УДК 004.934.5
DOI: 10.25559/SITITO.17.202104.896-905

Научная статья

Многоголосый синтез естественной речи с использованием генеративных потоков

Д. С. Обухов[1,2]
[1] ФГБОУ ВО «Новосибирский государственный технический университет», г. Новосибирск, Российская Федерация
Адрес: 630073, Российская Федерация, г. Новосибирск, пр-т. К. Маркса, д. 20
[2] Dasha.AI, г. Новосибирск, Российская Федерация
Адрес: 630005, Российская Федерация, г. Новосибирск, ул. Писарева, д. 32
bstodin@gmail.com

Аннотация
Современные системы синтеза речи формируют естественную речь и имеют высокую производительность. Модели с использованием генеративных потоков, среди прочих, показали впечатляющие результаты, позволяя формировать разнообразные произнесения заданного текста. Однако они ориентированы на синтез голосом одного заданного диктора. Не смотря на предложенные недавно техники для учета нескольких спикеров при обучении, качество многоголосого синтеза речи оставляет желать лучшего. В данной работе предложены техники и приемы, позволяющие повысить качество многоголосого синтеза при использовании акустических моделей на основе генеративных потоков. В качестве одной из таких техник предложено из внешней системы получать информацию о выравнивании по оси времени между речевым аудио сигналом и текстовой последовательностью. Такая информация позволяет определить, в какой момент времени какой именно звук был произнесен и является необходимой для рассматриваемой параллельной системы синтеза речи, т.к. позволяет разрешить проблему несоответствия длин входной и выходной последовательностей. Внешняя система получения таких выравниваний для записей разных спикеров более точна, чем внутренние эвристики при обучении, т.к. способна обучаться на большем объеме данных и поэтому обладает большей обобщающей способностью. Другая предложенная техника заключается в том, чтобы использовать полученные из внешней системы вещественные векторы фиксированной размерности, содержащие информацию о спикере, т.н. эмбеддинги спикера. В данной работе рассматриваются эмбеддинги спикера, полученные из системы для решения задачи верификации спикера. Такие представления спикера обладают тем свойством, что эмбеддинги, полученные из речевых фрагментов одного спикера, расположены рядом в пространстве, а эмбеддинги, полученные из речевых фрагментов разных спикеров – далеко друг от друга. Благодаря таким представлениям спикера система синтеза лучше формирует речь голосами разных дикторов.

Ключевые слова: многоголосый синтез речи, генеративные потоки, выравнивание текста и спектрограммы, эмбеддинги спикера

Автор заявляет об отсутствии конфликта интересов.

Для цитирования: Обухов Д. С. Многоголосый синтез естественной речи с использованием генеративных потоков // Современные информационные технологии и ИТ-образование. 2021. Т. 17, № 4. С. 896-905. doi: https://doi.org/10.25559/SITITO.17.202104.896-905

Original article

Multi Speaker Natural Speech Synthesis Using Generative Flows

D. S. Obukhov[a,b]
[a] Novosibirsk State Technical University, Novosibirsk, Russian Federation
Address: 20 Prospekt K. Marksa, Novosibirsk 630073, Russian Federation
[b] Dasha.AI, Novosibirsk, Russian Federation
Address: 32 Pisareva St., Novosibirsk 630005, Russian Federation
bstodin@gmail.com

Abstract

Modern speech synthesis systems generate natural speech and have high performance. Models using generative flows, among others, have shown impressive results, allowing you to form a variety of speech pronunciation from a given text. However, they are focused on synthesizing the voice of one given speaker. Despite the recently proposed techniques for taking into account several speakers in training, the quality of multi speaker speech synthesis leaves much to be desired. This paper proposes techniques to improve the quality of multi speaker synthesis using acoustic models based on generative flows. As one of such techniques, it is proposed to obtain information on the alignment along the time axis between a speech audio signal and a text sequence from an external system. Such forced alignments allow you to determine at what point in time which sound was uttered and is necessary for the considered parallel speech synthesis system, since it allows you to solve the problem of mismatching the lengths of the input and output sequences. An external alignment system is more accurate than internal heuristics for training, since it is able to learn on a larger amount of data and therefore has a greater generalizing ability. Another proposed technique is to use real vectors of fixed dimension obtained from the external system, containing information about the speaker, the speaker embeddings. In this paper, speaker embeddings obtained from the system for solving the problem of speaker verification are considered. Such representations of a speaker have the property that embeddings obtained from speech fragments of one speaker are located side by side in space, and embeddings obtained from speech fragments of different speakers are far from each other. Due to such representations of the speaker, the synthesis system better forms speech with the voices of different speakers.

Keywords: multi speaker speech synthesis, generative flows, forced alignments, speaker embeddings

The author declares no conflict of interest.

For citation: Obukhov D.S. Multi Speaker Natural Speech Synthesis Using Generative Flows. *Sovremennye informacionnye tehnologii i IT-obrazovanie = Modern Information Technologies and IT-Education.* 2021; 17(4):896-905. doi: https://doi.org/10.25559/SITITO.17.202104.896-905

Введение

Синтез речи одна из актуальных задач в сфере речевых технологий и области машинного обучения [1, 2, 3, 4, 5, 6, 7, 8, 9]. Данная технология позволяет формировать речевой сигнал по заданному печатному тексту. Одни из первых систем синтеза речи применялись в медицине и помогали при чтении больным с нарушениями зрения [10], а позже стали использоваться людьми с дислексией и другими трудностями чтения, а также детьми, не имеющими достаточной грамотности. Сейчас область применения синтеза речи стала значительно шире и нашла свое применение в голосовых колонках, умных ассистентах и различных задачах бизнеса. Умение синтезировать речь голосами разных людей, а также высокое качество синтезированных записей, являются необходимым требованием для ряда задач, например, диалоговых систем. Такой синтез речи голосами нескольких разных спикеров, неотличимый от человеческой речи, называют многоголосый естественной синтез речи.

