Обзор существующих алгоритмов преобразования текста в речь

Н.С. Киреев, Е.А. Ильюшин

***Аннотация*—Ученые достаточно давно работают над алгоритмами, позволяющими транслировать текст, написанный на естественном языке, в речь. Но качество работы этих алгоритмов оставляло желать лучшего до момента, когда применение методов глубокого обучения не стало возможным. С появлением необходимых вычислительных ресурсов и накопления достаточного количества данных для обучения, эти методы стали широко применять в машинном обучении в целом и, конечно, в синтезе речи в частности. Существенное улучшение качества работы алгоритмов трансляции текста в речь привело к их повсеместному применению, а именно в мобильных устройствах, умных колонках, голосовых помощниках и т.д. Но стоит отметить, что алгоритмы данного класса, разработанные на данный момент, не всегда корректно справляются с поставленной задачей. К примеру, не всегда могут правильно поставить ударение или озвучить нужные участки текста с необходимой интонацией. Таким образом исследование методов и средств, позволяющих синтезировать речь, приобрело еще большую актуальность.**

Существует множество различных способов синтеза речи по тексту, такие как параметрический синтез, компиляционный синтез, предметно-ориентированный синтез и полный синтез речи по правилам. Целью данной работы является обзор существующих алгоритмов трансляции текста в речь и проведение их сравнительного анализа. В качестве основных были рассмотрены алгоритмы: WaveNet, DeepVoice, Tacatron, DeepVoice 2, DeepVoice 3 и Tacatron 2. В ходе их сравнения было определено, что лучшими на текущий момент являются DeepVoice 3 и Tacatron 2, так как оценки качества их работы наиболее близки к профессионально записанной речи.

***Ключевые слова*—звук, синтез речи, обработка текста**

I. Введение

Технологии преобразования текста в речь существует достаточно давно, но все еще является актуальной сферой для научных исследований. Есть множество задач, которые решаются для создания схожей с человеческой синтезированной речи. К примеру, определение интонации, многоголосость систем, объединение пар фонем без слышимых недостатков, определение просодий и т.д. Но даже у современных систем существуют недостатки.

Во втором разделе статьи дадим определение понятию звук, а также, разберем его основные характеристики.

В третьем разделе дадим определение синтезу речи. Далее подробно опишем его основные этапы и рассмотрим способы их реализации.

В четвертом разделе представлены результаты исследований современных архитектур преобразования текста в речь, таких как WaveNet, DeepVoice, Tacotron, DeepVoice 2, DeepVoice 3 и Tacotron 2, а также их достоинства и недостатки.

В заключении мы обсудим проблемы современных подходов, а также дадим рекомендации с каких архитектур стоит начать погружение в такую область знаний, как синтеза речи.

II. Звук

Звук - физическое явление, представляющее собой распространение в виде упругих волн механических колебаний в твёрдой, жидкой или газообразной среде. В физиологии и психологии человека звук — это восприятие таких волн мозгом [1].

Как и любая волна, звук характеризуется амплитудой и частотой. Амплитуда характеризует громкость звука. Частота определяет тон, высоту. Обычный человек способен слышать звуковые колебания в диапазоне частот от 16—20 Гц до 15—20 кГц. Громкость звука сложным образом зависит от эффективного звукового давления, частоты и формы колебаний, а высота звука — не только от частоты, но и от величины звукового давления [1].

Среди слышимых звуков следует особо выделить фонетические, речевые звуки и фонемы (из которых состоит устная речь).

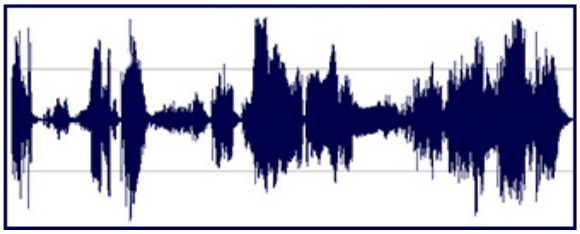

Рисунок 1 – Представление звука

III. Синтез речи

Синтез речи – искусственное производство человеческий речи. Система преобразования текста в речь преобразует текст на естественном языке в речь, остальные системы преобразуют символические

Статья получена 9 апреля 2020.

Н.С Киреев, МГУ им. М.В. Ломоносова, факультет вычислительной математики и кибернетики, Москва, Россия (e-mail: nikita.s.kireev@gmail.com).

