УДК 004.934.5

РАЗРАБОТКА ПОЛНОЦЕННОЙ СИСТЕМЫ СИНТЕЗА ГОЛОСА С ПОМОЩЬЮ НЕЙРОННОЙ СЕТИ

М.А. Ческидова, К.А. Сиденко, Т.В. Карпета

Генерация естественной речи из текста остается сложной задачей, несмотря на десятилетия исследований. Однако преобразование текста в речь (англ. «Text-To-Speech») получило большое распространение именно в последние несколько лет. Крупные компании используют TTS системы в своих голосовых помощниках, колл-центры автоматизируют обращения, люди всё чаще предпочитают слушать аудиокниги, вместо их прочтения. Помимо этого, синтез речи может служить помощником для людей, у которых есть серьезные проблемы со способностью к разговорной речи. В данной статье рассматривается ансамбль нейросетевых моделей, которые образуют полноценную систему синтеза речи, включающую в себя энкодер, синтезатор и вокодер.

Ключевые слова: синтез голоса, просодии голоса, энкодер, синтезатор, вокодер.

В последнее время достижения в области генерации речи из текста (text-to-speech) показывают, что нейросетевой подход добился определенных успехов. Можно производить качественную генерацию с выделением просодии голоса и его речевых особенностей. Проблема заключается в том, что в основном эти модели работают с такими языками как английский и китайский, а русскоязычные модели хоть и существуют, но далеки по качеству от иноязычных аналогов.

Данная работа направлена на то, чтобы разработать систему синтеза речи на русском языке. Smart Open Virtual Assistant (SOVA) – одни из первых предоставили полностью открытую TTS систему на русском языке. Представленная реализация основана на оригинальной архитектуре Tacotron [1], однако в ней реализованы различные подходы для улучшения качества синтезируемой речи. Это модуль для предобработки текстовых данных, расстановки ударений в словах, перевода символов в числовой вектор. На данный момент он поддерживает только русский и английский языки, но есть возможность добавления своих языков, а также добавления своих обработчиков текста.

Однако данная реализация не может похвастаться многоголосым синтезом речи. Несмотря на это, модель вполне может использоваться как точка

отправления для построения еще более качественной многоголосой системы синтеза речи на русском языке. Поэтому нашей задачей было собрать такую систему, которая могла бы генерировать большой объем речи разными голосами.

Для обучения подобному в систему добавляется энкодер – нейросетевая модель, которая преобразует сказанную речь в числовой вектор фиксированной длины. Благодаря ей возможно различие голосов разных дикторов.

Помимо добавления энкодера в системе был заменен вокодер. Вокодер WaveGlow [2] хоть и показывает отличные результаты, но является достаточно объемной моделью. Поэтому он был заменен на WaveGrad [3], который весит почти в 2 раза меньше своего аналога.

Общая схема генерации речи. Синтез речи состоит из 4 разных этапов: преобразование текста, выделение просодии голоса, создание мел-спектрограммы и преобразование мел-спектрограммы в синтезированную речь. Данные этапы представлены на рисунке 1.

На первом этапе происходит выделение просодии голоса при помощи энкодера. Просодии голоса – речевые особенности говорящего, включающие в себя тембр, высоту, громкость и другие особенности.

На втором этапе происходит создание мел-спектрограммы при помощи синтезатора. На вход подаются вектор текста и полученный вектор, содержащий просодии голоса. На выходе получается мел-спектрограмма, отображающая зависимость мощности сигнала от времени.

На третьем этапе происходит преобразование мел-спектрограммы в синтезированную речь при помощи вокодера. На вход подается мел-спектрограмма, а на выходе получается синтезированная аудиозапись.

