УДК 123.456.789

# Использование векторных представлений текста для синтеза эмоциональной речи

В. С. Болдаков

Исследуется возможность синтеза эмоциональной речи при помощи глобальных векторов стиля. Предлагается новый метод синтеза эмоциональной речи, использующий векторные представления эмоций на основе текстов. Демонстрируется реализация метода на авторегрессионной архитектуре Tacotron 2 и на базе трансформера FastSpeech 2.

*Ключевые слова*: синтез речи, нейронные сети, трансформер, Tacotron 2, FastSpeech 2.

## 1. Введение

Многие люди предпочитают использовать голосовой интерфейс в своей повседневной жизни, например – управление бытовой техникой при помощи умной колонки или создания таймера через голосового ассистента смартфона. Такой интерфейс предполагает необходимость взаимодействия с пользователем посредством голоса. Чтобы не ограничивать функциональность программного обеспечения, используя предзаписанные человеком фразы, необходим синтез речи.

В последние годы появилось большое количество исследований нейросетевых методов для качественного синтеза речи. Например, авторегрессионная архитектура Tacotron 2 [1] или архитектура на базе трансформера FastSpeech 2 [2]. В указанных работах синтезируется естественная речь, но нет возможности контролировать эмоции, с которыми произносится текст.

Существуют решения, позволяющие при синтезе задавать конкретную эмоцию из ограниченного дискретного распределения. В данной работе описывается новый метод синтеза речи с широким спектром эмоций из непрерывного распределения векторных представлений, полученных из текста с помощью модели, описанной в [3], с их последующим линейным преобразованием и нормированием.

В данной работе получено решение на базе архитектур Tacotron 2 и FastSpeech 2, позволяющее синтезировать эмоциональную речь. Данный способ получения эмоционального синтеза может быть применен к большинству существующих нейросетевых архитектур синтеза речи.

## 2. Эмоциональный синтез на базе Tacotron 2

### 2.1. Использование глобальных токенов стиля для синтеза эмоциональной речи

В ходе работы исследовался эмоциональный синтез речи на базе Tacotron 2 с использованием глобальных токенов стиля [4]. Такой подход предполагает обучение модели синтеза с добавлением токенов стиля, выделенных из референсного аудио, содержащего необходимый стиль.

Однако глобальные токены стиля не позволяли синтезировать речь с нужными эмоциями, основываясь на аудио с меткой необходимой эмоции. Токены стиля отлично справлялись с переносом скорости речи с референсного аудио. Чтобы избежать сохранения информации о скорости референсного аудио в векторном представлении глобального токена стиля рассмотренно распределение скорости речи аудио. В качестве скорости речи принималось количество фонем

на секунду аудио. На рис. 1 можно увидеть распределение количества фонем в секунду в выборке данных для обучения.

