенерировать с использованием вокодера. Этот процесс дает пользователю возможность создавать уникальные вложения и использовать их как эталоны для дальнейшего поколения голосовых данных [6].

Оценка успешности синтеза речи может быть выполнена на нескольких уровнях. Первым этапом является визуальная оценка мел-спектрограммы, где пользователь может обратить внимание на четкость и структуру звуковых паттернов. Удачный синтез обычно характеризуется четким отображением основных характеристик речи, таких как интонация, ритм и длительность фраз (рисунок 5).

Рисунок 5. Корректно сгенерированная речь

Однако важно отметить, что визуальная оценка мел-спектрограммы может быть ограничена, и для более объективной оценки необходимо использовать дополнительные метрики. Например, можно внедрить метрики качества звука, такие как частота дискретизации, отношение сигнал/шум и другие акустические характеристики. Эти метрики могут дать количественную оценку качества синтеза, а пользователь может использовать их для определения того, что нужно улучшить.

Выводы. Данная работа представляет обзор ключевых аспектов систем преобразования текста в речь, подчеркивая генеративную архитектуру и модели глубокого обучения, такие как WaveNet и Tacotron-2, в контексте обширного конвейера обработки данных. Важно отметить, что в статье также рассмотрен вопрос человеко-машинного взаимодействия, предоставляя пользователю возможность вводить текст для синтеза и управлять просодией. Анализ мел-спектрограмм играет ключевую роль в оценке успешности синтеза, а визуализация результатов через спектрограммы обеспечивает пользователю понятный и интуитивный способ взаимодействия с системой. Таким образом, данная работа не только рассматривает технические аспекты, но и акцентирует важность удобства и эффективности взаимодействия между человеком и машиной в области синтеза речи.

Список литературы:

1. Белоножко П.Е. Модификации архитектуры WaveNet для реализации вокодера в генеративной модели преобразования текста в речь // Научное обозрение. Технические науки. 2022. № 6. с. 37-42.
2. A. K. Yetkin, H. Köse. Speech-Based Emotion Analysis Using Log-Mel Spectrograms and MFCC Features // 2023 31st Signal Processing and Communications Applications Conference (SIU). 2023. pp. 1-4.
3. G. Ulutas, G. Tahaoglu, B. Ustubioglu. Forge Audio Detection Using Keypoint Features on Mel Spectrograms // 2022 45th International Conference on Telecommunications and Signal Processing (TSP). 2022. pp. 413-416.
4. J. Liu, Y. Guo, J. Chen. Speech Synthesis for Speaker Timbre Translation Across Languages. // 2022 4th International Conference on Control and Robotics (ICCR). 2022. pp. 320-324.
5. R. Jain, M. Y. Yiwere, D. Bigioi. A Text-to-Speech Pipeline, Evaluation Methodology, and Initial Fine-Tuning Results for Child Speech Synthesis. // IEEE Access. 2022. vol. 10. pp. 47628-47642.

6. W. Zhang, Y. Jia. A Study on Speech Emotion Recognition Model Based on Mel-Spectrogram and CapsNet. // 2021 3rd International Academic Exchange Conference on Science and Technology Innovation (IAECST). 2021. pp. 231-235.

References:

1. P.E. Belonozhko. Modifications of the WaveNet Architecture for the Implementation of a Vocoder in a Generatic Model of Text-to-Speech Conversion // Scientific Review. Technical Sciences. 2022. № 6. pp. 37-42.
2. A. K. Yetkin, H. Köse. Speech-Based Emotion Analysis Using Log-Mel Spectrograms and MFCC Features // 2023 31st Signal Processing and Communications Applications Conference (SIU). 2023. pp. 1-4.
3. G. Ulutas, G. Tahaoglu, B. Ustubioglu. Forge Audio Detection Using Keypoint Features on Mel Spectrograms // 2022 45th International Conference on Telecommunications and Signal Processing (TSP). 2022. pp. 413-416.
4. J. Liu, Y. Guo, J. Chen. Speech Synthesis for Speaker Timbre Translation Across Languages. // 2022 4th International Conference on Control and Robotics (ICCR). 2022. pp. 320-324.
5. R. Jain, M. Y. Yiwere, D. Bigioi. A Text-to-Speech Pipeline, Evaluation Methodology, and Initial Fine-Tuning Results for Child Speech Synthesis. // IEEE Access. 2022. vol. 10. pp. 47628-47642.
6. W. Zhang, Y. Jia. A Study on Speech Emotion Recognition Model Based on Mel-Spectrogram and CapsNet. // 2021 3rd International Academic Exchange Conference on Science and Technology Innovation (IAECST). 2021. pp. 231-235.