Е.А Ильюшин, МГУ им. М.В. Ломоносова, факультет вычислительной математики и кибернетики, Москва, Россия (e-mail: eugene.ilyushin@gmail.com).

лингвистические представления, например фонетическая транскрипция, в речь [2].

Рисунок 2 – 1 секунда сгенерированной речи

Синтезированная речь может быть создана путем объединения фрагментов записанной речи, которые хранятся в базе данных. В другом методе синтезатор включает модель речевого тракта и другие характеристики человеческого голоса, для создания полностью синтезированной речи.

О качестве синтезированной речи судят по ее сходству с человеческим голосом и по ее способности быть понятной.

Система преобразования текста в речь состоит из двух частей: front-end и back-end.

У front-end части есть две основные задачи:

- преобразование необработанного текста, содержащего символы (числа, сокращения) в эквивалент написанных слов, этот процесс также называют нормализацией текста, предварительной обработкой или токенизацией;
- назначение фонетической транскрипции каждому слову, а также деление и пометка текста на просодические единицы такие как, фразы и предложения - этот процесс называют преобразованием текста в фонему или преобразованием графемы в фонему

Фонетическая транскрипция и информация о просодии вместе составляют символическое лингвистическое представление, которое выводится интерфейсом.

Back-end часто называется синтезатором, эта часть системы преобразует символическое лингвистическое представление в звук. В некоторых системах эта часть включает вычисление целевой просодии (контур основного тона, длительность фонем), которая затем накладывается на основную речь.

Двумя основными методами синтеза речи являются конкенативный и формантный синтезы [3].

1. **Конкенативный синтез.**
   Конкенативный синтез основан на объединении сегментов записанной речи. Как правило конкенативный синтез синтезирует наиболее качественную синтезированную речь. Существует 3 основных подтипа конкенативного синтеза: синтез выбора единиц, дифтонгный синтез и домен-специфический синтез. Построение модели синтеза состоит из трех этапов: запись всех выбранных единиц речи во всех возможных контекстах, маркировка и сегментация единиц, выбор наиболее подходящих единиц. Этот подход является самым простым. Из минусов: необходимо иметь большее хранилище и невозможность применять различные изменения к голосу.

2. **Форматный синтез [3].**
   При этом методе синтезированная речь создается с использованием аддитивного метода и акустической модели. Этот метод также называют синтезом на основе правил.

3. **HMM синтез [4, 5].**
   Это синтез основанный на скрытых марковских моделях, так же называемый статистическим параметрическим синтезом. В этой системе частотный спектр (речевой тракт), основная частота (источник голоса) и длительность (просодия) речи моделируются одновременно. Речевые сигналы генерируются самим HMM на основе критерия максимального правдоподобия. Синтез состоит из двух основных частей (рис. 3):
   - на этапе обучения мы извлекаем параметры возбуждения, такие как основная частота и динамические характеристики (мел-частотные кепстральные коэффициенты (MFCC)) из базы данных, а затем оцениваем их при помощи одной из статистических моделей. Скрытая Марковская модель является наиболее используемой, эта модель является контекстно зависимой благодаря чему при обучении учитывается также лингвистический и просодический контексты;
   - на этапе синтеза текст преобразуется в контекстно зависимую последовательность меток, а затем строится скрытая Марковская модель высказывания по этой последовательности, после чего спектральные параметры и параметры возбуждения генерируются из полученного высказывания. В конце речевые сигналы синтезируются из этих параметров с использованием генерации возбуждения и фильтра синтеза речи.

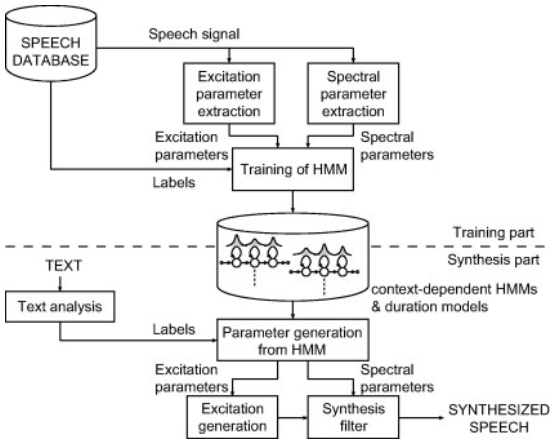

Рисунок 3 – модель HMM синтеза

Плюсами данного подхода является: небольшая занимаемая площадь, возможность изменять голос, синтез эмоций и независимость от языка. Самым большим недостатком является качество синтезированной речи.