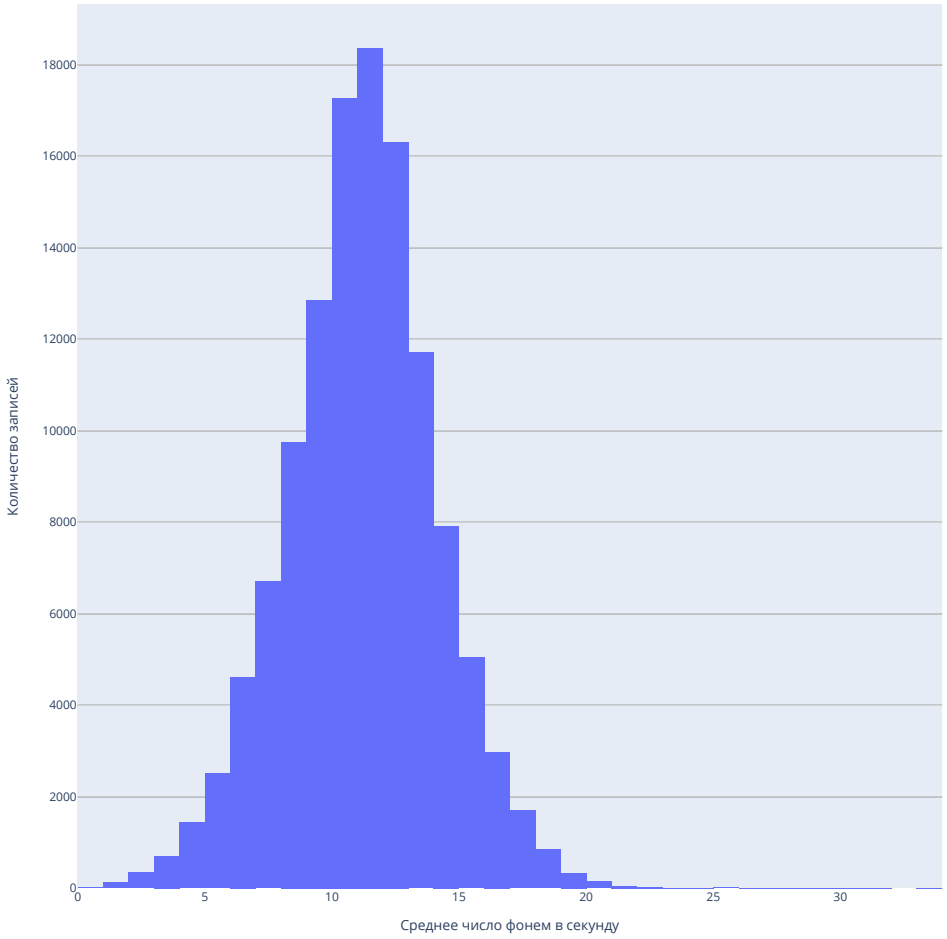


Рис. 1. Распределение количества фонем в секунду в выборке данных для обучения

Из распределения видно, что существуют аудиозаписи с существенно разной скоростью произношения, что должно сказываться на качестве обучения слоев выделения глобального токена стиля. Поэтому нормализованная характеристика скорости конкатенировалась к скрытому представлению фонемы, полученному из кодировщика Tacotron 2.

### 2.2. Обучение слоя предсказания продолжительностей фонем с учителем

Для обучения Tacotron 2 в оригинальном виде требуется много времени и ресурсов. Основная проблема заключается в том, что модель обучается строить продолжительности фонем без учителя. Можно ускорить процесс обучения и дистиллировать знания о продолжительностях фонем, полученные из модели Montreal Forced Aligner [5], обученной на тех же данных, что и в [6]. Это позволило значительно ускорить сходимость модели.

Выход слоя внимания в модели Tacotron 2 для каждого фрейма мелспектрограммы и для каждой фонемы предсказывает вероятность того, что какая-либо фонема или её часть содержится в каком-либо фрейме генерируемой мелспектрограммы. На рис. 2 и 3 можно увидеть предсказанные моделью вероятности содержания фреймом части фонемы с дистиллированными знаниями о продолжительностях и без. На рис. 4 показана истинная вероятность произно-

шения фонемы во фрейме.

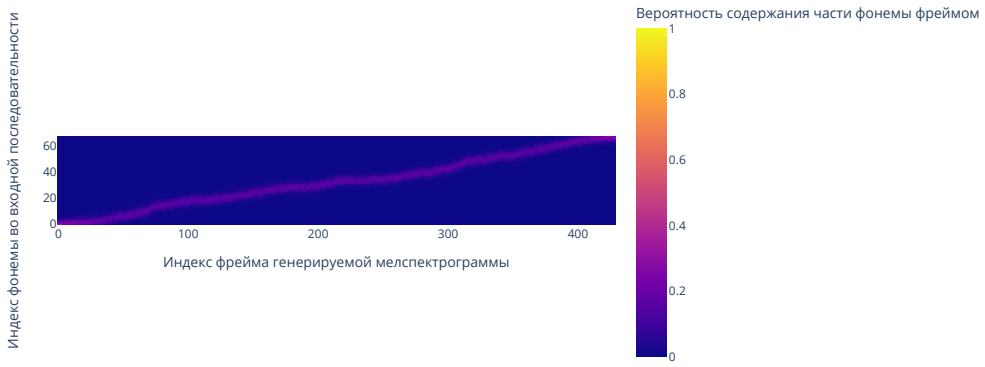


Рис. 2. Предсказанная моделью вероятность соответствия фрейма мелспектрограммы и фонемы без дистилляции знаний