В современных системах синтез речи происходит в несколько этапов [1, 2, 11, 12]. Первым делом, идет извлечение из текста лингвистических признаков – фонем. Затем следует отображение лингвистической информации в акустическую – спектрограмму, эту роль выполняет, т.н. акустическая модель. Вокодер генерирует аудио сигнал из спектрограммы, что является завершающим этапом синтеза речи. Особый интерес представляет именно акустическая модель, так как процесс перевода лингвистической информации в акустическую является наиболее сложной задачей и в наибольшей степени отвечает за качество синтезируемой речи.

Современные акустические модели основываются на нейронных сетях [1, 2, 5, 6, 7, 13, 14]. Есть как подходы, которые ориентированы на синтез только одним заданным голосом, так и на синтез несколькими голосами (многоголосые подходы). Преимущество многоголосных подходов в сравнении с построением нескольких систем, каждая из которых синтезирует речь своим голосом, заключается как в том, что многоголосая система работает вычислительно эффективнее, поскольку требуют загрузки лишь одной модели в память, так и в том, что для обучения многоголосой системы требуется меньше данных для каждого из голосов, на которых система обучается, поскольку модель видит данные всех спикеров в совокупности.

Нейросетевые подходы решения задачи синтеза речи делятся на авторегрессионные [11, 15, 16, 17, 18] и параллельные [2, 3, 4, 5, 6, 7]. В авторегрессионных подходах речь формируется последовательно, от первого фрейма спектрограммы к последнему, поэтому время синтеза значительно увеличивается по мере увеличения длины текстовой последовательности. В параллельных решениях спектрограмма строится сразу целиком для последовательности текста любой длины, за счет этого синтез речи выполняется значительно быстрее, при этом качество речи не уступает авторегрессионным подходам [3, 4, 5]. Более того, ряд специфичных для авторегрессионных решений ошибок [19], таких как пропуски и повторения слов [2], в параллельных подходах даже не возникает. Поэтому в данной работе рассматривается многоголосая параллельная акустическая модель синтеза речи.

Архитектура параллельных систем синтеза речи также разделяется на подтипы. Наиболее успешные из них на основе трансформеров [2, 3, 4, 6] и генеративных потоков [5, 20]. Трансформерные модели являются детерминированными, то есть для заданного текста модель возвращает всегда один и тот же результат. Модели на основе потоков – генеративными. При синтезе такие модели используют некоторый шум, который позволяет получить на выходе разнообразные сигналы для одного и того же заданного текста. Такое поведение более свойственно для человеческой речи. Более того, в работе [5] есть возможность влиять на степень разнообразности речи, а также контролировать некоторые параметры, например скорость произношения.

Несмотря на то, что в работе [5] получен синтез высокого качества, рассматривается в первую очередь синтез речи одним спикером, а также используется не вся возможная информацию, которую можно извлечь из данных. Авторы [5] предложили не использовать продолжительности фонем из внешних систем, как это сделано в [2, 3, 4], а вместо этого предложили свой алгоритм. Однако, в случае многоголосого синтеза речи этот алгоритм совершает больше ошибок, чем при использовании продолжительностей, полученных из внешних систем, так как внешние системы способны использовать дополнительные данные для обучения. В данной работе предложено совместить архитектуру на основе генеративных потоков для параллельного построения спектрограммы, а также технику использования внешних систем для построения продолжительностей систем. Кроме того, в настоящей работе предложен новый, для параллельных подходов, способ учета информации о спикере, что является ключевым моментом при переходе от одного спикера к нескольким.

Структура работы следующая: во второй главе описывается архитектура модели синтеза речи на основе генеративных потоков, а также приведена математическая модель, которая лежит в основе генеративных потоков. В третьей главе описана техника выравнивания последовательностей текста и спектрограммы. В четвертой главе предлагается подход учета информации о спикерах. В пятой главе приведены результаты экспериментов. В шестой главе рассматриваются возможности применения модели с предложенной архитектурой в других задачах, помимо синтеза речи. Заключение подводится в седьмой главе.

Синтез речи на основе генеративных потоков

Потоковые генеративные модели недавно были успешно предложены для задачи генерации изображений в работе [21], и чуть позднее в задачах синтеза речи [5, 20, 22, 23]. Они позволяют оценить правдоподобие данных, применяя обратимые преобразования. Генеративные потоки обучаются так, чтобы максимизировать это правдоподобие.

Акустическая модель синтеза речи в настоящей работе состоит из нескольких модулей: текстовый кодировщик, кодировщик спикера, потоковый декодировщик и модуль предсказания продолжительностей произнесения фонем.

Текстовый кодировщик f_{enc} отображает последовательность токенов фонем $t = t_{1:Ttext}$ в скрытое векторное представление $h = h_{1:Ttext}$, где T_{text} обозначает длину последовательности

фонем. После текстового кодировщика два линейных слоя используются для получения статистик математического ожидания $\mu = \mu_{1:Ttext}$ и средне-квадратичного отклонения $\sigma = \sigma_{1:Ttext}$ априорного распределения потокового декодера. В настоящей работе архитектура текстового кодировщика состоит из прямо направленных трансформер блоков. Заметим, что такая архитектура идентична предложенной в работе [2] за исключением того, что скрытая размерность и число фильтров в слоях были увеличены.

Кодировщик спикера кодирует информацию о спикере в виде вещественного вектора фиксированной длины – эмбеддинга (представления) s. В данной работе рассматриваются разные техники построения эмбеддингов спикера: обучаемый эмбеддинг спикера и эмбеддинг спикера из внешней системы. Подробнее об этом в главе 4.

