



Г.А. Горшков, В.М. Замшев, Т.В. Ящун
Вологодский государственный университет

ИМИТАЦИЯ ГОЛОСА СРЕДСТВАМИ ТЕХНОЛОГИИ ГОЛОСОВОЙ КОНВЕРСИИ RVC v2

В статье предлагаются результаты анализа современных средств голосовых конверсий с помощью нейросетей, описание проблем применения данных технологий и результаты исследования одного из способов имитации голосов на основе RVC v2. Предлагаются результаты оценивания временных и вычислительных затрат процесса имитации голоса средствами непрофессиональной вычислительной техники. Результаты эксперимента подтверждают предположение о широких возможностях современных нейросетей для имитации чужого голоса и поднимают ряд соответствующих этических и правовых проблем.

Нейронные сети, искусственный интеллект, голосовая конверсия, нейронная голосовая модель, преобразование голоса на основе поиска, этические и правовые проблемы.

В последние годы методы и технологии синтеза и клонирования голосов значительно развились. Существуют разнообразные подходы решения данной проблемы [1], одним из которых являются нейронные технологии. Так, например, в 2023 году проведен ряд исследований, направленных на усовершенствование методов синтеза вокала и речи, которые предлагают инновационные подходы к решению существующих проблем.

Одним из ключевых направлений является улучшение синтеза вокала. Например, в [2] представлена модель GIOTtal-flow LPC Filter (GOLF), которая использует физические характеристики человеческого голоса для синтеза, что позволяет значительно улучшить качество результата. Модель CoMoSpeech [3], представленная Гонконгским университетом науки и технологий и Гонконгским институтом науки и инноваций, использует методы консистентного моделирования для синтеза речи и вокала, обеспечивая высокое качество звука при минимальном времени обработки.

Результаты исследований в данной области являются ошеломительными. Действительно, на сегодняшний день любой человек, имеющий достаточно мощный персональный компьютер и данные голоса другого человека, может с легкостью создать нейронную голосовую модель (NVM). NVM (Neural Voice Model) может быть создана с использованием современных достижений в области нейронных сетей на основе элементарных образцов исходного материала – нескольких минут вокала или речи человека. Примеры применения и исследования в данной области демонстрируют успешные результаты.

Одной из технологий, проявивших незаурядные результаты в исследуемой области, является преобразование голоса на основе поиска (RVC). Технология RVC (Retrieval-Based Voice Conversion) в отличие от традиционных методов применяет передовые алго-

ритмы для клонирования голосов с замечательной точностью [4].

RVC v2 – инновационная технология голосовой конверсии на основе искусственного интеллекта, эволюция оригинальной модели RVC. Она улучшена за счет усовершенствованных алгоритмов, что обеспечивает большую точность и качество клонирования голосов. RVC v2 позволяет переносить чужой голос на собственное исполнение, создавать дуэты с любыми артистами, экспериментировать с вокалом, меняя тембр и стиль [5]. На сегодняшний день RVC v2 применяют для создания каверов на песни, используя голоса популярных исполнителей, а также для изменения голоса в реальном времени.

Несмотря на значительные достижения в области RVC, остается ряд проблем и вызовов, которые необходимо решить для широкого применения и дальнейшего развития этих технологий:

- возможность клонирования голосов вызывает серьезные этические и правовые вопросы;
- текущие модели RVC могут создавать неестественные или искаженные результаты – эта проблема в ближайшем будущем может быть успешно устранена;
- модели RVC часто испытывают сложности при работе с различными акцентами и интонациями – улучшение адаптивности алгоритмов к различным языковым и культурным особенностям остается важной задачей для исследователей.

В ходе исследования была поставлена задача определить, какие трудовые и временные затраты необходимы для имитации голоса средствами нейросети RVC v2 с использованием вычислительных мощностей домашнего персонального компьютера. В качестве испытуемого выступил один из студентов-исследователей Владимир. В проведенном исследовании можно выделить несколько этапов.

Этап 1. Подготовка большого количества аудио-материала для нейросети.

Этап 2. Загрузка и установка программы RVC v2 [6].

Этап 3. Обучение нейросети (рис. 2).