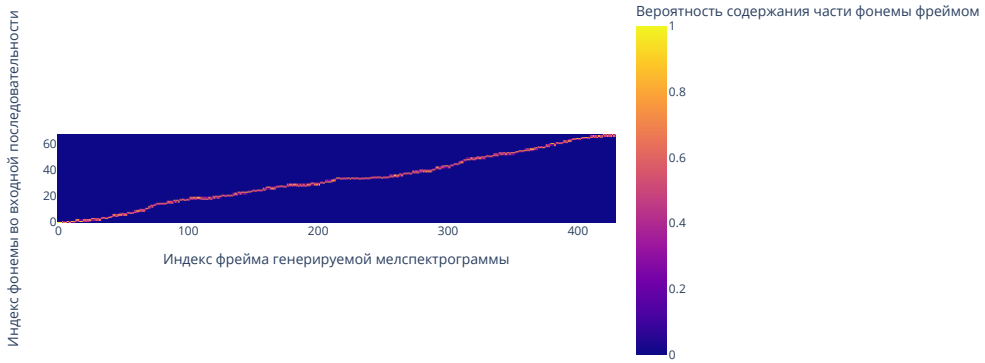


Рис. 3. Предсказанная моделью вероятность соответствия фрейма мелспектрограммы и фонемы с дистилляцией знаний

## 2.3. Архитектура модели эмоционального синтеза на базе глобальных токенов стиля

Итоговая архитектура модели эмоционального синтеза на базе глобальных токенов стиля показана на рис. 5.

Рассмотрим более подробно архитектуру кодировщика и декодировщика Tacotron 2. Они представляют собой последовательность сверток, линейных преобразований и рекуррентных слоев. Связь между кодировщиком и декодировщиком осуществляет механизм внимания, который и производит расчет вероятности произношения конкретной фонемы в момент каждого фрейма. На рис. 6 можно увидеть архитектуры кодировщика и декодировщика Tacotron 2.

Данная модификация не позволила синтезировать эмоциональную речь, т.к. глобальные токены стиля не выделяли информацию об эмоциях референсного аудио, но сохраняли информацию о длительности пауз или длительности отдельных фонем.

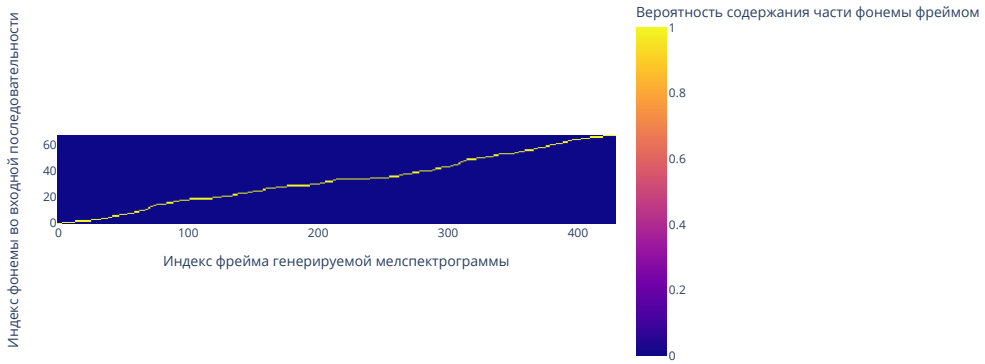


Рис. 4. Истинные соответствия фреймов мелспектрограммы и фонем


Рис. 5. Модификация архитектуры Tacotron 2, учитывающая скорость произношения, глобальный токен стиля и дистиллирующая знания о продолжительностях фонем

## 2.4. Использование векторных представлений эмоций для эмоционального синтеза речи

В результате обучения предыдущей модели был сделан вывод о том, что глобальный токен стиля не сохраняет информацию об эмоции и не является способом эффективного и предсказуемого изменения выходного аудио. Поэтому в данной работе предлагается оригинальный метод использования эмоциональных векторных представлений, извлеченных из предобученной модели, описанной в [3].

Чтобы понять, что наш набор данных содержит необходимое количество нужных эмоций, все полученные эмоции из модели [3] были разбиты на группы: Sad, Happy, Angry, Weary, Neutral. Таким образом, было проверено наличие в наборе данных записей, тексты которых модель считает содержащими эмоцию. На рис. 7 представлена частота распознанных эмоций: преобладают нейтральные тексты, что создает дисбаланс в выборке.