4. **Глубокое обучение [6].**
   Синтезаторы речи, основанные на глубоком обучении, используют Deep Neural Network

(глубокие нейронные сети), которые обучаются на записанных речевых данных. Примерами являются WaveNet от Google DeepMind, Tacatron от Google и DeepVoice (в основе которого лежит технология WaveNet) от Baidu.

Основные этапы синтеза речи [7]:

1. нормализация текста;
2. преобразование текста в фонетическое представление;
3. просодическое и эмоциональное содержание (изменять основной тон в зависимости от того какого типа предложение вопросительное, восклицательное или положительное).

IV. Современные архитектуры преобразования текста в речь

1. WaveNet.

WaveNet – генеративная модель для синтеза необработанного звука. Это полностью сверточная нейронная сеть, в которой каждый новый образец зависит от предыдущего. Основным отличием этой архтектуры (рис. 4) является причинно-следственные и расширенные (дилатационные) свертки [8].

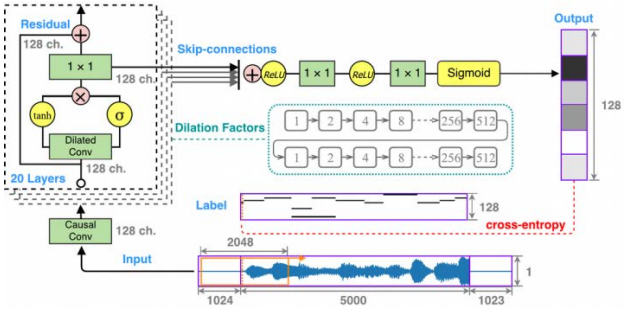

Рисунок 4 – архитектура WaveNet

На вход подаются ранее сгенерированные выборки. После причинной свертки существуют слои, в которые добавляются расширенные свертки, а затем из текста извлекаются лингвистические особенности.

Выходные данные этих слоев суммируются и обрабатываются далее посредством серии 1x1 сверток и активации, заканчивающегося softmax уровнем с 256 выводами.

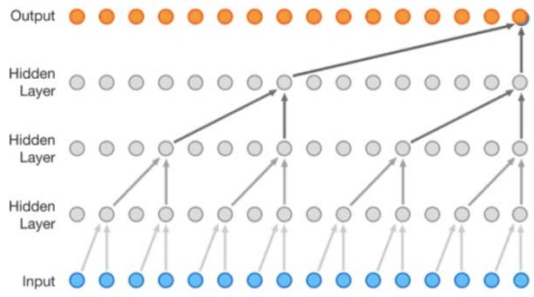

Рисунок 5 – иллюстрация концепции WaveNet

Из рисунка 5 видно, что входные данные используют ранее сгенерированные выборки для создания новых выходных данных через ряд скрытых слоев.

Особенности WaveNet:

- генерирует необработанные звуковые сигналы (более 16000 выборок в секунду);
- softmax слой моделирует условные распределения по отдельным аудио семплам;
- 4,21 оценка MOS для американского английского и 4,08 для китайского.

Минусы WaveNet:

- комплексная система, которая требует подготовки размеченных текстов;
- требует дополнительные лингвистические функции (например информацию об ударении или основной частоте);
- вычислительно дорогой синтез.

Так же был разработан механизм кэширования для удаления избыточных сверток, чтобы сократить расходы на вычислительные мощности. После этого скорость генерации WaveNet резко возросла.

2. DeepVoice.

Система DeepVoice состоит из нескольких независимо обученных моделей, объединённых в вычислительный конвейер. Предварительно обученные независимо друг от друга модели “Graphem-to-Phoneme” и “Segmentation” используются для создания функций, смешанных с обучающими наборами данных для обучения моделей “Audio-Synthesis”, “Duration Prediction” и “Fundamental Frequency”. Модель из графемы в фонему также используется при синтезировании речи наряду с последними [9].