По аналогии с работой [5] в настоящей работе моделируется условное распределение спектрограмм $P_x(x \mid t, s)$ путем преобразования условного априорного распределения $P_z(z \mid t, s)$ через потоковый декодеровщик $f_{dec} : z \to x$, где x, t и s обозначают входную спектрограмму, текстовую последовательность и информацию о спикере, соответственно. Используя правило замены переменных, можно вычислить логарифм правдоподобия данных следующим образом:

$$\log P_x(x \mid t, s) = \log P_z(z \mid t, s) + \log \left| \det \frac{\partial f_{dec}^{-1}(x)}{\partial x} \right| \quad (1)$$

Априорное распределение P_Z в формуле (1) является изотропным многомерным распределением Гаусса, и все статистические данные априорного распределения, μ и σ, полученные, как отмечено выше, текстовым кодировщиком, соотносятся со спектрограммой за счет выравнивания A, полученного из внешней системы. $A(j) = i$, если j-ая фонема произносится на i-ом фрейме спектрограммы. Таким образом, априорное распределение можно выразить следующим образом:

$$\log P_Z(z \mid t, s; \theta) = \sum_{j=1}^{Tmel} \log \mathrm{N}(z_j; \mu_{A(j)}, \sigma_{A(j)}) \quad (2)$$

где T_{mel} обозначает продолжительность спектрограммы.

На этапе обучения параметры модели подбираются так, чтобы максимизировать логарифм правдоподобия:

$$\max_{\theta} L(\theta) = \max_{\theta} \log P_X(x \mid t, s, A; \theta) \quad (3)$$

Важным свойством потокового декодировщика является его двунаправленность. На этапе инференса модели генерация спектрограммы x происходит за счет применения к z прямого прохода по декодировщику. На этапе обучения для вычисления функции потерь по формуле (3) используются формулы (1) и (2), а значения z получаются за счет применения к x обратного прохода по декодировщику. Двунаправленность декодера также позволяет дополнительно выполнять другие задачи помимо синтеза речи, об этом в главе 6.

По аналогии с системами FastSpeech [2], FastPitch [3], FastSpeech 2 [4] для того, чтобы на этапе инференса предсказывать продолжительности фонем обучается дополнительный модуль – предсказатель продолжительностей фонем f_{dur}. Для каждого токена входной последовательности данный модуль предсказывает число $\log w$ – логарифм количества фреймов, на протяжении которых будет длиться соответствующая фонема. Для получения продолжительности фреймов w округляется до ближайшего целого. Обучение модуля предсказания продолжительностей фонем достигается за счет минимизации среднеквадратичной ошибки между продолжительностями, полученными из выравниваний внешней системы $d = d_{1:Ttext}$ и предсказанными:

$$d_i = \sum_{j=1}^{Tmel} 1_{A(j)=i}, i = 1, \ldots, T_{text} \quad (4)$$

$$L_{dur} = \left\| f_{dur}(sg(f_{enc}(t)), sg(s)) - d \right\|_2^2 \quad (5)$$

Модуль предсказания продолжительности фонем аналогичен предложенному в работе [5], кроме того в настоящей работе градиенты не распространяются как на текстовый кодировщик, так и на кодировщик спикера, в формуле 5 это обозначается за счет оператора $sg(.)$.

Выравнивание текста и спектрограммы

Выравнивание текста и спектрограммы **A** используется на этапе обучения акустической модели синтеза речи для того, чтобы сопоставить фонемы из текста и фреймы из спектрограммы. Они используются на этапе обучения акустической модели, формулы (2) и (4).

Продолжительности фонем **d**, в отличии от выравниваний, не требуют спектрограммы. На этапе обучения продолжительности фонем могут быть получены из выравниваний по формуле (4). На этапе инференса они могут быть получены на основе текста и спикера модулем предсказания продолжительностей фонем.

Неточные выравнивания приводят к тому, что даже хорошо обучившись, качество синтеза будет низким, поскольку в местах ошибок модель обучается некорректным фонемам.

В работе [5] авторы предложили метод автоматического построения выравниваний между текстом и спектрограммой. Авторы акцентируют внимание на том, что использование их метода позволяет упростить процесс обучения, а также избавиться от ошибок, которые имеют внешние системы выравнивания. Тем не менее, их алгоритм не гарантирует точных выравниваний, а лишь находит решение, которое минимизирует ошибку для заданных данных.

В рамках данной работы было обнаружено, что при обучении многоголосных систем предложенный в работе [5] алгоритм – Monotonic Alignment Search (MAS) – допускает больше ошибок, чем внешние системы построения выравниваний. В частности, мы заметили, что фреймы, которые соответствуют последним фонемам некоторых слов, часто ошибочно соответствуют пробелам. Это обусловлено тем, что в случае многоголосого синтеза данные обучения гораздо более разнообразны, а среднее количество времени на одного спикера в данных обучения значительно меньше, чем в корпусе LJSpeech[1], на котором проводили эксперименты авторы работы [5]. В то же время, внешние системы построения выравниваний, не при-

[1] Ito K., Johnson L. The LJ Speech Dataset [Электронный ресурс]. URL: https://keithito.com/LJ-Speech-Dataset (дата обращения: 15.10.2021).

вязаны к данным, на которых происходит обучение акустической модели синтеза речи, за счет этого их можно обучить на порядок большем объеме и тем самым повысить их обобщающую способность.

В рамках данной работы обучена собственная модель для построения выравниваний, на основе смеси гауссовских моделей, на базе Kaldi Speech Recognition Toolkit [24]. Важно отметить, что при обучении этой модели мы дополнительно использовали пустые токены между каждой парой фонем, в последовательности соответствующей тексту. Пустым токеном является специальный дополнительный символ, который не соответствует никакой фонеме. Такой прием позволяет лучше обучить акустическую модель синтеза речи, даже если выравнивания из внешней системы содержат неточности, за счет того, что неточности становятся не на границах между фонемами, а на границах между фонемой и пустым токеном.

Учет информации о спикере

В рамках данной работы рассматривается два способа учета информации о спикере:

- обучаемые эмбеддинги спикеров;
- эмбеддинги спикеров из внешней системы.