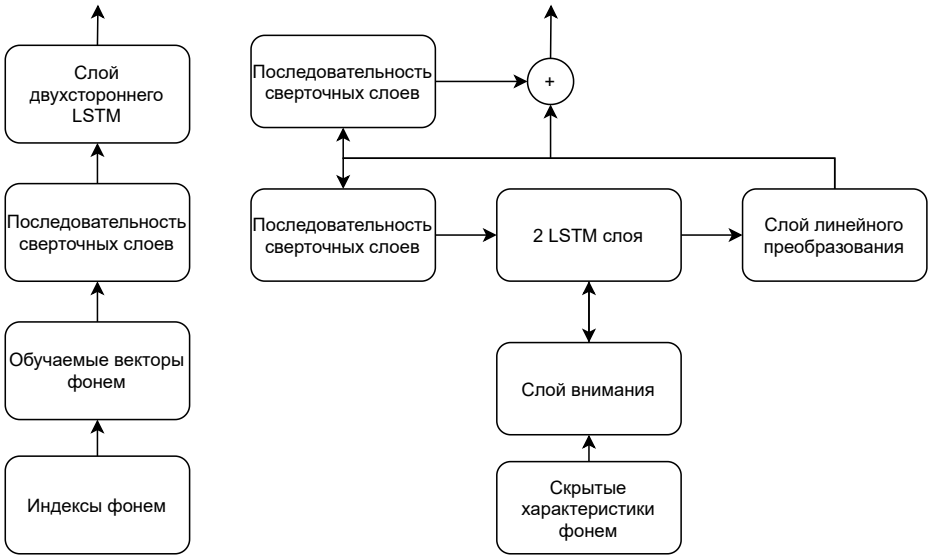


Рис. 6. Архитектуры кодировщика и декодировщика Tacotron 2 слева и справа соответственно

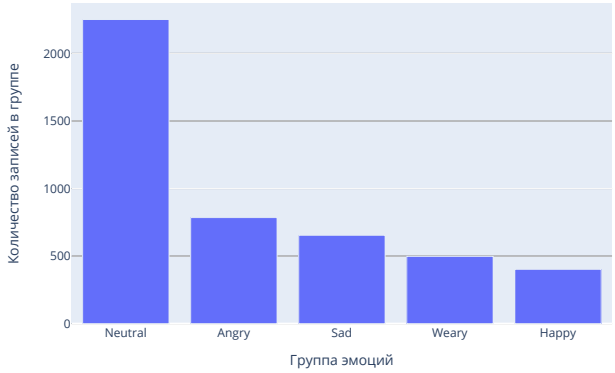


Рис. 7. Распределение предсказанных на основе текстов меток в случайной выборке дикторов из обучающего набора данных

Перед конкатенацией векторного представления эмоций к скрытым представлениям фонем происходит линейное преобразование конкатенированных векторов эмоций и диктора. Таким образом, мы достигаем, во-первых, уменьшения размерности большого вектора эмоций, а во-вторых, смешиваем его с векторным представлением диктора:

$$DecoderInput = Concat(EncoderOutput, W \times Concat(SpeakerEmb, EmotionEmb) + b),$$

где $W$, $b$ – обучаемые параметры линейного преобразования, $Concat$ – конкатенация векторов, $SpeakerEmb$ – обучаемый вектор, содержащий информацию о дикторе, $EmotionEmb$ – векторное представление эмоции, полученное из [3], $EncoderOutput$ – скрытые фонемные представления кодировщика Tacotron 2.

Данная модель способна синтезировать эмоциональную и связную речь [7]. Однако обучение такой модели требует большого количества эмоциональных данных, тогда как время обучения также сильно растет, поэтому качественно обучить Tacotron 2 с эмоциональными векторами на большом наборе данных не представляется возможным на имеющихся ресурсах.


Рис. 8. Модификация архитектуры Tacotron 2 для синтеза эмоциональной речи

## 3. Эмоциональный синтез на базе FastSpeech 2

Tacotron 2 требует слишком большого количества вычислительных ресурсов. Поэтому подобная же модификация применена к менее требовательному FastSpeech 2.

Для лучшего понимания контекста также рассмотрим архитектуры кодировщика и декодировщика FastSpeech 2. Кодировщик и декодировщик имеют общую архитектуру и отличаются количеством блоков: в кодировщике используется 4 блока, в декодировщике – 6. Также в кодировщике происходит преобразование индексов фонем в обучаемые векторы фонем, а в декодировщике происходит предсказание продолжительности каждой из фонем и их размножение до предсказанного значения фреймов. Подробнее это продемонстрировано на рис. 9.