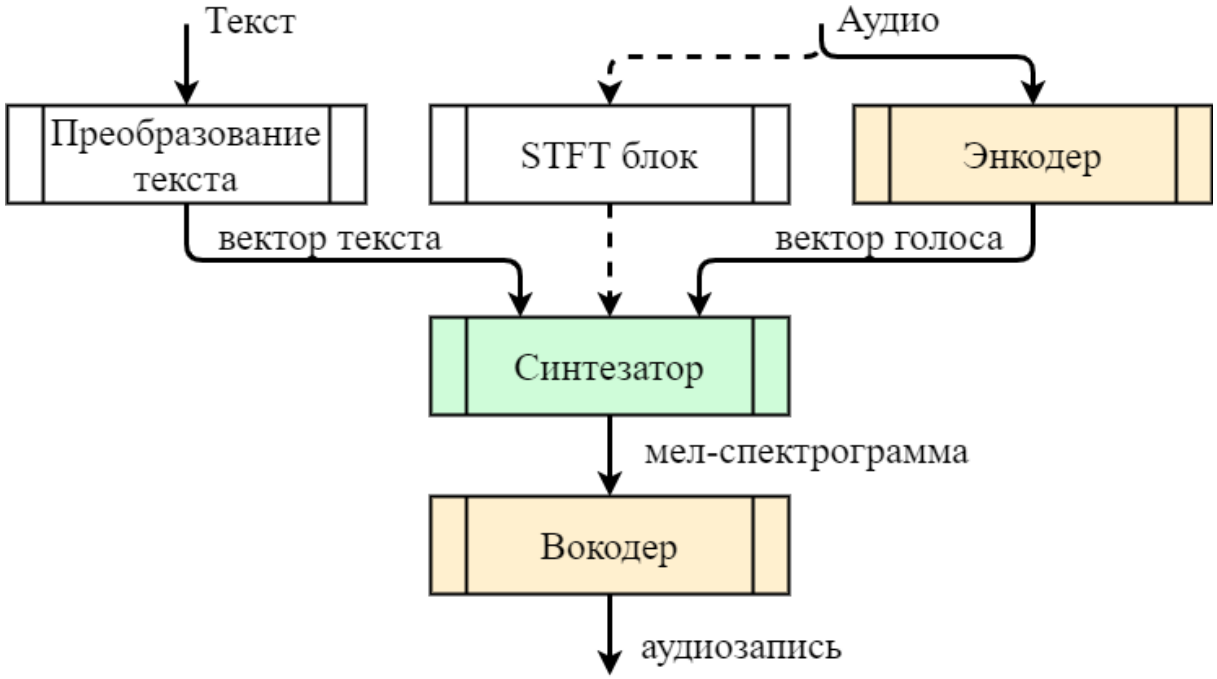

Рисунок 1 – Общая схема системы синтеза речи

Обучение энкодера. Для обучения в качестве источника набора данных использовались такие наборы данных как Common Voice, LibriSpeech и часть данных была собрана при помощи аудиокниг. Common Voice – многоязычный набор данных голосов с открытым исходным кодом, который в русскоязычной версии содержит около 124 часов аудиозаписей и 1638 носителей. Набор данных LibriSpeech содержит около 98 часов аудиоданных на русском языке. Собранные данные преобразуются путем удаления некачественных аудиозаписей с неразборчивой речью, удаляются фрагменты, содержащие шумы, и обрезаются отрезки, не содержащие речь.

Постановка задачи для энкодера выглядит следующим образом. Пусть X – матрица размером $N \times M$, где N – количество дикторов, M – количество высказываний, и x_{ij} – j-е высказывание i-го диктора, Y – матрица, в которой каждому элементу $y_{ij} \in Y$ соответствует вектор вещественных чисел, состоящий из 256 компонент $y_{ij} = \left(y_{ij}^1, y_{ij}^2, \dots, y_{ij}^{256}\right)$. Вектор говорящего определяется по следующей формуле:

$$y_{ij} = \frac{f(x_{ij}; w)}{\left|\left|f(x_{ij}; w)\right|\right|}, \tag{1}$$

где x_{ij} – j-е высказывание i-го диктора; w – веса.

