**© 2022 г. Д.С. ОБУХОВ (bstodin@gmail.com)**
**(Новосибирский государственный технический университет)**

# КЛОНИРОВАНИЕ И КОНВЕРСИЯ ПРОИЗВОЛЬНОГО ГОЛОСА С ИСПОЛЬЗОВАНИЕМ ГЕНЕРАТИВНЫХ ПОТОКОВ

С целью повышения качества формируемого речевого сигнала в данной работе предложен способ учета переменной во времени информации о спикере. Благодаря этой технике система синтезирует более естественную речь голосом, похожим на заданный целевой голос, как в задаче клонирования голоса, так и в задаче конверсии голоса.

*Ключевые слова*: клонирование голоса, конверсия голоса, синтез речи, потоковые генеративные модели, эмбеддинги спикера, частота основного тона.

**DOI:** 10.31857/S0005231022100087, **EDN:** AKGYHA

## 1. Введение

В настоящее время сфера применения синтеза речи стремительно расширяется и уже нашла свое применение в области медицины [1, 2], в голосовых колонках, умных ассистентах и других окружающих человека умных устройствах [3, 4], а также в различных задачах бизнеса [5, 6]. Одним из актуальных направлений развития синтеза речи сегодня является синтез голосом произвольного человека [7]. Умение генерировать речь с заданным голосом является необходимым требованием для ряда задач, например, диалоговых систем.

Современные подходы на основе глубокого обучения позволили эффективно и качественно формировать естественную речь голосом одного заданного диктора, представленного в наборе данных обучения. Предложенные недавно техники позволяют учитывать несколько дикторов при обучении, однако множество голосов, которыми формируется речь, по-прежнему остается ограниченным. Построение систем клонирования и конверсии произвольного голоса становится следующим вызовом в области формирования речевых сигналов.

Задача клонирования голоса подразумевает использование заданного образца речи человека для синтеза таким же голосом речевого сигнала с произвольным содержанием, заданным текстом [8]. Важной отличительной чертой клонирования голоса от обычного синтеза речи является то, что обученная модель может синтезировать речь голосами даже тех спикеров, которые не были представлены в наборе данных обучения.

Задача конверсии голоса заключается в преобразовании аудиосигнала с голосом исходного спикера в аудиосигнал с тем же лингвистическим содержанием, т.е. произнесенным текстом, но с произношением голосом целевого

спикера [9]. В зависимости от того, с какими голосами система может работать, конверсия голоса подразделяется на: один к одному, несколько к одному, несколько к нескольким, много к нескольким, много ко многим. Наибольший интерес представляет конверсия много ко многим, поскольку при таком типе конверсии происходит преобразование аудиосигнала с произвольными исходным и целевым голосами.

В совокупности задачи клонирования и конверсии голоса обеспечивают полный набор возможностей по преобразованию голоса речи — как для случая, когда исходная речь имеет текстовое представление, так и для случая, когда исходная речь задана в виде аудиосигнала.

За счет техники, предложенной в [10], которая заключается в использовании открытых представлений, так называемых эмбеддингов спикера, содержащих информацию о скрытых характеристиках спикера, многоголосый синтез речи можно обобщить на клонирование голоса. Современные системы синтеза речи имеют нейросетевую архитектуру [11], как правило, на основе трансформеров [12–14] и генеративных потоков [15–17]. Модели на основе генеративных потоков ко всему прочему позволяют выполнять задачу конверсии голоса за счет применения обратимых преобразований. Для построения системы, способной выполнять и синтез речи, и конверсию голоса, в настоящей работе предлагается использовать нейросетевую архитектуру на основе генеративных потоков с использованием открытых эмбеддингов спикера.

В [16, 17] авторы также используют генеративные потоки и обращают внимание на возможность выполнения конверсии голоса. Модели с использованием генеративных потоков недавно показали впечатляющие результаты в области синтеза речи, позволяя формировать разнообразные произнесения заданного текста. Однако в этих работах авторы делают основной акцент на качественный синтез речи голосом одного заданного диктора. Кроме того, работы [16, 17] без дополнительных модификаций не предусматривают возможность клонирования голоса. В отличие от [16, 17] решение, предложенное в настоящей работе, позволяет выполнять и клонирование голоса, и конверсию голоса.

Одним из недостатков моделей на основе генеративных потоков [16, 17] является монотонность синтезированной речи. Ранее в [12] было показано, что учет основного тона позволяет добиться более совершенного произношения заданным голосом. Частота основного тона является характеристикой спикера, которая, будучи переменной во времени и зависящей от лингвистического содержания речи, дополняет эмбеддинги спикера. В архитектурах моделей синтеза речи, предложенных в [12, 14], используется информация о частоте основного тона. Однако [12, 14] это трансформерные архитектуры, которые лишены возможности конверсии голоса, поэтому предложенный в этих работах подход хоть и может быть реализован в решениях на основе генеративных потоков, но не позволяет учитывать питч сигнала при выполнении конверсии голоса.

