УДК 004.896

# ПОСТРОЕНИЕ И ОЦЕНКА ЭФФЕКТИВНОСТИ НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ КЛОНИРОВАНИЯ ГОЛОСА

*Кукареко А.П., студент гр. 953502*

*Белорусский государственный университет информатики и радиоэлектроники*
*г. Минск, Республика Беларусь*

*Калугина М.А. – канд. физ.-мат. наук, доцент*

**Аннотация**: В работе рассматривается управляемый метод клонирования голоса, позволяющий контролировать, качественно и количественно оценивать точность различных параметров синтезированной речи. Демонстрируется возможность использования генеративной модели для клонирования таких стилистических характеристик голоса, как высота тона, темп и тембр речи, просодия, фонетические особенности русской речи. Производительность метода тестируется слоями глубокой свертки для моделирования кодеров, декодеров и вокодера на базе WaveNet. Эффективность построенной в результате модели сравнима с современными системами преобразования текста в речь (TTS) и конверсии голоса (VC) при использовании образцов речи без текстового сопровождения длиной 1–5 минут.

**Ключевые слова:** клонирование голоса, генеративная нейросеть, текст в речь, mel-спектрограмма, вокодер, перенос стиля, преобразование голоса, адаптация говорящего, нулевой выстрел.

Клонирование голоса – это задача научиться синтезировать голос произвольного человека из образцов записанной речи. Недавние исследования в этой области были сосредоточены на синтезе голоса человека только на основе нескольких эталонных образцов. Недостатком такого подхода является отсутствие возможности контролировать различные аспекты стиля речи.

Адаптация моделей преобразования текста в речь (TTS) с несколькими говорящими требует обучения на большом наборе данных, содержащем несколько минут текста речи тысяч говорящих. Большое разнообразие говорящих в обучающих данных важно для того, чтобы легко клонировать голос произвольных говорящих. Чтобы решить проблему разделения стиля и характеристик говорящего в большом наборе данных с несколькими говорящими, содержащем в основном нейтральную по стилю речь, предлагается модель клонирования голоса, которая обучается как на скрытой, так и на эвристической информации о стиле. С помощью количественной и качественной оценок демонстрируется то, что предлагаемая модель может заставить новый голос выражать эмоции, петь или копировать стиль заданной эталонной речи.

Рассматривается этап достижения следующей цели работы: создать TTS-модели клонирования выразительного голоса, обучающиеся на голосах и стилевых аспектах нескольких говорящих. Обработка стиля в моделях TTS осуществляется путем изучения словаря векторов скрытого стиля – Global Style Tokens (GST) [1]. Эмпирически было обнаружено, что GST позволяет незначительно контролировать стиль при обучении на большом наборе данных с несколькими говорящими и в основном нейтральной просодией.

Модуль GST обрабатывает целевые фрагменты записи конкретной фразы, на выходе получая стилевые характеристики. Для этого используется словарь «обучаемых» векторов. Во время вывода синтезатор голоса может быть настроен на разные эталонные аудиозаписи для создания стилевых вариантов речи для одного и того же текста. Манипулирование переменными скрытого стиля во время вывода позволяет управлять стилем синтезированной речи. Модель Меллотрон [2] использует комбинацию явных и скрытых переменных стиля, для получения более точного контроля над выразительными характеристиками синтезированной речи. В частности, Меллотрон использует сети синтеза спектрограмм по кривой основного тона, GST и идентификатор говорящего во время обучения. Во время логического вывода синтезатор может обучаться на «мелодической» информации – высоте и ритме эталонной речи. Демонстрируется, что использование явных характеристик кривой высоты тона во время тренировки позволяет обобщить вывод на различные гармонические и мелодические кривые высоты тона.

Используемый эвристический YIN-алгоритм обработки сигналов извлекает основные характеристики высоты тона [3]. Было обнаружено, что использование сочетания скрытой и эвристической информации о стиле в модели TTS не только обеспечивает детальный контроль над аспектами стиля синтезированной речи, но также позволяет масштабировать нейросеть до большого набора данных с несколькими говорящими для получения более естественного звука для произвольного говорящего. Общий обзор модели клонирования, обрабатывающей стилистические характеристики говорящего, получаемые из записи голоса данного текста, показан на рисунке 1. Во время обучения модели могут быть предоставлены эталонные параметры стиля говорящего для достижения более выразительного клонирования голоса. Все три основных компонента – кодер говорящего, синтезатор mel-спектрограммы и вокодер – обучаются отдельно.