После подсчета векторов дикторов необходимо вычислить их центроиды. Центроид – усредненный вектор высказываний диктора [4], вычисляемый по следующей формуле:

$$c_i = \frac{1}{M}\sum_{m=1}^{M} y_{im}, \qquad \text{при } k \neq i, \tag{2}$$

$$c_i^{-j} = \frac{1}{M-1}\sum_{\substack{m=1 \\ m \neq j}}^{M} y_{im}, \qquad \text{при } k = i, \tag{3}$$

где i – номер диктора; j – номер высказывания; k – номер диктора, $k \in [0; N]$.

Затем необходимо найти матрицу подобия между всеми векторами-высказываниями и центроидами по следующей формуле:

$$S_{ij,k} = \begin{cases} w \cdot \cos\left(y_{ij}, c_i^{-j}\right) + b, & \text{при } k = i, \\ w \cdot \cos\left(y_{ij}, c_k\right) + b, & \text{при } k \neq i. \end{cases} \tag{4}$$

Для того чтобы правильно выделять просодии и верифицировать говорящего, необходимо чтобы вектор вложения был близок к центроиду говорящего и далек от других говорящих [5]. Окончательная функция потерь рассчитывается по следующей формуле:

$$L_E = \sum_{i,j}\left(-S_{ij,i} + \ln\left(\sum_{k=1}^{N} exp\left(S_{ij,k}\right)\right)\right). \tag{5}$$

Обучение происходило методом обратного распространения ошибки, для оптимизации функции потерь был выбран оптимизатор Adam. Для определения качества нейросетевой модели была выбрана метрика, называемая метрикой равной частотой ошибок (EER), которая рассчитывается по следующей формуле:

$$EER = \frac{FAR + FRR}{2}, \tag{6}$$

где $FAR = \frac{FA}{FA+TR}$ – доля неправильно верифицированных дикторов – самозванцев, классифицируемых как подлинные; $FRR = \frac{FR}{FR+TA}$ – доля истинных дикторов, классифицируемых как самозванцы; данные значения показывают, когда нейросетевая модель правильно верифицировала диктора (TA) или отклонила диктора (TR), а также неправильно верифицировала диктора (FA) или неправильно отклонила диктора (FR).

Графики изменения значения функции потерь и метрики качества представлены на рисунке 2. В результате обучения нейронной сети на полученном графике видно, что при увеличении количества итераций уменьшалась функция потерь. В результате тестирования нейронной сети метрика, называемая равной частотой ошибок, уменьшилось, что свидетельствует о качестве обученной модели. В результате работы нейронной сети были получены графики, представленные на рисунке 3. Различные дикторы представлены различными цветами. Здесь представлено 50 дикторов, у каждого по 16 высказываний. В процессе обучения 16 высказываний одного диктора становятся настолько близки друг к другу, что отображаются практически в одну точку.

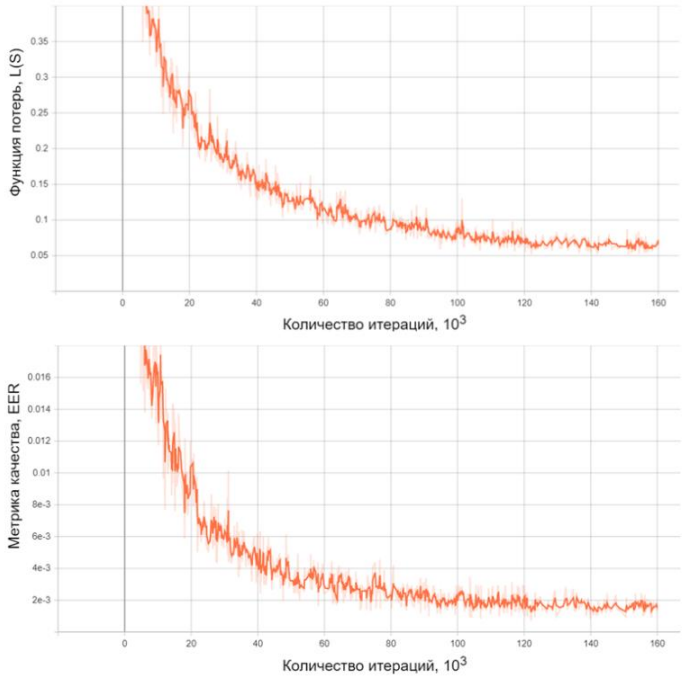

Рисунок 2 – Изменение значения функции потерь (сверху) и метрики качества (снизу)