Обучаемые эмбеддинги спикеров являются традиционным способом учета информации спикера. В упомянутых работах [2, 3, 4, 5] обобщение задачи на нескольких спикеров происходит именно таким образом. В рамках этого подхода, каждому уникальному идентификатору спикера из тренировочной выборки соответствует уникальный вектор характеристик, значения которого обновляются в процессе обучения.

Недостатком этого подхода является то, что такой способ учитывает только характеристики спикера общие для всех произнесенных записей. Однако, помимо этой общей информации, каждой записи соответствует уникальная информация, специфичная для заданного спикера. Например, один и тот же текст, человек может произнести с разным эмоциональным окрасом. Эту проблему решает пространство эмбеддингов спикера, в котором вектора полученные из записей одного спикера находились бы в одной окрестности, но не в одной точке, а вектора полученные из записей разных спикеров находились бы на большом расстоянии.

Ранее, в работе [25] авторы предложили использовать в синтезе речи эмбеддинги спикера, полученные из системы, обученной решать задачу верификации спикера. Задача верификации спикеров отвечает на бинарный вопрос, соответствует ли заданный фрагмент речи заданному спикеру или нет. Недавно в области этой задачи была опубликована модель ECAPA-TDNN с минимальной ошибкой среди существующих решений [26].

В настоящей работе в качестве эмбеддингов из внешней системы используется скрытое представление с предпоследнего слоя модели ECAPA-TDNN. В работе [26] эти эмбеддинги проходят через один линейный проективный слой и через финальную функцию активации. Поэтому в данной работе кодировщик спикера также состоит из одного линейного слоя. Кроме того, использование эмбеддингов спикера из внешней системы позволяет синтезировать речь даже спикерами, которые не встречались при обучении модели, об этом в главе 6.

Помимо выбранной модели получения эмбеддингов спикера и техники их учета, настоящая работа отличается от работы [25] тем, что рассматривается параллельная, а не авторегрессионная акустическая модель.

Эксперименты и результаты

Во всех экспериментах для обучения были использованы открытые англоязычные данные. В таблице 1 приведена информация по каждому используемому набору данных.

Таблица 1. **Используемые для обучения датасеты**

Table 1. **Datasets used for training**

Набор данных	Количество записей	Количество часов обучения	Среднее количество часов на спикера
Blizzard 2013 [27]	147249	198.2	4.4
HiFi-TTS dataset [28]	323978	291.7	29.2
LibriTTS [29]	375086	585.8	0.26
LJSpeech[1]	13100	23.9	23.9
M AI Labs[2]	69853	143.6	35.9

Все эксперименты были проведены на машине со следующей конфигурацией: CPU: AMD Ryzen Threadripper 2950X 16-Core Processor; GPU: 3x NVidia GeForce RTX 2080 Ti.

Для обучения акустических моделей для каждого из спикеров было использовано не более двух часов данных. Кроме того в обучающую выборку были включены только спикеры, для которых имелось не менее 30 минут записанной речи. Обучение акустической модели длилось три дня.

Для обучения модели выравниваний были использованы все данные, приведенные в таблице 1. Обучение модели выравниваний продолжалось 20 часов.

Для преобразования спектрограмм в аудио сигнал был использован вокодер HiFi-GAN [30].

Для оценки систем был проведен MOS (mean opinion score, усредненная оценка опрашиваемых) тест. В рамках MOS теста на естественность речи асессору предлагалось прослушать аудиозапись и оценить их по шкале от 1 до 5, где 1 – это речь совершенно неестественная, 5 – речь неотличима от человеческой. Каждую из записей оценивали по 15 раз. Всего в оценке принимало участие 80 записей для каждой из моделей.

В рамках MOS теста на похожесть голоса асессору было предложено оценить насколько голос в синтезируемой записи похож на голос в оригинальной записи. Оценивание также происходило по шкале от 1 до 5, и каждая из 80 записей сравнивалась с записями из оригинальной речи по 15 раз.

[1] Там же.

[2] Solak I. The M-AILABS Speech Dataset [Электронный ресурс] // caito, 2019. URL: https://www.caito.de/2019/01/the-m-ailabs-speech-dataset (дата обращения: 15.10.2021).

В таблице 2 приведено сравнение обученных систем, а также других открытых решений обученных на тех же данных без каких-либо изменений.

Таблица 2. **Сравнение систем многоголосого синтеза речи**
Table 2. **Comparison of polyphonic speech synthesis systems**

	Естественность речи	Похожесть голоса
Оригинальная речь	4.19 ± 0.058	4.035 ± 0.069
Glow-TTS [5]	3.643 ± 0.075	3.593 ± 0.082
FastPitch [2]	3.556 ± 0.084	3.785 ± 0.078
Предложенное решение	3.947 ± 0.066	3.686 ± 0.083

В первой строке таблицы 2 приведена оценка для оригинальной человеческой речи. Во второй и третьей строках приведены оценки для моделей Glow-TTS [5] и FastPitch [2], обученных на тех же данных, без каких-либо архитектурных изменений. В четвертой строке – оценки для предложенного решения, включающего все описанные техники – использование внешних выравниваний и внешних эмбеддингов спикера.

Кроме того, было дополнительно проведено исследование, показывающее насколько применение предложенных техник по отдельности позволяет повысить естественность речи и похожесть голоса. В рамках этого исследования были обучены модели с применением только части предложенных приемов и также был проведен MOS тест. Результаты приведены в таблице 3.

Таблица 3. **Исследование влияния предложенных техник**
Table 3. **Study of the influence of proposed techniques**

	Естественность речи	Похожесть голоса
Оригинальная речь	4.163 ± 0.067	3.872 ± 0.05
Предложенное решение	3.859 ± 0.076	3.751 ± 0.053
Без использования эмбеддингов спикера	3.786 ± 0.086	3.701 ± 0.055
Без использования внешних продолжительностей	3.639 ± 0.085	3.639 ± 0.056

Другие применения

Предложенная архитектура позволяет использовать обученную для синтеза речи модель для других задач, такие как клонирование голоса и конверсия голоса.