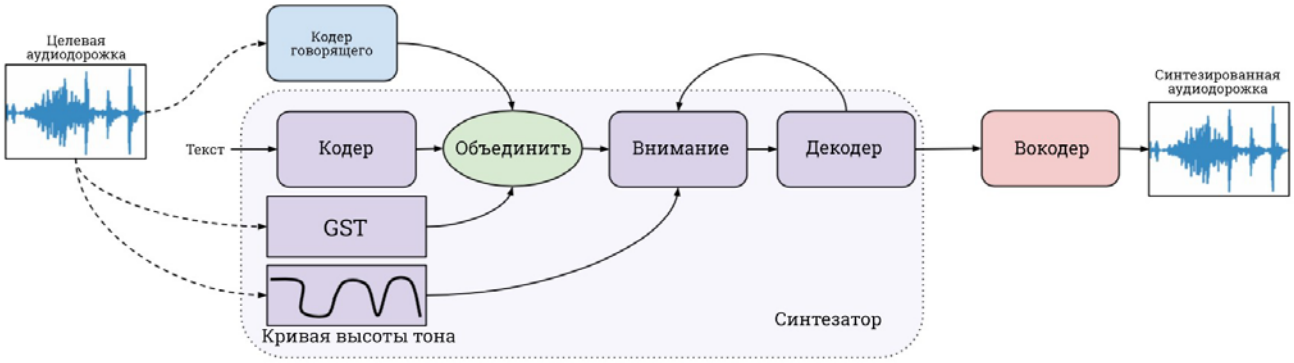


Рисунок 1 – Модель клонирования голоса Tacotron-2 TTS

Нейросеть представляет собой стек из 3 слоев LSTM по 256 ячеек в каждом слое, которые работают с мел-спектрограммами с 40 каналами. Итоговый результат достигается путем проецирования вывода LSTM на последнем уровне на 256 измерений с последующей нормализацией $L_2$. Кодер обучен оптимизировать общие потери при сравнении говорящих, т. е. достигать сходства между голосами от одного и того же говорящего. Во время обучения каждое высказывание разбивается на более мелкие сегменты по 1600 мс с перекрытием в 1000 мс между последовательными сегментами. Mel-спектрограммы состоят из текста ($t$), закодированного голоса ($s$), кривой высоты тона ($f_0$) и параметра скрытого стиля, полученного из GST($z$). Синтезатор представляет собой генеративную модель $g(t, s, f_0, z; W)$, параметризованную обучаемыми весами $W$, обученную для оптимизации функции потерь $L$, которая учитывает различия между сгенерированной и реальной достоверной спектрограммой:

$$\mathbb{E}_{(t_i, a_i)\sim D}\{L(g(t_i, s_i, f_{0i}, z_i; W), y_i)\},$$

где $D$ – массив данных, содержащий пары текста и аудио ($t_i$, $a_i$).

В работе применяется два подхода для клонирования голоса нового говорящего из нескольких речевых образцов с текстом и без него.

Обучение с нулевого выстрела: $L$-2 нормализация закодированного голоса для отдельных образцов целевого говорящего. Поскольку закодированный голос получается непосредственно из аудиодорожки, нам не требуется текст для клонирования голоса произвольного говорящего.

Адаптация модели: когда доступны расшифрованные образцы голоса, можно точно настроить модель синтеза, используя пары текста и звука. Как показано в [1], тонкая настройка может значительно улучшить показатели сходства говорящих в клонированной речи. Изучаются следующие два метода адаптации модели: полная адаптация – точная настройка всех параметров модели синтеза на текстовой и звуковой парах произвольного говорящего; адаптационный декодер – тонкая настройка только параметров декодера модели синтеза. Преимущество адаптации только параметров декодера в том, что для этого требуется меньше параметров модели, зависящих от говорящего, и кодер может использоваться для всех говорящих. В обеих приведенных выше настройках адаптации модель настраивается на 100–200 итераций с помощью оптимизатора Adam со скоростью обучения 1e-4.