Предложенный в данной работе подход на основе потоковых генеративных моделей позволяет выполнять задачу клонирования голоса за счет использования полученных из внешней системы вещественных векторов фиксиро-

ванной размерности, содержащих информацию о спикере, т.н. эмбеддингов спикера. За счет своих архитектурных возможностей генеративные потоки позволяют одновременно с этим решать задачу конверсии голоса, таким образом обеспечивая полный набор возможностей по преобразованию голоса речи — как для случая, когда исходная речь имеет текстовое представление, так и для случая, когда исходная речь задана в виде аудиосигнала. С целью улучшения конверсии голоса в настоящей работе предложен новый способ учета частоты основного тона.

Таким образом, вклад автора в данной работе следующий:

— объединены предложенные техники использования внешних эмбеддингов спикера и декодировщика на основе генеративных потоков для создания модели, способной одновременно выполнять задачи синтеза речи несколькими голосами, клонирования голоса и конверсии голоса;
— предложен новый способ учета информации о частоте основного тона для задач синтеза речи, клонирования и конверсии голоса для решений на основе генеративных потоков.

Структура работы следующая: во втором разделе приведена архитектура предложенной модели и описан процесс ее обучения, также описана математическая модель, которая лежит в основе генеративных потоков. В третьем разделе описан процесс выполнения клонирования голоса. В четвертом разделе описан процесс выполнения задачи конверсии голоса. В пятом разделе описаны возможные техники и предложен новый подход для учета информации о частоте основного тона. В шестом разделе приведены результаты экспериментов. Заключение дано в седьмом разделе.

## 2. Архитектура модели

Предложенная система для выполнения синтеза речи, а также клонирования и конверсии голоса схематично изображена на рис. 1. На этом рисунке синим цветом показан сценарий синтеза речи и клонирования голоса, оранже-

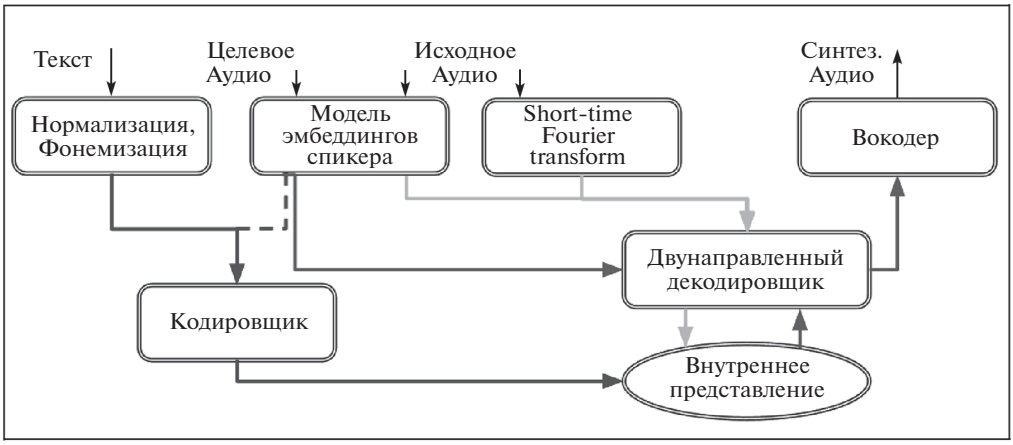


Рис. 1. Архитектура предложенной системы для выполнения синтеза речи, клонирования и конверсии голоса.

вым цветом показан сценарий конверсии голоса, красным цветом обозначены преобразования, относящиеся ко всем сценариям.

Технически, в предложенном подходе синтез речи и клонирование голоса выполняются одинаково, разница лишь в том, что при выполнении клонирования голоса для получения эмбеддинга спикера используется целевое аудио с голосом спикера, который не встречался в данных обучения модели. Таким образом, при выполнении сценария синтеза речи и клонирования голоса, текст нормализуется, т.е. приводится к упрощенному формату за счет раскрытия чисел, сокращений и пр., и фонемизируется, т.е. преобразуется в последовательность звуков, соответствующих произнесению этого текста. Затем последовательность фонем подается на вход в кодировщик, после чего полученное внутреннее представление декодируется с использованием эмбеддинга, полученного из целевого аудиосигнала. На последнем шаге полученная спектрограмма преобразуется в аудиосигнал. Подробнее сценарий клонирования голоса описан в разделе 3.

Сценарий конверсии голоса выполняется другим образом. Спектрограмма исходного аудиосигнала декодируется в обратном направлении с использованием эмбеддинга спикера, полученного из этого же сигнала, а затем полученное внутреннее представление декодируется в прямом направлении с использованием эмбеддинга, полученного из целевого сигнала. Полученная спектрограмма преобразуется в аудиосигнал. Подробнее сценарий конверсии голоса описан в разделе 4.