Так же как и в Tacotron 2, у оригинального FastSpeech 2 нет никакой возможности получить информацию о скорости произношения – в данных, как уже было показано выше, есть большой разброс относительно этого параметра. Однако в данном случае из-за природы архитектуры трансформера предлагается прибавлять к выходу энкодера обучаемый вектор квантированного значения скорости. Таким образом, достигается необходимая размерность данных, и декодировщику передается информация о скорости аудио. Добавление же эмоционального вектора происходит через линейное преобразование вектора, снижающее размерность вектора эмоций до размера внутреннего представления фонем и последующего нормирования вектора, чтобы избежать взрыва градиентов и ускорить сходимость:

$$y = \frac{W \times EmotionEmb - \mathbf{E}[W \times EmotionEmb]}{\sqrt{\mathbf{D}[W \times EmotionEmb] + \epsilon}} \times \gamma + \beta,$$

где $EmotionEmb$ – полученное векторное представление эмоций, $W$ – обучаемая матрица линейного преобразования, $\beta, \gamma$ – обучаемые параметры.

Архитектура модифицированного FastSpeech 2 для синтеза эмоционального текста показана на рис. 10.

Модификация данной архитектуры лучше справляется с синтезом эмоциональной речи [8].

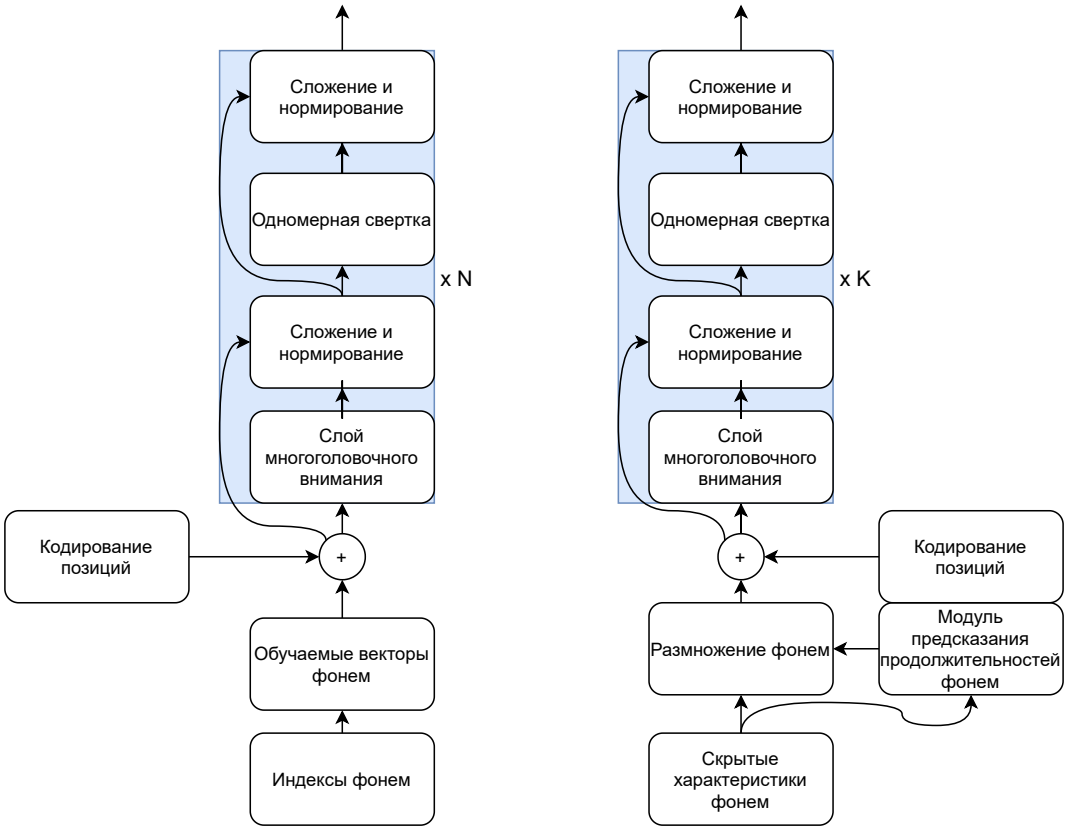


Рис. 9. Архитектура кодировщика и декодировщика FastSpeech 2 слева и справа соответственно


Рис. 10. Модификация архитектуры FastSpeech 2 для синтеза эмоциональной речи

## 4. Набор данных

Модели обучались на смеси из внутреннего приватного набора данных и внешних открытых данных. Из открытых данных использовался набор данных «The LJ Speech Dataset» [9]. Все данные были пересемплированы с частотой дискретизации 16 кГц.