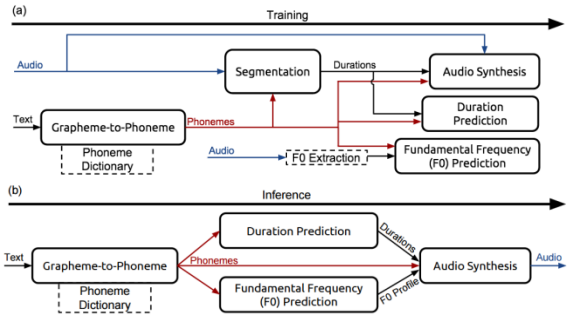

Рисунок 6 – архитектура DeepVoice

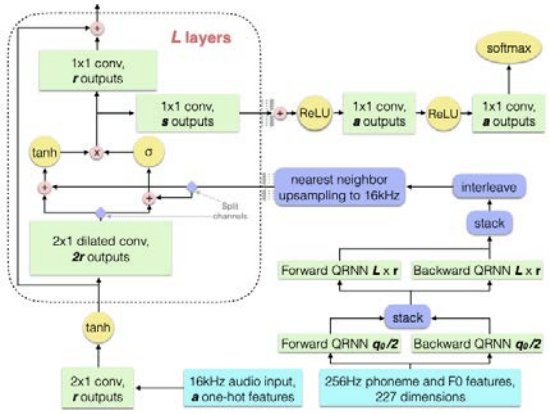

Рисунок 7 – часть аудио синтеза DeepVoice

Часть аудио синтеза (рис. 7) имеет модифицированную архитектуру WaveNet.

DeepVoice закладывает основу для синтеза речи в реальном времени, но оценка MOS достаточно низка – 2,67 для американского английского.

3. Tacotron.

Tacotron – модель типа end-to-end состоящая из кодера и декодера с механизмом внимания. Тренировка этой архитектуры достаточно проста, так как требует только пары текст-аудио [10].

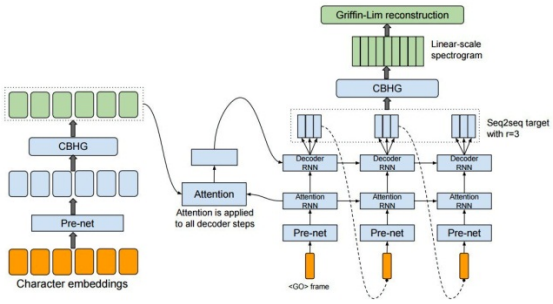

Рисунок 8 – архитектура Tacotron

Генерация речи начинается с подачи необработанного текста на вход кодера. Первый уровень кодировщика – встраивание символов. Вложения для ввода текста передаются в двухслойную предварительную сеть (рис. 8 pre-net). Следующим этапом является модуль CBHG (рис. 9).

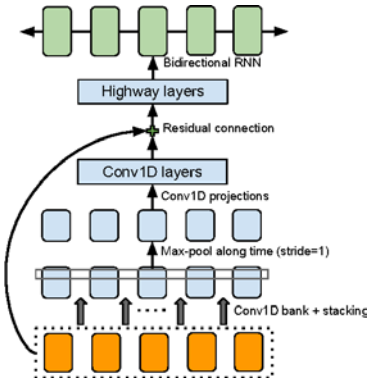

Рисунок 9 – модуль CBHG

CBHG – банк одномерной свертки, сеть магистралей и двунаправленной GRU (управляемый рекуррентный блок). Изначально он был разработан для задачи перевода. Первый этап в модуле выполняет набор сверток разных размеров на входе, результаты складываются в единый тензор. После этого применяется максимальное объединение к полученному тензору. Объединённые значения проходят через несколько слоев одномерных сверток. Результаты, объединённые с входными данными, отправляются на уровень извлечения максимально значимых функций, а затем эти функции передаются в двунаправленный GRU.

Основными задачами, решаемыми декодером, являются, прогнозирование мел-частотных кепстральных коэффициентов и прогнозирование линейной спектрограммы.

Для получения мел-частотных кепстральных коэффициентов, входные данные декодера преобразуются через предварительную сеть – слой внимания GRU и двухслойный декодер RNN (рекуррентная нейронная сеть), а также с помощью GRU. Тут также принимает участие модуль CBHG, который преобразует мел-частотные кепстральные коэффициенты в линейную спектрограмму.

Плюсы архитектуры Tacotron:

- end-to-end модель;
- оценка MOS – 3,82 для американского английского.

4. DeepVoice 2.

Улучшением по сравнению с предыдущей версией стало внедрение в синтезатор речи встроенной функции мультиспикера. Это значит, что модель умеет генерировать речь из сотни уникальных голосов [11].