Два важных свойства обеспечивают выполнение описанного сценария. Во-первых, внутреннее представление не зависит от характеристик спикера, а зависит только от лингвистического содержания. Во-вторых, декодировщик является двунаправленным. Это означает, что в прямом направлении декодировщик преобразует внутреннее представление в спектрограмму, а в обратном направлении, наоборот, преобразует спектрограмму во внутреннее представление, причем эти преобразования происходят без потерь. Далее по ходу этого раздела будет показано, за счет чего данные свойства достигаются.

На приведенной схеме верхние четыре блока не являются обучаемыми. Это либо детерминированные преобразования, как в случае с нормализацией текста и преобразованием Фурье, либо преобразования, выполненные предобученными моделями. Для корректной работы предложенной системы требуется обучить центральную часть — акустическую модель, которая включает кодировщик и декодировщик, а также несколько дополнительных модулей.

Предложенная акустическая модель состоит из нескольких основных модулей: текстовый кодировщик, потоковый декодировщик и модуль предсказания продолжительностей произнесения фонем. На рис. 2 приведена схема взаимодействия этих компонент во время обучения системы. В описанной схеме обучения рассматривается случай, когда дополнительная информация о частоте основного тона не используется. Подходы учета этой информации описаны в разделе 5.

Текстовый кодировщик отображает последовательность токенов фонем $x = x_{1:Ttext}$ в скрытое векторное представление $h = h_{1:Ttext}$. После текстового кодировщика два линейных слоя используются для получения стати-

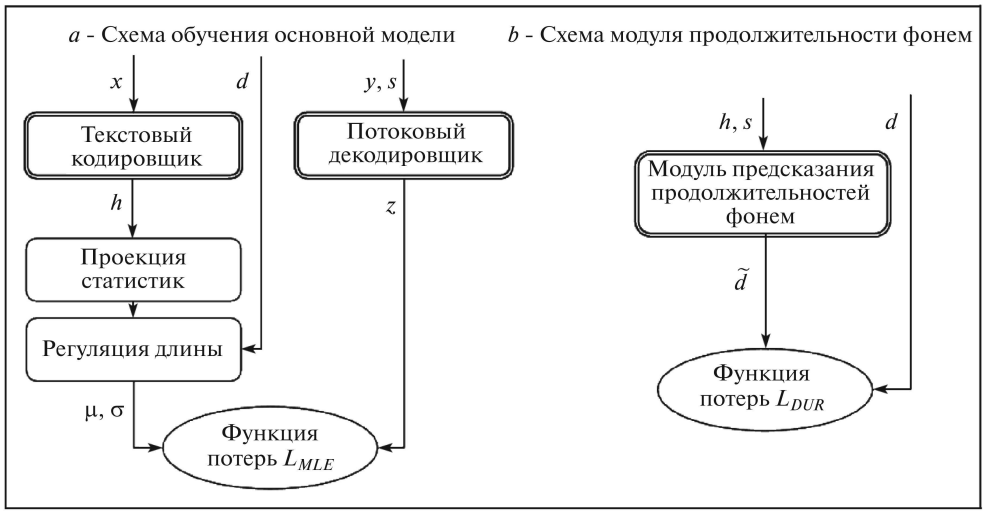


Рис. 2. Схема обучения предложенной модели.

стик $\mu = \mu_{1:Ttext}$ и $\sigma = \sigma_{1:Ttext}$ априорного распределения потокового декодера. В настоящей работе архитектура текстового кодировщика состоит из прямо направленных трансформер блоков. Заметим, что такая архитектура идентична предложенной в [16] за исключением того, что скрытая размерность и число фильтров в слоях были увеличены.

По аналогии с [16] в данной работе моделируется условное распределение спектрограмм $P_y(y \mid t, s)$ путем преобразования условного априорного распределения $P_z(z \mid t, s)$ через потоковый декодер $f_{dec} : z \to y$, где $y$, $t$ и $s$ обозначают входную спектрограмму, текстовую последовательность и информацию о спикере соответственно.

Потоковый декодировщик представляет из себя последовательность потоковых слоев, которые применяют обратимые преобразования. Такие обратимые преобразования гарантируют важное свойство потокового декодировщика — его двунаправленность. На рис. 2 направление работы декодировщика обозначено от спектрограммы $y$ к внутреннему представлению $z$. В разделах 3 и 4 при выполнении клонирования и конверсии голоса будет показано, как декодировщик работает в обратном направлении.