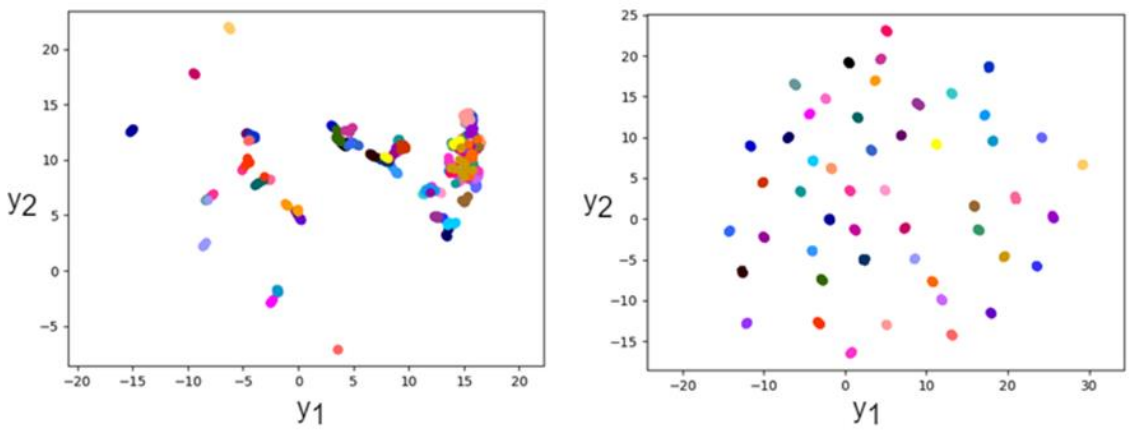

Рисунок 3 – Отображение векторов после 1000 (слева)
и 140000 (справа) итераций

Обучение синтезатора. Для обучения в качестве источника данных использовались такие наборы как Common Voice Corpus 6.1, RUSLAN, Natasha и часть данных была собрана при помощи аудиокниг. Набор данных Common Voice Corpus 6.1 включает более 80000 аудио на русском языке. Наборы данных RUSLAN и Natasha включают в себя 22200 и 11500 записей мужского и женского голоса. На вход нейросетевой модели подается текст и мел-спектрограмма, полученная из исходной аудиозаписи. Мел-спектрограмма – это спектрограмма, где частота выражена в мелах (рисунок 4).