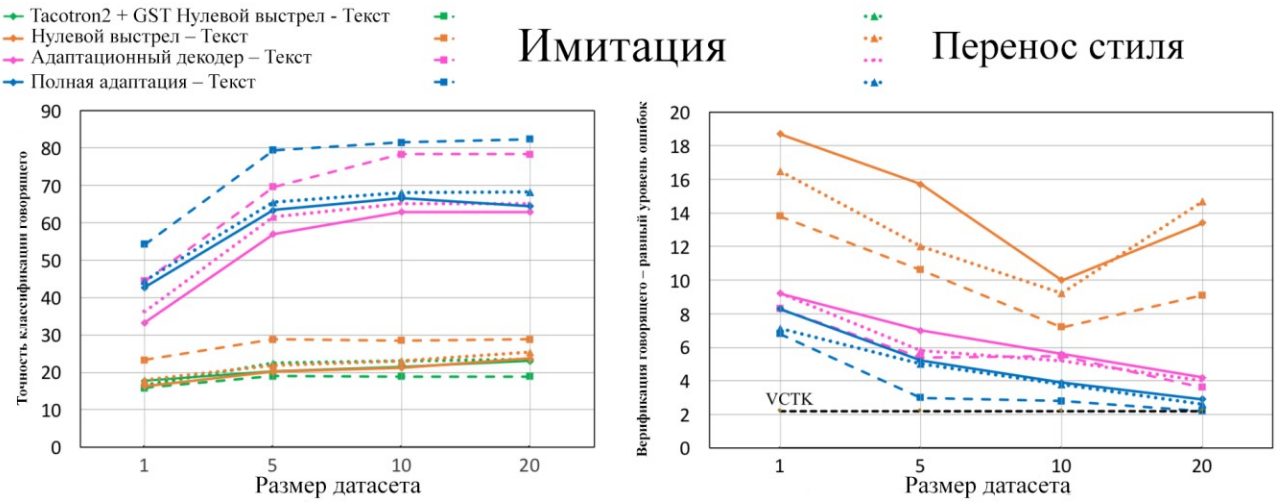


Рисунок 2 – Оценка эффективности моделей клонирования голоса в отношении точности классификации говорящего

Для достижения поставленной цели решались следующие задачи клонирования:

1. Текст – клонирование текста речи непосредственно из текста. Сначала синтезируется речь для данного текста на основе модели TTS с одним говорящим: Tacotron 2 + WaveGlow, обученную на наборе данных Russian Single. Затем выводится кривая высоты тона синтезированной речи с помощью YIN-алгоритма.

2. Имитация – восстановление образца целевого говорящего. Используется пара текста и звука целевого говорящего, нейросеть пытается восстановить звук из его обработанного представления. Все формирующие стиль переменные: высота тона, ритм, просодия – получены из образца речи, которую мы пытаемся имитировать. Эта задача позволяет количественно оценить метрики стилевого сходства.

3. Передача стиля – передача высоты тона и ритма речи от другого говорящего. Для этой задачи используются примеры из набора данных M-AILABS и Voxforge+audiobook с одним говорящим. Этот набор данных содержит выразительные аудиокниги, прочитанные одним говорящим с большим разбросом эмоций и высоты тона. Данный эталонный стиль аудио используется для извлечения высоты тона и ритма.

Для описанных выше задач клонирования оцениваются три аспекта клонированной речи: сходство говорящего с целевым говорящим, сходство стиля с эталонным стилем и естественность речи.

Для получения точности классификации говорящих классификатор говорящего обучается на наборе данных VCTK, чтобы классифицировать данное высказывание как одного из 78 говорящих. Классификатор – двухуровневая нейронная сеть с 256 скрытыми блоками, которая принимает на вход кодированный голос, полученный с помощью нашей предварительно обученной сети кодера. Подобно [4], классификатор достигает 100% точности на наборе, содержащем 200 примеров из набора данных VCTK. Клонируется 25 образцов речи на говорящего для каждой задачи. На рисунке 2 (слева) показаны кривые точности классификации говорящих для всех задач и методов клонирования в зависимости от количества размера выборки. Были получены следующие выводы: адаптация модели значительно превосходит методику клонирования голоса с нулевым выстрелом, поскольку она позволяет модели подстраиваться под характеристики произвольного говорящего.