Генеративные потоки позволяют оценить правдоподобие данных и обучаются так, чтобы максимизировать это правдоподобие. Используя замену переменных, можно вычислить логарифм правдоподобия данных следующим образом:

$$\log P_y(y \mid c) = \log P_z(z \mid c) + \log \left| \det \frac{\partial f_{dec}^{-1}(x)}{\partial x} \right|. \tag{1}$$

Априорное распределение $P_z$ в (1) является изотропным многомерным распределением Гаусса, и процесс обучения выстраивается так, чтобы его статистики соответствовали статистическим данным априорного распределения, $\mu$ и $\sigma$, полученным из текстового кодировщика $f_{enc}$.

Таким образом, априорное распределение можно выразить следующим образом:

$$\log P_z(z \mid c; \theta, A) = \sum_{j=1}^{Tmel} \log N\left(z_j; \mu_{A(j)}, \sigma_{A(j)}\right), \tag{2}$$

где $T_{mel}$ обозначает продолжительность спектрограммы.

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

$$\max_{\theta, A} L(\theta, A) = \max_{\theta, A} \log P_y(y \mid c; \theta, A). \tag{3}$$

Для обучения предложенной модели по формулам (2) и (3) статистики априорного распределения потокового декодера требуется выровнять по фреймам спектрограммы, т.е. сопоставить индексы этих двух последовательностей. На схеме рис. 2 этот блок обозначен как регуляция длины. Индексы статистик соотносятся со спектрограммой за счет выравнивания $A$, полученного из внешней системы. $A(j) = i$, если $j$-я фонема произносится на $i$-м фрейме спектрограммы:

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

Значения $d$ можно интерпретировать как продолжительности фонем, поскольку до регуляции длины векторы статистик имеют такую же длину, как и последовательность токенов фонем, которая подается на вход в текстовый кодировщик.

По аналогии с системами FastPitch [12], FastSpeech [13], FastSpeech 2 [14] для того, чтобы предсказывать продолжительности фонем при выполнении клонирования голоса, обучается дополнительный модуль — предсказатель продолжительностей фонем, рис. 2,*б*. Для каждого токена входной последовательности данный модуль предсказывает число $\log d$ — логарифм количества фреймов, на протяжении которых будет длиться соответствующая фонема. Для получения продолжительности фреймов $d$ округляется до ближайшего целого. Обучение модуля предсказания продолжительностей фонем достигается за счет минимизации среднеквадратичной ошибки между продолжительностями, полученными из выравниваний внешней системы и предсказанными:

$$L_{dur} = MSE(d, \overline{d}). \tag{5}$$

Модуль предсказания продолжительности фонем аналогичен предложенному в [13].

Заметим, что хоть на рис. 2 есть другие входы, помимо токенов текстовой последовательности $x$ и спектрограммы $y$, а именно продолжительности

фонем $d$ и эмбеддинги спикера $s$, однако для их получения требуется только аудио и текст.

Для построения продолжительностей фонем по формуле (4) используются выравнивания, полученные также из внешней системы. В рамках данной работы обучена собственная модель для построения выравниваний на основе смеси гауссовских моделей, на базе Kaldi Speech Recognition Toolkit [18].

Построение эмбеддингов спикера осуществляется за счет сторонней модели ECAPA-TDNN [19], так как имеет минимальную ошибку на задаче верификации спикера. Поскольку для построения эмбеддингов спикера используется предобученная модель, постольку по тексту настоящей работы они называются внешними. Для построения такого эмбеддинга необходим только аудиосигнал.

## 3. Клонирование голоса

Задача клонирования голоса заключается в том, чтобы синтезировать речевой сигнал образцом голоса, который не присутствовал в тренировочных данных модели синтеза речи. Образец с речью 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, возможно получить эмбеддинг для голоса любого произвольного спикера.

Более подробно процедура клонирования голоса изображена на рис. 3.

Сначала для заданного речевого сигнала с голосом целевого спикера строится вектор с характеристиками этого спикера, эмбеддинг спикер $s$. Текст,