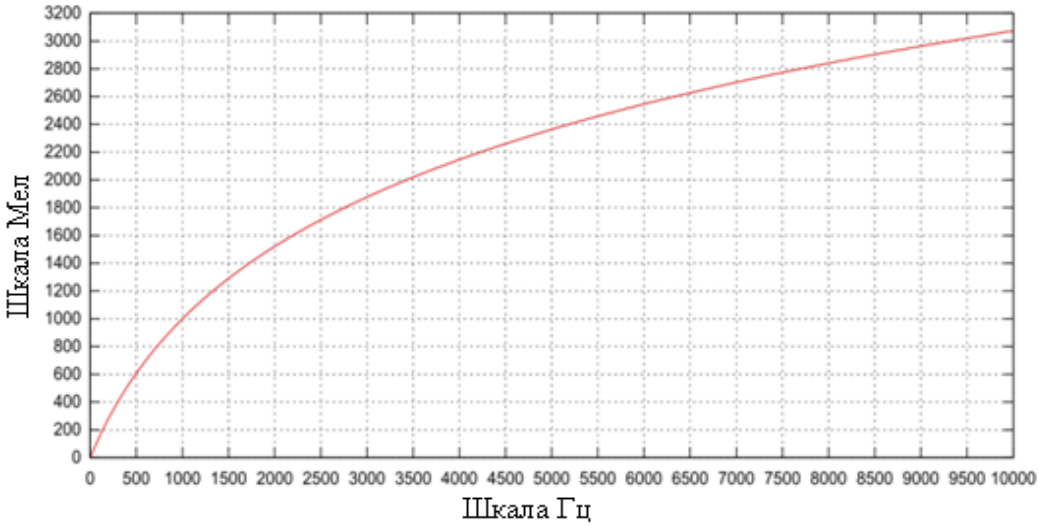

Рисунок 4 – График зависимости частоты от мел

В качестве основной функции потерь используется линейная комбинация трех функций потерь, представленных в формуле (7).

$$L_S = L_{mel} + L_{gate} + L_{ssim} \quad (7)$$

где L_{mel} – среднеквадратическая ошибка между синтезированной мел-спектрограммой и оригиналом; L_{gate} – бинарная кросс-энтропия с сигмоидальной функцией; L_{ssim} – функция ошибки, основанная на индексе структурного сходства.

Бинарная кросс-энтропия с сигмоидальной функцией, которая представлена в формуле (8), служит для контроля своевременной остановки синтеза мел-спектрограмм.

$$L_{gate} = -\frac{1}{n}\sum_{i=1}^{n}(y_i \cdot \log_2 \sigma(x_i) + (1 - y_i) \cdot \log_2\big(1 - \sigma(x_i)\big), \quad (8)$$

$$\sigma(x) = \frac{1}{1 + e^{-x}}, \quad (9)$$

где n – число точек на временном отрезке; x, y – соответствуют символам окончания синтеза для исходной мел-спектрограммы и синтезированной.

Среднеквадратичная ошибка (англ. «Mean Squared Error»), которая представлена в формуле (10), используется для оценки среднеквадратического различия между предсказанной мел-спектрограммой и исходной.

$$L_{mel} = MSE(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2, \quad (10)$$

где x и y соответствуют оригинальной и синтезированной мел-спектрограмме; n – число точек данных по всем переменным.

Индекс структурного сходства (Structural Similarity Index Measure), который выражается формулой (11), определяет различия между двумя изображениями. Область значений находится на промежутке от 0 до 1, где 1 означает полное сходство двух изображений, а 0 полное несоответствие.

$$L_{ssim} = 1 - SSIM(x, y) = 1 - \frac{\big(2\mu_x\mu_y + c_1\big)\big(2\sigma_{xy} + c_2\big)}{\big(\mu_x^2 + \mu_y^2 + c_1\big)\big(\sigma_x^2 + \sigma_y^2 + c_2\big)}, \quad (11)$$

где μ_x и μ_y – средние значения картинок x и y; σ_x и σ_y – среднеквадратичные отклонения для картинок x и y; σ_{xy} – ковариация x и y; $c_1 = (k_1 L)^2$, $c_2 = (k_2 L)^2$ – поправочные коэффициенты для стабилизации деления с малым знаменателем; $L = (2^{(N \text{ бит на пиксель})} - 1)$ – динамический диапазон пикселей; $k_1 = 0.01$ и $k_2 = 0.003$ – константы.

За основу модели синтезатора была взята архитектура нейронной сети Sova-TTS. Обучение происходило методом обратного распространения ошибки, для оптимизации функции потерь был выбран алгоритм Ranger. В результате обучения нейронной сети были получены графики изменения значения функции потерь на тренировочной и валидационной выборках, представленные на рисунке 5. Из графиков видно, что при увеличении количества итераций уменьшается функция потерь. Валидационная выборка не участвует в процессе обучения, поэтому оценка на её основе будет более репрезентативной и точной.