Обучение производилось на основе небольшой инструкции и содержало следующие шаги:

- установка названия проекта;
- определение числа процессов ЦП, используемых для «вытаскивания» тональностей и обработки данных, например: если имеется 8 ядер и 16 потоков, то можно указать значение близкое к 16 (рис. 1);

выбор папки с аудиофайлами человека, голос которого будет имитировать нейросеть (рис. 2);

проработка данных в течение некоторого времени;

- выбор нужного алгоритма обучения – «вытаскивание черт» (рис. 2);

- тренировка нейросети – тренировка модели и индекса черт с указанием нужных настроек (частота сохранения, полное количество эпох) (рис. 3);

просмотр и анализ статистики процесса тренировки модели с учетом данных по каждой эпохе (рис. 4).

Рис. 1. Определение числа процессов ЦП

Рис. 2. Выбор папки с аудиофайлами и выбор алгоритма обучения

Шаг 3: Заполните остальные настройки тренировки и начните тренировать модель и индекс

Частота сохранения (save_every_epoch): <input type="range" value="5"/>	Полное количество эпох (total_epoch): <input type="range" value="300"/>	Размер пачки для видеокарты: <input type="range" value="12"/>	Сохранять только последний файл '.ckpt', чтобы сохранить место на диске: ☒ Да ☐ Нет
Кэшировать все тренировочные сетки в видеопамять. Кэширование маленьких датасетов (меньше 10 минут) может ускорить тренировку, но кэширование больших, наоборот, займёт много видеопамяти и не сильно ускорит тренировку: ☐ Да ☒ Нет		Сохранять маленькую финальную модель в папку 'weights' на каждой точке сохранения: ☒ Да ☐ Нет	

Путь к натренированной базовой модели G: pretrained_v2/f0G_40k.pth	Тренировать модель	Тренировать индекс черт	Тренировка одним нажатием
Путь к натренированной базовой модели D: pretrained_v2/f0D_40k.pth			

Рис. 3. Тренировка нейросети в программе RVC v2

```

F0:StarostaVovan:Saving model and optimizer state at epoch 10 to ./logs\StarostaVovan\G_2333333.pth
F0:StarostaVovan:Saving model and optimizer state at epoch 10 to ./logs\StarostaVovan\D_2333333.pth
F0:StarostaVovan:saving ckpt StarostaVovan_e10:Success.
F0:StarostaVovan:====> Epoch: 10 [2024-05-17 16:06:26] | (0:01:29.370525)
F0:StarostaVovan:====> Epoch: 11 [2024-05-17 16:07:51] | (0:01:24.432717)
F0:StarostaVovan:Train Epoch: 12 [49%]
F0:StarostaVovan:[1000, 9.98501030820433e-05]
F0:StarostaVovan:loss_disc=4.031, loss_gen=3.201, loss_fm=8.646, loss_mel=19.939, loss_kl=1.856
F0:StarostaVovan:====> Epoch: 12 [2024-05-17 16:09:15] | (0:01:23.980102)
  
```

Рис. 4. Информация о процессе тренировки нейросети

Model Slot Configuration		
	[14]	StarostaVovan_e300_s26100
	model:	StarostaVovan_e300_s...
	index:	added_IVF3111_Flat_n...
	info:	f0, 40000, 768, pyTorchRVCv2, 0, 0, 0.5
blank	[15]	blank
blank	[16]	blank