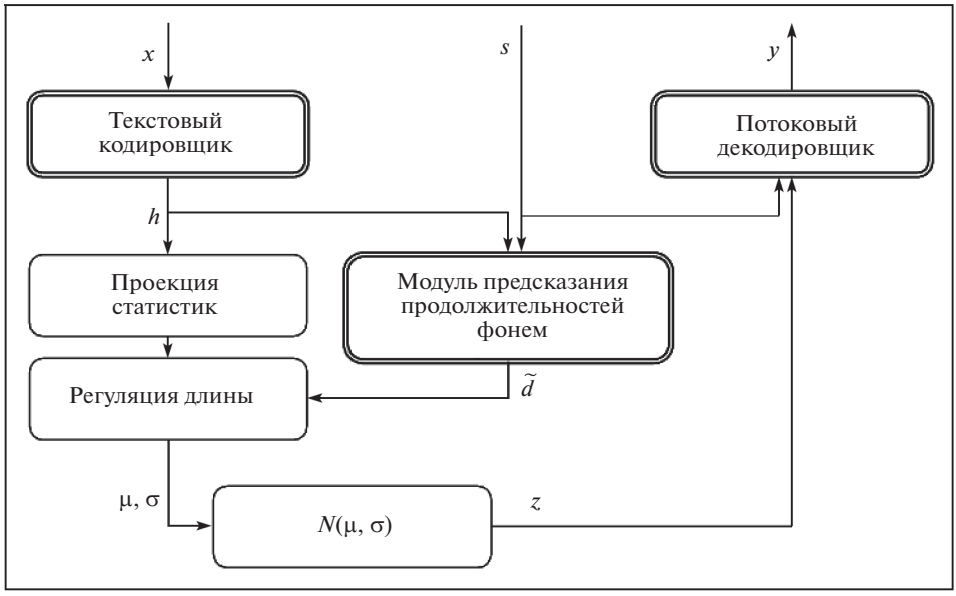


Рис. 3. Схема выполнения клонирования голоса.

который требуется озвучить, представляется в виде последовательности токенов $x$ и направляется в текстовый кодировщик, как и во время обучения. Выход текстового кодировщика $h$ используется в двух местах. Во-первых, вместе с эмбеддингом спикера $s$ для предсказания продолжительностей $\overline{d}$. Во-вторых, для построения статистик $\mu$ и $\sigma$, длина которых регулируется за счет продолжительностей $\overline{d}$. Направление действия потокового декодировщика во время работы клонирования голоса меняется на противоположное относительно обучения. Случайная величина из гауссовского распределения $N(\mu, \sigma)$ проходит через потоковый декодировщик, а эмбеддинг спикера учитывается в нем как дополнительное глобальное условие. Выходом декодировщика является спектрограмма $y$. Для того чтобы получить аудиосигнал из спектрограммы, в настоящей работе во всех экспериментах был использован вокодер HiFi-GAN [20].

## 4. Конверсия голоса

Задача конверсии голоса заключается в преобразовании аудиосигнала с голосом исходного спикера в аудиосигнал с тем же лингвистическим содержанием, но произношением голосом целевого спикера. При этом конверсия голоса позволяет копировать естественную интонацию и тембр голоса исходного диктора. Образцы исходного аудиосигнала и сигнала с речью целевого голоса обычно прилагаются в виде аудиофайла.

Схема выполнения конверсии голоса изображена на рис. 4.

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 аудиосигнала $x$, в котором содержится речь исходного спикера $s_{source}$:

$$z = f_{dec}^{-1}(y \,|\, s_{source})\,. \tag{6}$$

Это представление не зависит от спикера, поскольку при обучении требовалось, чтобы апостериорное распределение являлось изотропным многомерным распределением Гаусса со статистиками, полученными из текстового

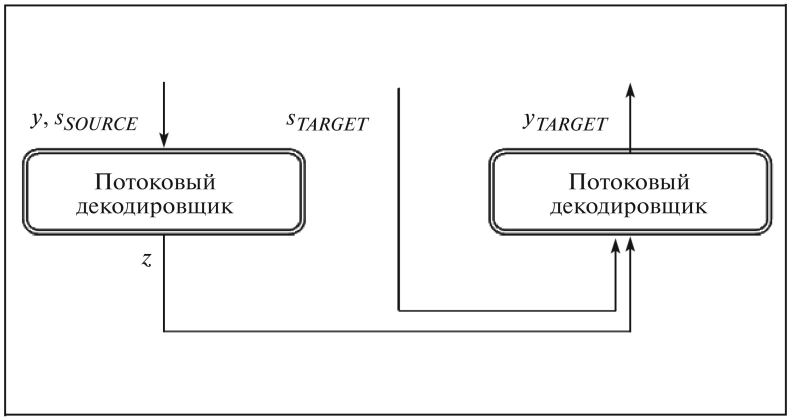


Рис. 4. Схема выполнения конверсии голоса.

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 аудиосигнал, с голосом целевого спикера и исходным лингвистическим содержанием:

$$y_{target} = f_{dec}(z \mid s_{target}). \tag{7}$$

Поскольку в процессе выполнения этих преобразований не происходит изменений продолжительностей фонем, постольку целевой голос сохранит темп речи, близкий к темпу речи исходного диктора.

Как уже упоминалось выше, модель ECAPA-TDNN позволяет получить эмбеддинги даже для спикеров, не представленных в данных обучения предложенной модели. За счет этого предложенный подход позволяет выполнять конверсию любого произвольного голоса в произвольный целевой голос, даже если образцы речи с этими голосами не встречались в данных обучения.

## 5. Учет частоты основного тона

В описанном выше подходе вся информация, специфичная для спикера, сосредоточена в одном эмбеддинге спикера. Однако это представление фиксированного размера и не зависящее от времени. Такое поведение не является достаточным для описания всей вариативности человеческой речи, поскольку в естественной речи присутствует и переменная во времени, специфичная каждому голосу информация.