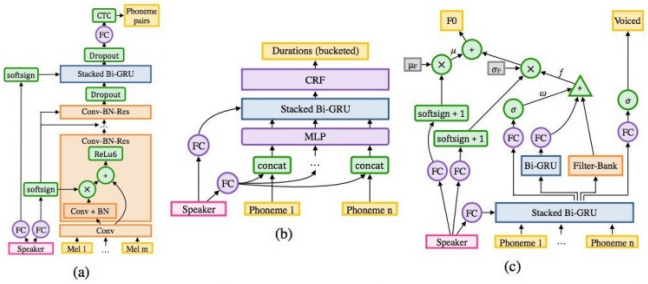

Рисунок 10 – архитектура DeepVoice 2
a – модель сегментации фонем, b – модель длительности, в – частотная модель

Модель сегментации является сверточно-рекуррентной сетью. Самым значимым отличием по сравнению с DeepVoice является включение пакета нормализации и остаточных соединений в сверточные слои [12].

Модель длительности предназначена для прогнозирования длительности фонем, сгруппированных в сегменты, с использованием модели условных случайных полей.

Модель частотного прогнозирования построена из сверточных и GRU уровней для прогнозирования основной частоты.

Плюсы DeepVoice 2:

- может выучить множество уникальных голосов;
- оценка MOS – 3,53 для американского английского.

5. DeepVoice 3.

В DeepVoice 3 была реализована мультисистема, в которой нет блоков сегментации, длительности и частоты. Одна модель генерирует мел-спектрограмму или другие аудио функции, которые должны быть декодированы в аудио сигнал WaveNet или другого вокодера [13].

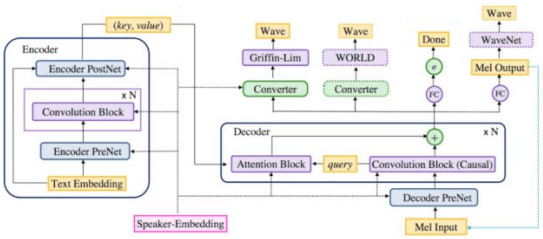

Рисунок 11 – архитектура DeepVoice 3

DeepVoice 3 – полностью сверточная архитектура

кодера-декодера с механизмом внимания.

Блок свертки (рис. 12) состоит из одномерной свертки и GRU. Этот блок обрабатывает, для кодирования, скрытые представления текста и аудио.

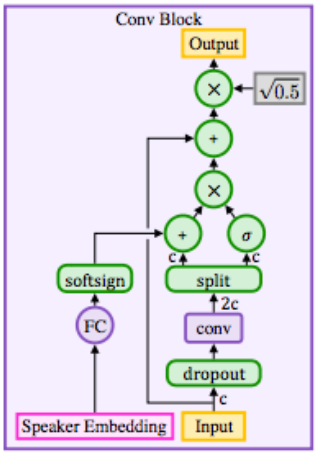

Рисунок 12 – блок свертки

Блок внимания (рис. 13) является монотонным. Он использует вектор запроса (скрытые состояния декодера) и векторы ключей для каждого временного шага от кодера для вычисления веса внимания, а затем выводит вектор контекста, вычисленный как средневзвешенное значение полученных весов. Значение вектора здесь означает масштабированную сумму основного вектора и вектора встраивания текста.

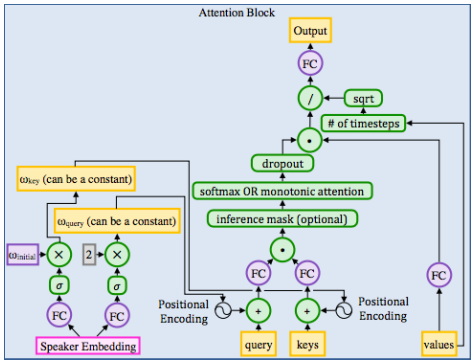

Рисунок 13 – блок внимания

Плюсы DeepVoice 3:

- преобразует входной текст в спектрограммы;
- модель использует механизм внимания, чтобы ввести монотонное выравнивание;
- модель быстро обучается (достаточно 500000 итераций. К примеру, Tacatron требует около двух миллионов итераций);
- оценка MOS – 3,78

6. Tacatron 2.

Архитектура Tacatron 2 (рис. 14) аналогична первой версии. Кодер стал более простым, остались только слои свертки и двунаправленный LSTM (сеть долгой краткосрочной памяти) для кодирования вложений символов. Аддитивное внимание было заменено на механизм чувствительного к расположению внимания. Часть декодера стала проще [14].