Для клонирования голоса с нулевым выстрелом с использованием Tacotron2-GST, предлагаемая модель достигает высокой точности классификации говорящих для задач клонирования текста и переноса стиля. Точность предложенной модели немного выше для задачи имитации по сравнению с другими задачами как для адаптации модели, так и для клонирования голоса (нулевой выстрел). Это означает, что согласование кривой высоты тона целевого говорящего улучшает специфические для говорящего характеристики клонированной речи.

Для оценки сходства говорящих используется косинусоидное сходство аудиодорожек, полученных с помощью кодера говорящего. Равная частота ошибок (EER) – это точка, в которой частота ложноположительных и ложноотрицательных значений системы проверки говорящего равны. Оценка производится на случайно выбранных 20 говорящих в наборе данных VCTK. Используются 5 речевых образцов для каждого говорящего в системе проверки говорящего и синтезируются 50 речевых образцов для каждого говорящего для каждой задачи клонирования. EER оцениваются путем сравнения выборки одного и того же говорящего с выборкой другого говорящего. В таблице 1 показаны значения EER для различных методов и задач клонирования с использованием предложенной модели, а также те, которые были оценены с использованием реальных данных. Наблюдения за метрикой EER (ОВТ – ошибка высоты тона, ОГ – ошибка голосоведения, ОФ – ошибка фрейминга) аналогичны наблюдениям за метрикой точности.

Таблица 1 – Стилистическое сходство в задачах имитации и переноса стиля голоса

| Модель | ОВТ | ОГ | ОФ | Перенос стиля |
|---|---|---|---|---|
| Tacotron2 + GST – Нулевой выстрел | 20,35% | 26,38% | 29,47% | 2,69 ± 0,11 |
| Нулевой выстрел | 3,71% | 10,64% | 11,77% | 3,15 ± 0,11 |
| Адаптационный декодер | 2,39% | 11,61% | 12,49% | 3,29 ± 0,10 |
| Полная адаптация | 2,97% | 12,59% | 13,59% | 3,41 ± 0,10 |

Предлагаемые модели значительно превосходят базовые показатели Tacotron 2 + GST, что ясно указывает на важность согласования кривой высоты тона для точной передачи стиля, и могут быть использованы для создания персонализированных голосовых ассистентов, озвучки мультфильмов, рекламы, помощи людям, потерявшим голос.

***Список использованных источников***

*1. Y. Wang, E. Battenberg, Rif A. Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis – arXiv, 2018.*

*2. Rafael Valle, Jason Li, Ryan Prenger, and Bryan Catanzaro. Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens – ICASSP, 2020.*

*3. Alain De Cheveigné, Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music, 2002.*

*4. Sercan Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. Neural voice cloning with a few samples – NeurIPS, 2018.*

*5. Лобанов Б. М., Цирульник Л. И. Компьютерный синтез и клонирование речи // Минск : Белорусская Наука, 2008.*

UDC 004.896

# NEURAL NETWORK MODELS FOR VOICE CLONING: BUILDING AND EFFICIENCY ESTIMATION

*Kukareko A.P.*

*Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus*

*Kalugina M.A. – PhD in Physics and Mathematics*

We present a controlled method of voice cloning, which makes it possible to control, evaluate the accuracy of various parameters of synthesized speech in quantity and quality. The possibility of using a generative model for cloning such stylistic characteristics of a voice as pitch, tempo and timbre of speech, prosody, phonetic features of Russian speech is demonstrated. The performance of the method is tested by deep convolution layers to simulate WaveNet-based encoders, decoders, and vocoder. The efficiency of the resulting model is comparable to modern text-to-speech (TTS) and voice conversion (VC) systems when using 1–5 minutes speech samples without text supervision.

**Keywords:** voice cloning, generative neural network, text to speech, mel spectrogram, vocoder, style transfer, voice conversion, speaker adaptation.