Частота основного тона является характеристикой переменной во времени и специфичной для заданного спикера. Поэтому учет частоты основного тона должен дополнить информацию из эмбеддинга спикера и сделать синтезируемую речь более естественной.

В настоящей работе рассматриваются три следующие стратегии учета частоты основного тона:

— не учитывать;
— добавлять к выходам энкодера, по аналогии с подходами FastPitch [12] и FastSpeech2 [14];
— добавлять как локальное условие в декодер (предложенный подход).

Первая стратегия является базовым вариантом для сравнения.

Вторая стратегия используется в работах FastPitch [12] и FastSpeech2 [14]. Идея заключается в том, чтобы добавлять нормализованные значения частоты основного тона, либо полученные из них векторные представления, к выходам энкодера $h$. На этапе инференса при таком подходе требуется предсказывать значения частоты основного тона, для этого обучается дополнительный модуль предсказания частоты основного тона.

Недостатком такого подхода в предложенной архитектуре на основе генеративных потоков является то, что статистики $\mu$ и $\sigma$, полученные из представления $h$, становятся зависимыми от частоты основного тона, а значит и от переменных во времени характеристик спикера. Это не только нарушает базовую идею, заложенную в подход, так как представление $z$ перестает быть

независимым от спикера, но и ограничивает возможности эффективного выполнения конверсии голоса.

Поэтому в настоящей работе предложена третья стратегия учета частоты основного тона спикера. Вместо того, чтобы учитывать ее на выходе кодировщика, здесь предполагается учитывать частоту основного тона в декодере как дополнительное локальное условие [21]. На этапе инференса по-прежнему потребуется предсказывать значения частоты основного тона, и для этого, как и во втором подходе, необходимо обучить дополнительный модуль предсказания частоты основного тона. Однако принципиальное отличие в том, что статистики $\mu$ и $\sigma$ априорного распределения потокового декодера, как и внутреннее представление $z$, больше не зависят от характеристик спикера.

## 6. Эксперименты и результаты

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

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

На предварительном этапе до начала обучения тексты из набора данных были нормализованы и фонемизированы. Нормализация включала раскрытие сокращений, чисел, аббревиатур и специальных знаков. Фонемизация текста заключается в преобразовании заданного текста в последовательность фонем с учетом фонетических, морфологических и грамматических особенностей языка. В данной работе нормализация и фонемизация выполнялись с использованием инструмента Kyubyong/g2p [27]. Инструмент [27] также позволяет расставлять ударения в словах.

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

Для оценки предложенного подхода на задаче клонирования речи был проведен MOS (mean opinion score, усредненная оценка опрашиваемых) тест на естественность речи и похожесть голоса.

В рамках MOS теста на естественность речи асессору предлагалось прослушать аудиозапись и оценить их по шкале от 1 до 5, где 1 — это речь

**Таблица 1.** Используемые для обучения датасеты

| Набор данных | Количество записей | Количество часов обучения | Среднее количество часов на спикера |
|---|---|---|---|
| Blizzard 2013 [22] | 147 249 | 198,2 | 4,4 |
| HiFi-TTS dataset [23] | 323 978 | 291,7 | 29,2 |
| LibriTTS [24] | 375 086 | 585,8 | 0,26 |
| LJSpeech [25] | 13 100 | 23,9 | 23,9 |
| M AI Labs [26] | 69 853 | 143,6 | 35,9 |

**Таблица 2.** Сравнение предложенных систем на задаче клонирования голоса

| | Естественность речи | Похожесть голоса |
|---|---|---|
| Оригинальная речь | $3{,}969 \pm 0{,}034$ | $4{,}037 \pm 0{,}043$ |
| Без учета частоты основного тона | $3{,}711 \pm 0{,}045$ | $3{,}101 \pm 0{,}053$ |
| Учет частоты основного тона сразу после кодировщика | $3{,}745 \pm 0{,}043$ | $3{,}237 \pm 0{,}058$ |
| Учет частоты основного тона как локальное условие в декодировщике | $3{,}795 \pm 0{,}04$ | $3{,}306 \pm 0{,}062$ |

**Таблица 3.** Анализ MOS теста по оценке качества клонирования голоса предложенного решения по категориям

| | Естественность речи | Похожесть голоса |
|---|---|---|
| Детские голоса | $3{,}94 \pm 0{,}081$ | $3{,}24 \pm 0{,}144$ |
| Женские голоса | $3{,}883 \pm 0{,}082$ | $3{,}433 \pm 0{,}121$ |
| Мужские голоса | $3{,}64 \pm 0{,}099$ | $3{,}167 \pm 0{,}152$ |

совершенно неестественная, 5 — речь не отличима от человеческой. Каждую из записей оценивали по 20 раз. Всего в оценке принимало участие 75 записей для каждой из моделей, по 3 записи для каждого из 5 мужских, 5 женских и 5 детских голосов.