Задача клонирования голоса заключается в том, чтобы синтезировать речевой сигнал образцом голоса, который не присутствовал в тренировочных данных модели синтеза речи. Образец целевого голоса обычно прилагается в виде аудиофайла. Предложенный подход позволяет использовать эмбеддинг спикера, полученный из аудиофайла с образцом целевого голоса, для синтеза речи заданным голосом. За счет того, что модель ECAPA-TDNN не ограничена никаким фиксированным набором спикеров, эмбеддинг для голоса произвольного спикера может быть получен. Таким образом, предложенный подход обобщается на задачу клонирования голоса.

Задача конверсии голоса заключается в преобразовании аудио сигнала с голосом исходного спикера в аудио сигнал с тем же лингвистическим содержанием, но произношением голосом целевого спикера. Модель ECAPA-TDNN позволяет получить эмбеддинги спикеров с исходным и целевым голосами – s_{source}, s_{target}. За счет двунаправленности потокового декодера, не составляет труда получить представление z для первоначального аудио сигнала x, в котором содержится речь исходного спикера s_{source}:

$$z = f_{dec}^{-1}(x \mid s_{source}) \quad (6)$$

Это представление не зависит от спикера, поскольку при обучении мы требовали, чтобы апостериорное распределения являлось изотропным многомерным распределением Гаусса со статистиками, полученными из текстового энкодера. В свою очередь, эти статистики не зависят от спикера, а зависят лишь от лингвистического содержания. Таким образом, применение прямого прохода по декодировщику, с условием заданным в виде эмбеддинга целевого спикера s_{target} позволяет получить аудио сигнал, с голосом целевого спикера и исходным лингвистическим содержанием:

$$x_{target} = f_{dec}(z \mid s_{target}) \quad (7)$$

Заключение

В данной работе были предложены техники по улучшению качества многоголосого синтеза речи. Использование внешних выравниваний текста и спектрограммы, а также внешних эмбеддингов спикера, позволило повысить естественность синтезируемой речи и синтезировать речь более похожим на заданный голос. Несмотря на это, результаты также показывают, что речь человека звучит более естественно, а степень клонирования голоса остается недостаточно высокой. В будущем можно использовать еще больше специфичной для дикторов информации из аудио, например, питч и энергию сигнала, для дальнейшего повышения естественности синтеза, а также придания синтезируемой речи звучания голосом заданного диктора. Также имеется интересный, как для исследования, так для повышения качества синтезируемой речи, сценарий использования данных из разных языков для обучения одной многоязыковой модели. Кроме того, предложенная архитектура предоставляет возможности для последующих экспериментов с задачами клонирования и конверсии голоса.

Благодарности

Автор выражает благодарность компании Dasha.AI за помощь в обеспечении вычислительными ресурсами для проведения экспериментов, а также выделенные средства для проведения оценочного тестирования синтезированных фрагментов речи.

Acknowledgements

The author thanks the Dasha.AI company for the assistance in providing computing resources for the experiments, as well as for the allocated funds for the evaluation testing of synthesized speech fragments.

Список использованных источников

[1] A Survey on Neural Speech Synthesis / X. Tan [и др.] // arXiv:2106.15561. 2021. doi: https://doi.org/10.48550/arXiv.2106.15561

[2] FastSpeech: fast, robust and controllable text to speech / Y. Ren [и др.] // Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, 2019. Article number: 285. P. 3171-3180. URL: https://dl.acm.org/doi/abs/10.5555/3454287.3454572 (дата обращения: 15.10.2021).

[3] Łańcucki A. Fastpitch: Parallel Text-to-Speech with Pitch Prediction // ICASSP 2021 – 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, ON, Canada: IEEE Press, 2021. P. 6588-6592. doi: https://doi.org/10.1109/ICASSP39728.2021.9413889

[4] FastSpeech 2: Fast and High-Quality End-to-End Text to Speech / Y. Ren [и др.] // arXiv:2006.0455. 2020. doi: https://doi.org/10.48550/arXiv.2006.04558

[5] Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search / J. Kim [и др.] // Advances in Neural Information Processing Systems (NeurIPS 2020) ; H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin (eds.). Vol. 33. Curran Associates, Inc., 2020. P. 8067-8077. URL: https://arxiv.org/pdf/2005.11129.pdf (дата обращения: 15.10.2021).

[6] Alignments: Efficient Feed-Forward Text-to-Speech System Without Explicit Alignment / Z. Zeng [и др.] // ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE Press, 2020. P. 6714-6718. doi: https://doi.org/10.1109/ICASSP40776.2020.9054119

[7] AdaSpeech: Adaptive Text to Speech for Custom Voice / M. Chen [и др.] // Proceedings of the Ninth International Conference on Learning Representations (ICLR 2021). Vienna, Austria, 2021. P. 1-10. URL: https://openreview.net/pdf?id=Drynvt7gg4L (дата обращения: 15.10.2021).

[8] Киреев Н. С., Ильюшин Е. А. Обзор существующих алгоритмов преобразования текста в речь // International Journal of Open Information Technologies. 2020. Т. 8, № 7. С. 84-90. URL: https://elibrary.ru/item.asp?id=43128091 (дата обращения: 15.10.2021).

[9] Лань Г., Фадеев А. С., Моргунов А. Н. Синтез фрагментов голоса человека на основе модели реконструкции частотных спектров // Доклады Томского государственного университета систем управления и радиоэлектроники. 2021. Т. 24, № 2. С. 14-20. doi: https://doi.org/10.21293/1818-0442-2021-24-2-14-20

[10] Cooper F. S., Gaitenby J. H., Nye P. W. Evolution of reading machines for the blind: Haskins Laboratories' research as a case history // Journal of Rehabilitation Research and Development. 1984. Vol. 21, issue 1. P. 51-87.