Эмпирически получено, что качество звука также повышается от числа итераций обучения. На каждом этапе валидации, сохранялась пара из оригинальной и синтезированной мел-спектрограмм. На основе полученных мел-спектрограмм генерировалось аудио при помощи вокодера.

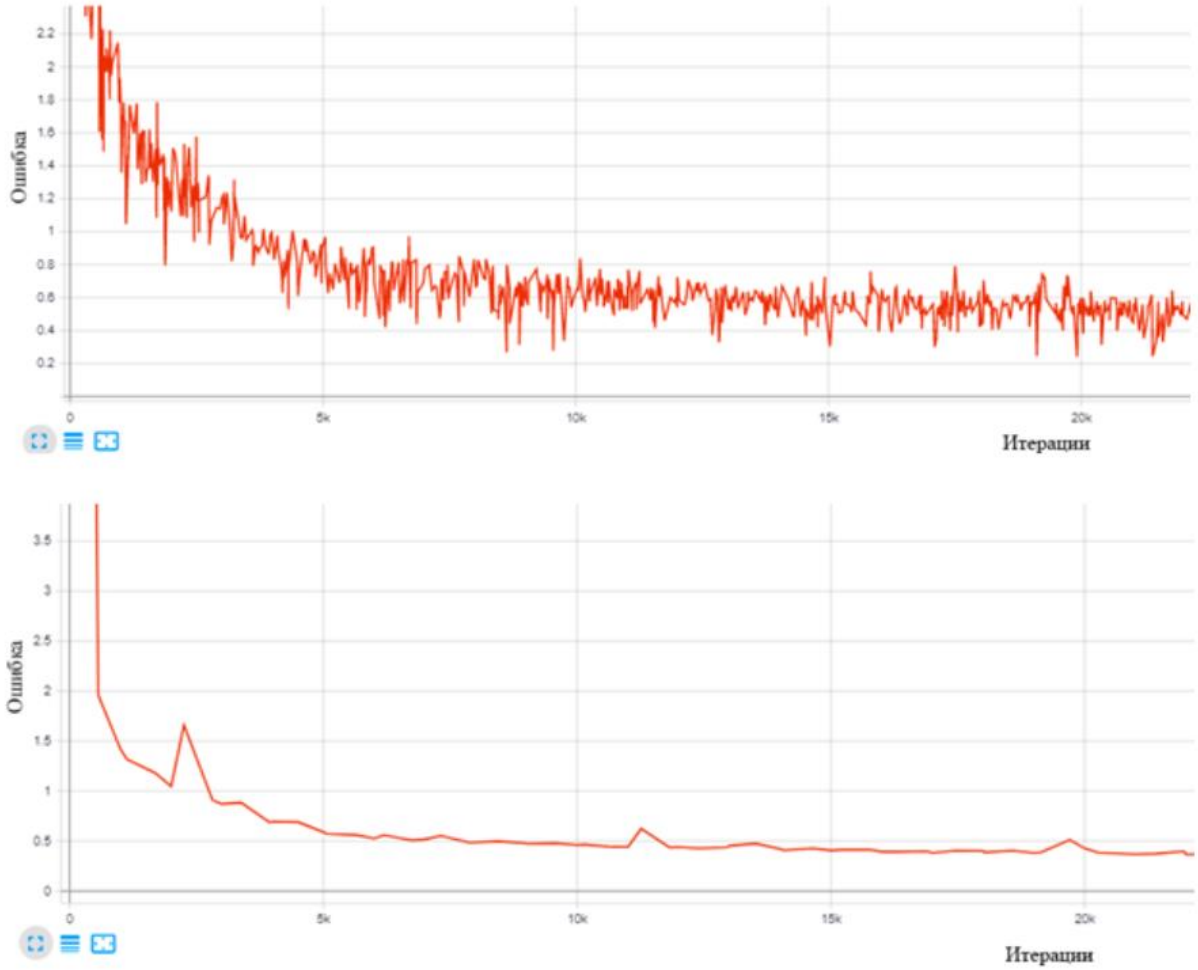

Рисунок 5 – График обучения нейронной сети на тренировочной (сверху) и на валидационной выборках (снизу)

Обучение вокодера. Для обучения в качестве источника данных использовались такие наборы как LJ speech data, Common Voice Corpus 6.1 и часть данных была собрана при помощи аудиокниг. Набор данных LJ speech data состоит из 13100 аудио фрагментов одного диктора.