В рамках MOS теста на похожесть голоса асессору требовалось оценить, насколько голос в двух предложенных записях похож. Одна из предложенных записей являлась целевой записью с речью человека. Оценивание также происходило по шкале от 1 до 5, и каждая из тех же 75 записей сравнивалась с записями из оригинальной речи по 20 раз.

В табл. 2 приведено сравнение предложенных систем на задаче клонирования голоса.

В первой строке табл. 2 приведена оценка для оригинальной человеческой речи. В последующих строках приведены оценки для синтезированных записей в зависимости от способа учета в обученной модели частоты основного тона. Во второй строке оценивались записи, полученные из модели, которая не учитывает частоту основного тона никаким образом. В третьей строке приведена оценка для модели, в которой частота основного тона учитывается в представлениях, полученных из текстового кодировщика. В четвертой строке приведена оценка для модели, в которой учет частоты основного тона происходит в потоковом декодировщике.

Анализ результатов этого MOS теста показал, что система работает хуже для голосов, которые в меньшем объеме были представлены в данных обучения акустической модели, табл. 3. Так, для мужских и детских голосов результаты похожести голоса ниже, чем для женских голосов, которые присутствовали в данных обучения в большей степени. Интересно, что для детских голосов естественность речи при этом высока, но это достигается за счет того, что эти голоса больше звучат как женские, чем детские.

**Таблица 4.** Результаты MOS теста по оценке качества многоголосого синтеза речи

| | Естественность речи | Похожесть голоса |
|---|---|---|
| Оригинальная речь | $4{,}163 \pm 0{,}067$ | $3{,}872 \pm 0{,}05$ |
| Предложенное решение | $3{,}859 \pm 0{,}076$ | $3{,}751 \pm 0{,}053$ |
| Модель FastPitch [12] | $3{,}556 \pm 0{,}084$ | $3{,}785 \pm 0{,}058$ |
| Модель FastSpeech 2 [14] | $3{,}965 \pm 0{,}065$ | $3{,}701 \pm 0{,}067$ |
| Модель Glow-TTS [16] | $3{,}639 \pm 0{,}056$ | $3{,}639 \pm 0{,}056$ |

Кроме того, было проведено сравнение с другими решениями на задаче синтеза речи. В табл. 4 приведены результаты MOS теста на задаче синтеза речи. Помимо предложенного решения, в сравнении рассматривались упомянутые модели из [12, 14, 16]. Поскольку результаты при учете частоты основного тона в декодировщике оказались лучше, постольку далее в сравнении с другими системами рассматривался именно этот подход.

Несмотря на то что на задаче синтеза речи предложенное решение не является лучшим по всем критериям, работы [12, 14, 16] не позволяют выполнять клонирование голоса и работы [12, 14], не позволяют выполнять конверсию голоса.

## 7. Заключение

В данной работе была предложена архитектура, позволяющая выполнять задачи синтеза речи, клонирования и конверсии голоса. За счет использования внешних эмбеддингов спикера, предложенная архитектура, единожды обучившись, позволяет выполнять данные задачи даже с голосами спикеров, которые не встречались при обучении. Также предложена техника учета частоты основного тона, за счет которой удалось повысить естественность синтезированной речи и синтезировать речь, более похожую на заданный голос. Несмотря на это, результаты показывают, что речь человека звучит более естественно, а степень клонирования голоса остается недостаточно высокой.

Подход, предложенный в данной работе, является вычислительно эффективным, поскольку использует не авторегрессионный метод генерации последовательности, за счет чего асимптотика генерации аудиосигнала относительно входной последовательности является линейной. За счет того, что предложенная система требует однократного обучения, она является простой в использовании и внедрении в другие продукты.

В будущем можно улучшить качество клонирования и конверсии голоса за счет использования дополнительной информации из аудио, например, энергии сигнала, а также за счет увеличения объема данных обучения, в том числе и за счет данных из разных языков.

## СПИСОК ЛИТЕРАТУРЫ

1. *Cooper F.S., Gaitenby J.H., Nye P.W.* Evolution of reading machines for the blind: Haskins Laboratories' research as a case history // J. of Rehabilit. Res. Development. 1984. No. 21.1. P. 51–87.