[11] Tacotron: Towards End-to-End Speech Synthesis / Y. Wang [и др.] // Proc. Interspeech. 2017. P. 4006-4010. doi: https://doi.org/10.21437/Interspeech.2017-1452

[12] Плюхина Г. А. Компьютерное моделирование иноязычного речевого акта // Научный резерв. 2019. № 1(5). С. 91-97. URL: https://elibrary.ru/item.asp?id=38469159 (дата обращения: 15.10.2021).

[13] Калиев А., Рыбин С. В. Синтез речи: прошлое и настоящее // Компьютерные инструменты в образовании. 2019. № 1. С. 5-28. doi: https://doi.org/10.32603/2071-2340-2019-1-5-28

[14] Израилова Э. С. Особенности машинного обучения средствами CNN в рамках синтеза речи // Вестник ГГНТУ. Технические науки. 2019. Т. 15, № 2(16). С. 29-35. doi: https://doi.org/10.34708/GSTOU.2019.16.2.004

[15] Mellotron: Multispeaker Expressive Voice Synthesis by Conditioning on Rhythm, Pitch and Global Style Tokens / R. Valle [и др.] // ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE Press, 2020. P. 6189-6193. doi: https://doi.org/10.1109/ICASSP40776.2020.9054556

[16] Flowtron: an Autoregressive Flow-based Generative Network for Text-to-Speech Synthesis / R. Valle, K. J. Shih, R. Prenger, B. Catanzaro // Proceedings of the Ninth International Conference on Learning Representations (ICLR 2021). Vienna, Austria, 2021. P. 1-17. URL: https://openreview.net/pdf?id=Ig53hpHxS4 (дата обращения: 15.10.2021).

[17] Using Vaes and Normalizing Flows for One-Shot Text-To-Speech Synthesis of Expressive Speech / V. Aggarwal [и др.] // ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE Press, 2020. P. 6179-6183. doi: https://doi.org/10.1109/ICASSP40776.2020.9053678

[18] Вовк И. Ю., Гогорян В. С. Система синтеза естественной речи для русского языка на базе глубоких нейронных сетей // Новые информационные технологии в автоматизированных системах. 2019. № 22. С. 142-150. URL: https://www.elibrary.ru/item.asp?id=41200093 (дата обращения: 15.10.2021).

[19] Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks / S. Bengio [и др.] // Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS'15). Vol. 1. Cambridge, MA, USA: MIT Press, 2015. P. 1171-1179. URL: https://proceedings.neurips.cc/paper/2015/file/e995f98d56967d946471af29d7bf99f1-Paper.pdf (дата обращения: 15.10.2021).

[20] Flow-TTS: A Non-Autoregressive Network for Text to Speech Based on Flow / C. Miao [и др.] // ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE Press, 2020. P. 7209-7213. doi: https://doi.org/10.1109/ICASSP40776.2020.9054484

[21] Kingma D. P., Dhariwal P. Glow: Generative Flow with Invertible 1x1 Convolutions // Advances in Neural Information Processing Systems (NeurIPS 2018) ; S. Bengio et al. (eds.). Vol. 31. Curran Associates, Inc., 2018. P. 1-15. URL: https://arxiv.org/pdf/1807.03039.pdf (дата обращения: 15.10.2021).

[22] Prenger R., Valle R., Catanzaro B. Waveglow: A Flow-based Generative Network for Speech Synthesis // ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE Press, 2019. P. 3617-3621. doi: https://doi.org/10.1109/ICASSP.2019.8683143

[23] Deep Voice: Real-time Neural Text-to-Speech / S. Ö. Arık [и др.] // Proceedings of the 34th International Conference on Machine Learning (PMLR) ; D. Precup, Y. W. Teh (eds.). 2017. Vol. 70. P. 195-204. URL: http://proceedings.mlr.press/v70/arik17a/arik17a.pdf (дата обращения: 15.10.2021).

[24] The Kaldi Speech Recognition Toolkit / D. Povey [и др.] // IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society, 2011. URL: https://infoscience.epfl.ch/record/192584 (дата обращения: 15.10.2021).

[25] Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis / Y. Jia [и др.] // Advances in Neural Information Processing Systems (NeurIPS 2018) ; S. Bengio et al. (eds.). Vol. 31. Curran Associates, Inc., 2018. P. 1-15. URL: https://arxiv.org/pdf/1806.04558.pdf (дата обращения: 15.10.2021).

[26] Desplanques B., Thienpondt J., Demuynck K. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification // Proc. Interspeech. 2020. P. 3830-3834. doi: https://doi.org/10.21437/Interspeech.2020-2650

[27] King S., Karaiskos V. The Blizzard Challenge 2013 // Proc. Blizzard Challenge 2013 Workshop. LTI at Carnegie Mellon University, 2013. URL: http://festvox.org/blizzard/bc2013/summary_Blizzard2013.pdf (дата обращения: 15.10.2021).

[28] Hi-Fi Multi-Speaker English TTS Dataset / E. Bakhturina, V. Lavrukhin, B. Ginsburg, Y. Zhang // Proc. Interspeech. 2021. P. 2776-2780. doi: https://doi.org/10.21437/Interspeech.2021-1599

[29] LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech / H. Zen [и др.] // arXiv:1904.02882. 2019. URL: https://arxiv.org/pdf/1904.02882.pdf (дата обращения: 15.10.2021).

[30] Kong J., Kim J., Bae J. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis // Advances in Neural Information Processing Systems (NeurIPS 2020) ; H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin (eds.). Vol. 33. Curran Associates, Inc., 2020. P. 1-12. URL: https://arxiv.org/pdf/2010.05646.pdf (дата обращения: 15.10.2021).

Поступила 15.10.2021; одобрена после рецензирования 29.11.2021; принята к публикации 05.12.2021.