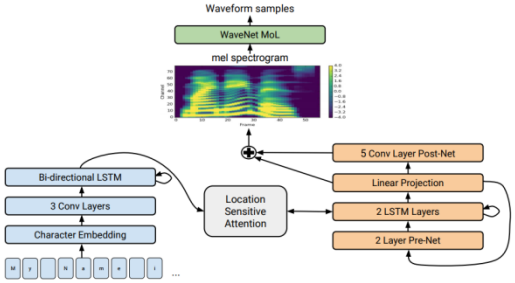

Рисунок 14 – архитектура tacotron 2

Главное обновление Tacatron 2 то, что он использует WaveNet MoL для обработки прогнозов спектрограмм и формирования речевых сигналов. WaveNet MoL отличается от WaveNet тем, что он использует 10-компонентные распределения Mixture of Logistics для генерации аудио сэмплов.

Оценка MOS – 4,53 для американского английского. Этот показатель максимально приближен к MOS оценке профессионально записанной речи, которая составляет 4,58.

V. Заключение

На сегодняшний день существует множество архитектур преобразования текста в речь. К сожалению, все еще существуют проблемы, с которыми сталкиваются разработчики данных систем. Самой главной проблемой является выразительность синтезированной речи. Для решения данной проблемы необходимо: улучшить определение системой правильность простановки ударения (к примеру слова омографы, одинаковые по написанию, но с разными ударениями), определение правильной интонации (восклицательные, вопросительные или положительные предложения). Также стоят задачи улучшения нормализации текста и преобразования текста в фонетическое представление.

Начать изучение данной область стоит с более современных систем, таких как: Tacotron 2 и Deep Voice 3. У данных систем есть возможность скачать исходный код, так как он находится в открытом доступе. Также в системах можно отметить качественность синтезированной речи, которая приближенна к профессионально записанной речи.

Библиография

[1] WikipediA [Электронный ресурс]/ Sound/ URL: https://en.wikipedia.org/wiki/Sound (дата обращения 27.02.2020)
[2] WikipediA [Электронный ресурс]/ Speech synthesis/ URL: https://en.wikipedia.org/wiki/Speech_synthesis (дата обращения 27.02.2020)
[3] WikipediA [Электронный ресурс]/ Синтез речи/ URL: https://ru.wikipedia.org/wiki/Синтез_речи (дата обращения 27.02.2020)
[4] WikipediA [Электронный ресурс]/ Human voice/ URL: https://en.wikipedia.org/wiki/Human_voice (дата обращения 27.02.2020)
[5] WikipediA [Электронный ресурс]/ Голос/ URL: https://ru.wikipedia.org/wiki/Голос (дата обращения 27.02.2020)
[6] Google Patents [Электронный ресурс]/ Intonation adjustment in text-to-speech systems/ Shankar Narayan/ URL: https://patents.google.com/patent/US5642466A/en (дата обращения 27.02.2020)
[7] Google Patents [Электронный ресурс]/ Text to speech/ Edwin R. AddisonH. Donald WilsonGary MarpleAnthony H. HandalNancy