Т а б л и ц а 1. Использованная для обучения выборка данных

| Набор данных | Количество дикторов | Продолжительность |
|---|---|---|
| The LJ Speech Dataset | 1 | 24 часа |
| Внутренний набор данных | 178 | 45.5 часов |
| Всего | 179 | 69.5 часов |

## 5. Результаты

В результате данной работы получен метод синтеза эмоциональной речи, который может быть применен к большинству существующих нейросетевых архитектур. Также для модели FastSpeech 2 проведено тестирование качества синтеза. Каждый из 70 носителей языка для каждого из 70 аудио поставил оценку от 0 до 5, характеризующую естественность речи.

Т а б л и ц а 2. Средние значения естественности синтезированной и оригинальной речи

| Оценка речи, синтезированной FastSpeech 2 с эмоциями | Оценка речи, произнесенной человеком |
|---|---|
| 4.159 | 4.279 |

В дальнейшем возможно исследование применения векторных представлений эмоций, полученных с помощью более современных и точных моделей, например, на основе BERT [10].

## Литература

1. *Shen J., Pang R., Weiss R. et al.* Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions // 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, April 15 – April 20, 2018. P. 4779–4783.
2. *Ren Y., Hu C., Tan X. et al.* FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. [Электронный ресурс]. URL: `https://arxiv.org/abs/2006.04558` (дата обращения: 09.09.2021).
3. *Felbo B., Mislove A., Sogaard A. et al.* Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm // Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, September, 2017. P. 1615–1625.
4. *Wang Y., Stanton D., Zhang Y. et al.* Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis. [Электронный ресурс]. URL: `https://arxiv.org/abs/1803.09017` (дата обращения: 19.09.2021).
5. *McAuliffe M., Socolof M., Mihuc S. et al.* Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi // INTERSPEECH 2017: Conference of the International Speech Communication Association, Stockholm, Sweden, August 20 – August 24, 2017. P. 498–502.
6. *Zhu X., Zhang Y., Yang S. et al.* Pre-Alignment Guided Attention for Improving Training Efficiency and Model Stability in End-to-End Speech Synthesis // IEEE Access. 2019. V. 7. P. 65955–65964.
7. *Болдаков В. С.* Примеры синтеза эмоциональной речи на базе Tacotron 2. [Электронный ресурс]. URL: `https://bit.ly/3nOPHRN` (дата обращения: 09.09.2021).
8. *Болдаков В. С.* Примеры синтеза эмоциональной речи на базе FastSpeech 2. [Электронный ресурс]. URL: `https://bit.ly/39i0T15` (дата обращения: 09.09.2021).
9. *Ito K., Johnson L.* The LJ Speech Dataset. [Электронный ресурс]. URL: `https://keithito.com/LJ-Speech-Dataset/` (дата обращения: 09.09.2021).

10. *Luo L., Wang Y. et al.* EmotionX-HSU: Adopting Pre-trained BERT for Emotion Classification. [Электронный ресурс]. URL: https://arxiv.org/pdf/1907.09669.pdf (дата обращения: 19.09.2021).

*Статья поступила в редакцию 26.09.2021;*
*переработанный вариант – 31.10.2021.*

**Болдаков Валерий Сергеевич**
аспирант кафедры прикладной математики и кибернетики СибГУТИ (630102, Новосибирск, ул. Кирова, 86), инженер-исследователь отдела машинного обучения Dasha AI (NY 10022, New York, 885 Third Avenue, 24th floor), e-mail: valboldakov@gmail.com.

**Emotional Speech Synthesis with Emotion Embeddings**

**V. Boldakov**

Several neural network architectures provide high-quality speech synthesis. Several neural network architectures provide high-quality speech synthesis. In this article, emotional speech synthesis with global style tokens is researched. A novel method of emotional speech synthesis with emotional text embeddings is described.

*Keywords*: speech synthesis, neural networks, transformer, Tacotron 2, FastSpeech 2.