Об авторе:

Обухов Дмитрий Сергеевич, аспирант, младший научный сотрудник, ФГБОУ ВО «Новосибирский государственный технический университет» (630073, Российская Федерация, г. Новосибирск, пр-т. К. Маркса, д. 20); инженер-исследователь, Dasha.AI (630005, Российская Федерация, г. Новосибирск, ул. Писарева, д. 32), **ORCID: https://orcid.org/0000-0001-5963-6327,** bstodin@gmail.com

Автор прочитал и одобрил окончательный вариант рукописи.

References

[1] Tan X., Qin T., Soong F., Liu T.-Y. A Survey on Neural Speech Synthesis. arXiv:2106.15561. 2021. (In Eng.) doi: https://doi.org/10.48550/arXiv.2106.15561

[2] Ren Y., Ruan Y., Tan X., Qin T., Zhao S., Zhao Z., Liu T.-Y. FastSpeech: fast, robust and controllable text to speech. *Proceedings of the 33rd International Conference on Neural Information Processing Systems.* Curran Associates Inc., Red Hook, NY, USA; 2019. Article number: 285. p. 3171-3180. Available at: https://dl.acm.org/doi/abs/10.5555/3454287.3454572 (accessed 15.10.2021). (In Eng.)

[3] Łańcucki A. Fastpitch: Parallel Text-to-Speech with Pitch Prediction. *ICASSP 2021 – 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE Press, Toronto, ON, Canada; 2021. p. 6588-6592. (In Eng.) doi: https://doi.org/10.1109/ICASSP39728.2021.9413889

[4] Ren Y., Hu C., Tan X., Qin T., Zhao S., Zhao Z., Liu T.- Y. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. arXiv:2006.0455. 2020. (In Eng.) doi: https://doi.org/10.48550/arXiv.2006.04558

[5] Kim J., Kim S., Kong J., Yoon S. Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search. In: Larochelle H., Ranzato M., Hadsell R., Balcan M.F., Lin H. (eds.) *Advances in Neural Information Processing Systems (NeurIPS 2020).* Vol. 33. Curran Associates, Inc.; 2020. p. 8067-8077. Available at: https://arxiv.org/pdf/2005.11129.pdf (accessed 15.10.2021). (In Eng.)

[6] Zeng Z., Wang J., Cheng N., Xia T., Xiao J. Alignтts: Efficient Feed-Forward Text-to-Speech System Without Explicit Alignment. *ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE Press, Barcelona, Spain; 2020. p. 6714-6718. (In Eng.) doi: https://doi.org/10.1109/ICASSP40776.2020.9054119

[7] Chen M., Tan X., Li B., Liu Y., Qin T., Zhao S., Liu T.-Y. AdaSpeech: Adaptive Text to Speech for Custom Voice. *Proceedings of the Ninth International Conference on Learning Representations (ICLR 2021).* Vienna, Austria; 2021. p. 1-10. Available at: https://openreview.net/pdf?id=Drynvt7gg4L (accessed 15.10.2021). (In Eng.)

[8] Kireev N.S., Ilyushin E.A. Review of existing text-to-speech algorithms. *International Journal of Open Information Technologies*. 2020; 8(7):84-90. Available at: https://elibrary.ru/item.asp?id=43128091 (accessed 15.10.2021). (In Russ., abstract in Eng.)

[9] Lan G., Fadeev A.S., Morgunov A.N. Synthesis of human voice fragments based on frequency spectra reconstruction. *Doklady Tomskogo gosudarstvennogo universiteta sistem upravlenija i radiojelektroniki* = Proceedings of the TUSUR University. 2021; 24(2):14-20. (In Russ., abstract in Eng.) doi: https://doi.org/10.21293/1818-0442-2021-24-2-14-20