Krebs/ URL: https://patents.google.com/patent/US6865533B2/en (дата обращения 27.02.2020)
[8] arXiv.org [Электронный ресурс]/ WAVENET: A GENERATIVE MODEL FOR RAW AUDIO/ Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu/ URL: https://arxiv.org/pdf/1609.03499.pdf (дата обращения 11.02.2020)
[9] arXiv.org [Электронный ресурс]/ Deep Voice: Real-time Neural Text-to-Speech/ Sercan O. Arik, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, Shubho Sengupta, Mohammad Shoeybi/ URL: https://arxiv.org/pdf/1702.07825.pdf (дата обращения 11.02.2020)
[10] arXiv.org [Электронный ресурс]/ Tacotron: Towards End-to-End Speech Synthesis/ Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, Rif A. Saurous/ URL: https://arxiv.org/pdf/1703.10135.pdf (дата обращения 27.02.2020)
[11] B arXiv.org [Электронный ресурс]/ Deep Voice 2: Multi-Speaker Neural Text-to-Speech/ Sercan Arik, Gregory Diamos, Andrew Gibiansky, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, Yanqi Zhou/ URL: https://arxiv.org/pdf/1705.08947.pdf (дата обращения 27.02.2020)
[12] Semantic Scholar [Электронный ресурс]/ Speech Enhancement for a Noise-Robust Text-to-Speech Synthesis System using Deep Recurrent Neural Networks/ Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, Junichi Yamagishi/ URL: https://pdfs.semanticscholar.org/ed99/08f71d6521a45093ffc0f9365315c1183604.pdf (дата обращения 27.02.2020)
[13] arXiv.org [Электронный ресурс]/ Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning/ Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O. Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, John Miller/ URL: https://arxiv.org/pdf/1710.07654.pdf (дата обращения 27.02.2020)
[14] arXiv.org [Электронный ресурс]/ Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions/ Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, Yonghui Wu / URL: https://arxiv.org/pdf/1712.05884.pdf (дата обращения 20.02.2020)
[15] Google Patents [Электронный ресурс]/ Method and system for statistic-based distance definition in text-to-speech conversion/ Wei Z W ZhangXi Jun MaLing JinHai Xin Chai/ URL: https://patents.google.com/patent/US7590540B2/en (дата обращения 27.02.2020)
[16] Google Patents [Электронный ресурс]/ System and method for distributed text-to-speech synthesis and intelligibility/ Jun XuTeck Chee LEE/ URL: https://patents.google.com/patent/US9761219B2/en (дата обращения 27.02.2020)
[17] Google Patents [Электронный ресурс]/ Systems and methods for text-to-speech synthesis using spoken example/ Andy AaronRaimo BakisEllen M. EideWael M. Hamza/ URL: https://patents.google.com/patent/US8886538B2/en (дата обращения 27.02.2020)
[18] Google Patents [Электронный ресурс]/ System and method of performing user-specific automatic speech recognition/ Bojana GajicShrikanth Sambasivan NarayananSarangarajan ParthasarathyRichard Cameron RoseAaron Edward Rosenberg / URL: https://patents.google.com/patent/US9058810B2/en (дата обращения 27.02.2020)
[19] arXiv.org [Электронный ресурс]/ Deep Speech: Scaling up end-to-end speech recognition/ Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, Andrew Y. Ng/ URL: https://arxiv.org/pdf/1412.5567.pdf (дата обращения 27.02.2020)
[20] arXiv.org [Электронный ресурс]/ Segmental Recurrent Neural Networks for End-to-end Speech Recognition/ Liang Lu, Lingpeng Kong, Chris Dyer, Noah A. Smith, and Steve Renals/ URL: https://arxiv.org/pdf/1603.00223.pdf (дата обращения 27.02.2020)
[21] CiteSeerX [Электронный ресурс]/ Hidden Markov Model based Speech Synthesis: A Review/ Sangramsing Kayte, Monica Mundada, Jayesh Gujrath/ URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.740.1357&rep=rep1&type=pdf (дата обращения 27.02.2020)
[22] Semantic Scholar [Электронный ресурс]/ Speech Enhancement for a Noise-Robust Text-to-Speech Synthesis System using Deep Recurrent Neural Networks/ Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, Junichi Yamagishi/ URL: https://pdfs.semanticscholar.org/ed99/08f71d6521a45093ffc0f9365315c1183604.pdf (дата обращения 27.02.2020)

Review of existing text-to-speech algorithms

N.S. Kireev, E.A. Ilyushin

***Annotation* — Scientists have long been working on algorithms for translate text written in natural language into speech. But the quality of work these algorithms left much to be desired until the moment when the application of deep learning methods was not possible. With the advent of the necessary computing resources and the accumulation of a sufficient amount of data for training, these methods have become widely used in machine learning in general and, of course, in speech synthesis in particular. A significant improvement in the quality of the work of text-to-speech algorithms has led to their widespread use, namely in mobile devices, smart speakers, voice assistants, etc. But it is worth noting that the algorithms of this class, developed at the moment, do not always correctly cope with the task. For example, they cannot always correctly emphasize or voice the necessary parts of the text with the necessary intonation. Thus, the study of methods and means of synthesizing speech has become even more relevant.**

There are many different ways to synthesize speech by text, such as parametric synthesis, compilation synthesis, subject-oriented synthesis, and full speech synthesis by the rules. The purpose of this work is to review existing algorithms for translating text to speech and conducting their comparative analysis. The main algorithms were considered: WaveNet, DeepVoice, Tacatron, DeepVoice 2, DeepVoice 3 and Tacatron 2. In the course of their comparison, it was determined that the best at the moment are DeepVoice 3 and Tacatron 2, since the assessments of the quality of their work are closest to professionally recorded speech.