2. *Miyabe M., Yoshino T.* Development of multilingual medical reception support system with text-to-speech function to combine utterance data with voice synthesis / ICIC '10: Proceedings of the 3rd international conference on Intercultural collaboration. 2010. P. 195–198.
3. *Kargathara A., Vaidya K., Kumbharana C.K.* Analyzing Desktop and Mobile Application for Text to Speech Conversation / Rising Threats in Expert Applications and Solutions. 2020. P. 331–337.
4. *Sokol K., Flach P.* Glass-Box: Explaining AI Decisions With Counterfactual Statements Through Conversation With a Voice-enabled Virtual Assistant / Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. 2018. P. 5868–5870.
5. *Hoy M.B.* Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants // Medical Reference Services Quarterly. 2018. No. 37. P. 81–88.
6. *Nasirian F., Ahmadian M., Lee O.* AI-Based Voice Assistant Systems: Evaluating from the Interaction and Trust Perspectives / Twenty-third Americas Conference on Information Systems. 2017.
7. *Obukhov D.S.* Многоголосый синтез естественной речи с использованием генеративных потоков // Современные информационные технологии и ИТ-образование. 2021. No. 17.4.
8. *Xie Q., Tian X., Liu G., et. al.* The Multi-Speaker Multi-Style Voice Cloning Challenge 2021 // International Conference on Acoustics, Speech, and Signal Processing. 2021.
9. *Sisman B., Yamagishi J., King S., Li H.* An overview of voice conversion and its challenges: From statistical modeling to deep learning // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2020.
10. *Jia Y., Zhang Y., Weiss R.J., et. al.* Transfer learning from speaker verification to multispeaker text-to-speech synthesis / Conference on Neural Information Processing Systems. 2018.
11. *Tan X., Qin T., Soong F., Liu T.-Y.* A survey on neural speech synthesis / arXiv preprint arXiv:2106.15561. 2021. URL: https://arxiv.org/pdf/2106.15561.pdf (дата обращения: 22.01.2022).
12. *Lancucki A.* Fastpitch: Parallel text-to-speech with pitch prediction / arXiv preprint arXiv:2006.06873. 2020. URL: https://arxiv.org/pdf/2006.06873.pdf (дата обращения: 22.01.2022).
13. *Ren Y., Ruan Y., Tan X., Qin T., Zhao S., Zhao Z., Liu T.-Y.* Fastspeech: Fast, robust and controllable text to speech / In Advances in Neural Information Processing Systems. 2019. P. 3165–3174.
14. *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. URL: https://arxiv.org/pdf/2006.04558.pdf (дата обращения: 22.01.2022).
15. *Valle R., Shih K., Prenger R., Catanzaro B.* Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis / arXiv preprint arXiv:2005.05957. 2020. URL: https://arxiv.org/pdf/2005.05957.pdf (дата обращения: 22.01.2022).
16. *Kim J., Kim S., Kong J., Yoon S.* Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search / In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems. 2020.
17. *Kim J., Kong J., Son J.* Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech / arXiv preprint arXiv:2106.06103. 2021. URL: https://arxiv.org/pdf/2112.02418.pdf (дата обращения: 22.01.2022).

18. *Povey D., Ghoshal A., Boulianne G., et. al.* The Kaldi Speech Recognition Toolkit / In IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. 2011.
19. *Desplanques B., Thienpondt J., Demuynck K.* Ecapatdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification, arXiv preprint arXiv:2005.07143. 2020. URL: https://arxiv.org/pdf/2005.07143.pdf (дата обращения: 22.01.2022).
20. *Kong J., Kim J., Bae J.* Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis / Advances in Neural Information Processing Systems. 2020.
21. *Oord A., Dieleman S., Zen H., et. al.* Wavenet: A generative model for raw audio / arXiv:1609.03499. 2016. URL: https://arxiv.org/pdf/1609.03499.pdf (дата обращения: 22.01.2022).
22. *King S., Karaiskos V.* The blizzard challenge 2013 / Proc. Blizzard Challenge workshop 2013. 2013.
23. *Bakhturina E., Lavrukhin V., Ginsburg B., Zhang Y.* Hi-fi multi-speaker english tts dataset / arXiv preprint arXiv:2104.01497. 2021.
URL: https://arxiv.org/pdf/2104.01497.pdf (дата обращения: 22.01.2022).
24. *Zen H., Dang V., Clark R. et. al.* LibriTTS: A corpus derived from LibriSpeech for text-to-speech / arXiv preprint arXiv:1904.02882. 2019.
URL: https://arxiv.org/abs/1904.02882 (дата обращения: 22.01.2022).
25. *Ito K., Johnson L.*, The LJ speech dataset / Электронный ресурс: The LJ Speech Dataset. 2017. URL: https://keithito.com/LJ-Speech-Dataset/ (дата обращения: 22.01.2022).
26. *Solak I.* The M-AILABS Speech Dataset / Электронный ресурс: The M-AILABS Speech Dataset. 2019. URL: https://www.caito.de/2019/01/the-m-ailabs- speech-dataset (дата обращения: 22.01.2022).
27. *Kyubyong P., Jongseok K.* g2pE: A Simple Python Module for English Grapheme To Phoneme Conversion / Электронный ресурс: GitHub repository. 2018. URL: https://github.com/Kyubyong/g2p (дата обращения: 22.04.2022).

*Статья представлена к публикации членом редколлегии А.А. Лазаревым.*

Поступила в редакцию 22.01.2022

После доработки 25.04.2022

Принята к публикации 29.06.2022