[10] Cooper F.S., Gaitenby J.H., Nye P.W. Evolution of reading machines for the blind: Haskins Laboratories' research as a case history. *Journal of Rehabilitation Research and Development.* 1984; 21(1):51-87. (In Eng.)
[11] Wang Y., Skerry-Ryan R.J., Stanton D., Wu Y., Weiss R.J., Jaitly N., Yang Z., Xiao Y., Chen Z., Bengio S., Le Q., Agiomyrgiannakis Y., Clark R., Saurous R.A. Tacotron: Towards End-to-End Speech Synthesis. *Proc. Interspeech.* 2017. p. 4006-4010. (In Eng.) doi: https://doi.org/10.21437/Interspeech.2017-1452
[12] Plyuhina G.A. Computer modeling of foreign speech act. *Nauchnyij rezerv* = Scientific reserve. 2019; (1):91-97. Available at: https://elibrary.ru/item.asp?id=38469159 (accessed 15.10.2021). (In Russ., abstract in Eng.)
[13] Kaliev A., Rybin S.V. Speech synthesis: past and present. *Komp'juternye instrumenty v obrazovanii* = Computer Tools in Education. 2019; (1):5-28. (In Russ., abstract in Eng.) doi: https://doi.org/10.32603/2071-2340-2019-1-5-28
[14] Izrailova E.S. Features of machine learning by CNN within the speech synthesis. *Vestnik GGNTU. Tehnicheskie nauki* = Herald Of GSTOU. Engineering Sciences. 2019; 15(2):29-35. (In Russ., abstract in Eng.) doi: https://doi.org/10.34708/GSTOU.2019.16.2.004
[15] Valle R., Li J., Prenger R., Catanzaro B. Mellotron: Multispeaker Expressive Voice Synthesis by Conditioning on Rhythm, Pitch and Global Style Tokens. *ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE Press, Barcelona, Spain; 2020. p. 6189-6193. (In Eng.) doi: https://doi.org/10.1109/ICASSP40776.2020.9054556
[16] Valle R., Shih K.J., Prenger R., Catanzaro B. Flowtron: an Autoregressive Flow-based Generative Network for Text-to-Speech Synthesis. *Proceedings of the Ninth International Conference on Learning Representations (ICLR 2021).* Vienna, Austria; 2021. p. 1-17. Available at: https://openreview.net/pdf?id=Ig53hpHxS4 (accessed 15.10.2021). (In Eng.)
[17] Aggarwal V., Cotescu M., Prateek N., Lorenzo-Trueba J., Barra-Chicote R. Using Vaes and Normalizing Flows for One-Shot Text-To-Speech Synthesis of Expressive Speech. *ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE Press; 2020. p. 6179-6183. (In Eng.) doi: https://doi.org/10.1109/ICASSP40776.2020.9053678
[18] Vovk I.Yu., Gogoryan V.S. Natural speech synthesis system for the Russian language based on deep neural networks. *Novye informacionnye tehnologii v avtomatizirovannyh sistemah* = New information technologies in automated systems. 2019; (22):142-150. Available at: https://www.elibrary.ru/item.asp?id=41200093 (accessed 15.10.2021). (In Russ., abstract in Eng.)
[19] Bengio S., Vinyals O., Jaitly N., Shazeer N. Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks. *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS'15).* Vol. 1. MIT Press, Cambridge, MA, USA; 2015. p. 1171-1179. Available at: https://proceedings.neurips.cc/paper/2015/file/e995f98d56967d946471af29d7bf99f1-Paper.pdf (accessed 15.10.2021). (In Eng.)
[20] Miao C., Liang S., Chen M., Ma J., Wang S., Xiao J. Flow-TTS: A Non-Autoregressive Network for Text to Speech Based on Flow. *ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE Press; 2020. p. 7209-7213. (In Eng.) doi: https://doi.org/10.1109/ICASSP40776.2020.9054484
[21] Kingma D.P., Dhariwal P. Glow: Generative Flow with Invertible 1x1 Convolutions. In: Bengio S., Wallach H., Larochelle H., Grauman K., Cesa-Bianchi N., Garnett R. (eds.) *Advances in Neural Information Processing Systems (NeurIPS 2018).* Vol. 31. Curran Associates, Inc.; 2018. p. 1-15. Available at: https://arxiv.org/pdf/1807.03039.pdf (accessed 15.10.2021). (In Eng.)
[22] Prenger R., Valle R., Catanzaro B. Waveglow: A Flow-based Generative Network for Speech Synthesis. *ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE Press; 2019. p. 3617-3621. (In Eng.) doi: https://doi.org/10.1109/ICASSP.2019.8683143
[23] Arık S.Ö., Chrzanowski M., Coates A., Diamos G., Gibiansky A., Kang Y., Li X., Miller J., Ng A., Raiman J., Sengupta S., Shoeybi M. Deep Voice: Real-time Neural Text-to-Speech. In: Precup D., Teh Y.W. (eds.) *Proceedings of the 34th International Conference on Machine Learning (PMLR).* 2017; 70:195-204. Available at: http://proceedings.mlr.press/v70/arik17a/arik17a.pdf (accessed 15.10.2021). (In Eng.)
[24] Povey D., Ghoshal A., Boulianne G., Burget L., Glembek O., Goel N., Hannemann M., Motlicek P., Qian Y., Schwarz P., Silovsky J., Stemmer G., Vesely K. The Kaldi Speech Recognition Toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.* IEEE Signal Processing Society; 2011. Available at: https://infoscience.epfl.ch/record/192584 (accessed 15.10.2021). (In Eng.)
[25] Jia Y., Zhang Y., Weiss R.J., Wang Q., Shen J., Ren F., Chen Z., Nguyen P., Pang R., Moreno I.L. Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis. In: Bengio S., Wallach H., Larochelle H., Grauman K., Cesa-Bianchi N., Garnett R. (eds.) *Advances in Neural Information Processing Systems (NeurIPS 2018).* Vol. 31. Curran Associates, Inc.; 2018. p. 1-15. Available at: https://arxiv.org/pdf/1806.04558.pdf (accessed 15.10.2021). (In Eng.)
[26] Desplanques B., Thienpondt J., Demuynck K. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. *Proc. Interspeech.* 2020. p. 3830-3834. (In Eng.) doi: https://doi.org/10.21437/Interspeech.2020-2650
[27] King S., Karaiskos V. The Blizzard Challenge 2013. *Proc. Blizzard Challenge 2013 Workshop*. LTI at Carnegie Mellon University; 2013. Available at: http://festvox.org/blizzard/bc2013/summary_Blizzard2013.pdf (accessed 15.10.2021). (In Eng.)
[28] Bakhturina E., Lavrukhin V., Ginsburg B., Zhang Y. Hi-Fi Multi-Speaker English TTS Dataset. *Proc. Interspeech.* 2021. p. 2776-2780. (In Eng.) doi: https://doi.org/10.21437/Interspeech.2021-1599
[29] Zen H., Dang V., Clark R., Zhang Y., Weiss R.J., Jia Y., Chen Z., We Y. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. arXiv:1904.02882. 2019. Available at: https://arxiv.org/pdf/1904.02882.pdf (accessed 15.10.2021). (In Eng.)
[30] Kong J., Kim J., Bae J. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. In: Larochelle H., Ranzato M., Hadsell R., Balcan M.F., Lin H. (eds.) *Advances in Neural Information Processing Systems (NeurIPS 2020).* Vol. 33. Curran Associates, Inc.; 2020. p. 1-12. Available at: https://arxiv.org/pdf/2010.05646.pdf (accessed 15.10.2021). (In Eng.)

Submitted 15.10.2021; approved after reviewing 29.11.2021; accepted for publication 05.12.2021.

About the author:

Dmitry S. Obukhov, Postgraduate Student, Junior researcher, Novosibirsk State Technical University (20 Prospekt K. Marksa, Novosibirsk 630073, Russian Federation); Research engineer, Dasha.AI (32 Pisareva St., Novosibirsk 630005, Russian Federation), **ORCID: https://orcid.org/0000-0001-5963-6327,** bstodin@gmail.com

The author has read and approved the final manuscript.