***Keywords*—speech, text-to-speech, speech synthesis, nlp**

REFERENCES

[1] Sound [Online]. Available: https://en.wikipedia.org/wiki/Sound

[2] Speech synthesis [Online]. Available: https://en.wikipedia.org/wiki/Speech_synthesis

[3] Sintez rechi [Online]. Available: https://ru.wikipedia.org/wiki/Sintez_rechi

[4] Human voice [Online]. Available: https://en.wikipedia.org/wiki/Human_voice

[5] Voice [Online]. Available: https://ru.wikipedia.org/wiki/Voice

[6] Shankar Narayan. (1997, June 24). Intonation adjustment in text-to-speech systems/ Shankar Narayan [Online]. Available: https://patents.google.com/patent/US5642466A/en

[7] Edwin R. AddisonH. Donald WilsonGary MarpleAnthony H. HandalNancy Krebs. (2005 March 8). Text to speech [Online]. Available: https://patents.google.com/patent/US6865533B2/en

[8] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu. (2016, September 19). WAVENET: A GENERATIVE MODEL FOR RAW AUDIO [Online]. Available: https://arxiv.org/pdf/1609.03499.pdf

[9] Sercan O. Arik, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, Shubho Sengupta, Mohammad Shoeybi. (2017, March 7). Deep Voice: Real-time Neural Text-to-Speech [Online]. Available: https://arxiv.org/pdf/1702.07825.pdf

[10] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, Rif A. Saurous. (2017, April 6). Tacotron: Towards End-to-End Speech Synthesis [Online]. Available: https://arxiv.org/pdf/1703.10135.pdf

[11] Sercan Arik, Gregory Diamos, Andrew Gibiansky, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, Yanqi Zhou. (2017, Septembere 20). Deep Voice 2: Multi-Speaker Neural Text-to-Speech [Online]. Available: https://arxiv.org/pdf/1705.08947.pdf

[12] Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, Junichi Yamagishi. (2016, September 8). Speech Enhancement for a Noise-Robust Text-to-Speech Synthesis System using Deep Recurrent Neural Networks [Online]. Available: https://pdfs.semanticscholar.org/ed99/08f71d6521a45093ffc0f9365315c1183604.pdf

[13] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O. Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, John Miller. (2018, February 22). Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning [Online]. Available: https://arxiv.org/pdf/1710.07654.pdf

[14] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, Yonghui Wu. (2018, September 16). Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions [Online]. Available: https://arxiv.org/pdf/1712.05884.pdf

[15] Wei Z W ZhangXi Jun MaLing JinHai Xin Chai. (2009, September 15). Method and system for statistic-based distance definition in text-to-speech conversion [Online]. Available: https://patents.google.com/patent/US7590540B2/en

[16] Jun XuTeck Chee LEE. (2017, September 12). System and method for distributed text-to-speech synthesis and intelligibility [Online]. Available: https://patents.google.com/patent/US9761219B2/en

[17] Andy AaronRaimo BakisEllen M. EideWael M. Hamza. (2014, November 11). Systems and methods for text-to-speech synthesis using spoken example [Online]. Available: https://patents.google.com/patent/US8886538B2/en

[18] Bojana GajicShrikanth Sambasivan NarayananSarangarajan ParthasarathyRichard Cameron RoseAaron Edward Rosenberg. (2015, June 16). System and method of performing user-specific automatic speech recognition [Online]. Available: https://patents.google.com/patent/US9058810B2/en

[19] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, Andrew Y. Ng. (2014, December 19). Deep Speech: Scaling up end-to-end speech recognition [Online]. Available: https://arxiv.org/pdf/1412.5567.pdf

[20] Liang Lu, Lingpeng Kong, Chris Dyer, Noah A. Smith, and Steve Renals. (2016, June 20). Segmental Recurrent Neural Networks for End-to-end Speech Recognition [Online]. Available: https://arxiv.org/pdf/1603.00223.pdf

[21] Sangramsing Kayte, Monica Mundada, Jayesh Gujrath. (2015, November). Hidden Markov Model based Speech Synthesis: A Review [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.740.1357&rep=rep1&type=pdf

[22] Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, Junichi Yamagishi. (2016, September 8-12). Speech Enhancement for a Noise-Robust Text-to-Speech Synthesis System using Deep Recurrent Neural Networks [Online]. Available: https://pdfs.semanticscholar.org/ed99/08f71d6521a45093ffc0f9365315c1183604.pdf