В качестве функции потерь для обучения вокодера использовалась функция, представленная в формуле (12).

$$L_W = \frac{\sum_{i=1}^{n} \left| y_{true} - y_{predicted} \right|}{n}, \quad (12)$$

где y_{true} – это мел-спектрограмма аудио, которое подается на вход, представленная в формате массива весов, которые относятся к признакам аудио; $y_{predicted}$ – это мел-спектрограмма синтезированного аудио, так же представленная в формате массива весов; n – количество всех аудио.

Для оценивания качества сгенерированной речи использовалась оценка Mean Opinion Score (MOS) – усредненная оценка разборчивости речи. Это оценка от 1 до 5, где 5 означает что сгенерированная речь имеет хорошее качество и почти или совсем неотличима от человеческой, а 1 ставится, когда речь имеет очень низкое качество и совсем не похожа на человеческую.

Обучение происходило методом обратного распространения ошибки, для оптимизации функции потерь был выбран оптимизатор Adam. График изменения значения функции потерь представлен на рисунке 6.

Рисунок 6 – График обучения нейронной сети

Результат обучения нейронной сети WaveGrad был сравнен с результатами других нейросетевых вокодеров при помощи оценки MOS. Данные об оценках можно увидеть в таблице 1. В опросе приняли участие 100 человек, каждому из которых были даны на прослушивание аудиозаписи, сгенерированные вокодерами. По результатам сравнения было выявлено, что использованный вокодер WaveGrad генерирует аудио, которые воспринимаются человеческим слухом, как более приближенные к настоящему человеческому голосу.

Таблица 1

Вокодеры	MOS
Оригинальное аудио	4.45±0.04
WaveGrad	4.43±0.07
WaveGlow	4.13±0.247
WaveNet	3.97±0.05
WaveRNN	4.02±0.02

Выводы. В данной работе была описана система синтеза речи, которая является объединением 3 нейронных сетей, сочетающая в себе модифицированную модель нейронной сети Sova-TTS, энкодер для выделения просодии голоса и вокодер для синтеза аудио. Основным направлением работы было обеспечение поддержки нескольких дикторов, для синтеза речи. Приведенная архитектура позволяет синтезировать аудио голосом конкретного человека, с частичной передачей эмоциональности и просодии голоса. Одно из главных преимуществ собранной системы – заменяемость отдельных мо-

дулей, на более современные, без изменения общей архитектуры. В результате разработки были получены результаты, превосходящие большинство аналогов синтеза речи на русском языке.

Литература

1. J. Shen, *Natural tts synthesis by conditioning wavenet on mel spectrogram predictions* (International conference on acoustics, speech and signal processing, Calgary, Alberta, Canada, 2018), pp. 4779 – 4783.

2. R. Prenger, R. Valle and B. Catanzaro, *Waveglow: A flow-based generative network for speech synthesis* (International conference on acoustics, speech and signal processing, Brighton, East Sussex, UK, 2019), pp. 3617 – 3621.

3. Chen, N. WaveGrad: Estimating Gradients For Waveform Generation / N. Chen. – Режим доступа: https://arxiv.org/pdf/2009.00713.pdf. – 2020. – 15 p.

4. Arasteh, S.T. Generalized LSTM-based End-to-End Text-Independent Speaker Verification / S.T. Arasteh. – Режим доступа: https://arxiv.org/pdf/1803.05427.pdf. – 2020. – 7 p.

5. L. Wan, *Generalized end-to-end loss for speaker verification* (International conference on acoustics, speech and signal processing, Calgary, Alberta, Canada, 2018), pp. 4879 – 4883.