Рис. 5. Результаты обучения нейросети

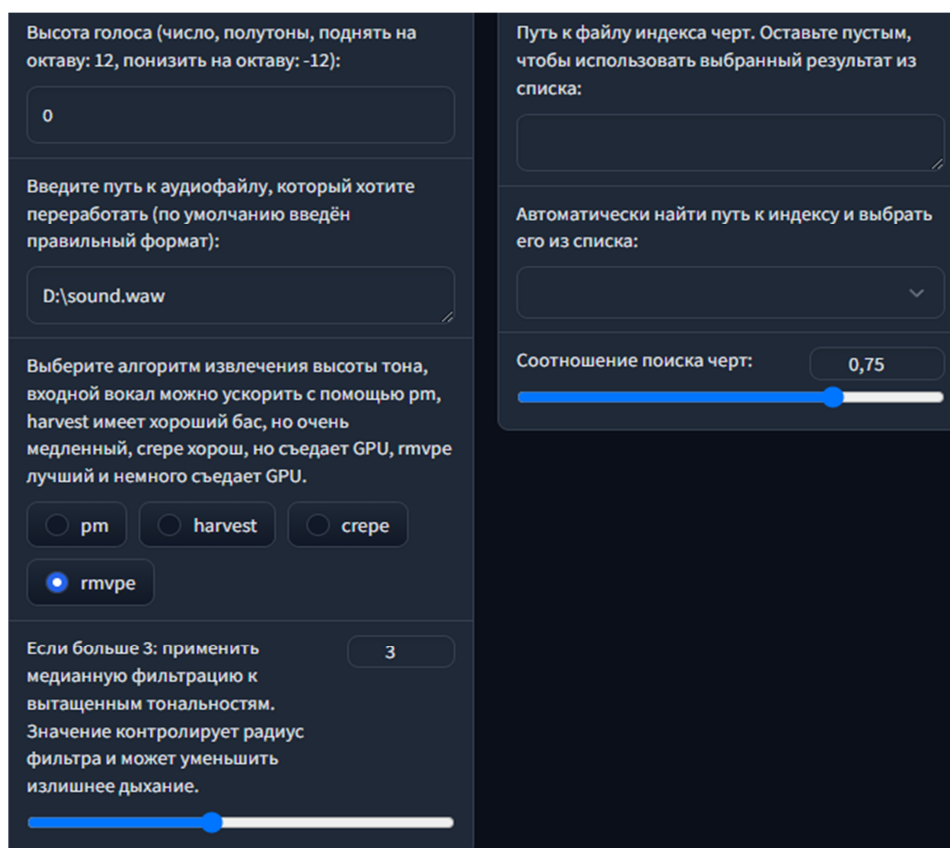


Рис. 6. Установка настроек для обработки вокала на основе полученной модели

Этап 4. Оценка временных и производственных затрат.

Нейросеть прошла 300 эпох. На 1 эпоху пришлось примерно 1,5 минуты. В результате была получена модель голоса Владимира и его индекс. Было потрачено примерно 8 часов, без учета времени на знакомство с программой и подбор нужных настроек.

Нагрузка на вычислительные устройства была не очень значительна:

- загрузка центрального процессора (AMD Ryzen 5700G) – 50 %;
- нагрузка на видеокарту (Nvidia GeForce RTX3060) – 70 %.

Параллельно с обучением нейросети была возможность выполнять низкозатратные процессы. Тренировка нейросети не мешала выполнению рутинных работ: просмотр фильмов, работа с приложениями MS Office, например MS Word или MS PowerPoint, прослушивание музыки и т.д. Вместе с тем использование других программ приводило к перезапуску системы.

Этап 5. Анализ результатов обучения нейросети.

Полученная обученная модель и индекс предоставили возможность говорить голосом Владимира в программе Voice Changer [7] (рис. 5). То есть первоначальный результат был достигнут – создана возможность что-то сказать или отправить голосовое сообщение симитированным голосом.

Улучшение имеющихся результатов возможно путем проведения дополнительных эпох обучения нейросети.

Этап 6. Применение обученной нейросети.

Следующим этапом исследования являлось применение нейросети для имитации вокала Владимира в известном музыкальном хите:

- с помощью программы Ultimate Vocal Remover [8] выполнено разделение инструментальной и вокальной составляющих песни;
- обработан вокал в программе RVC v2 на основе полученных модели и индекса с применением различных настроек – конвертирование было проведено неоднократно с разными настройками для достижения оптимального результата (рис. 6);
- загрузка обработанного вокала;
- сведение вокала с инструментальной частью песни с использованием соответствующего программного сервиса [9] и получение кавер-версии от нейросети.

Ознакомиться с полученными результатами имитации голоса можно по ссылке: https://drive.google.com/drive/folders/1zbeHVE_4oz8RylAMlv5D4O9NbiW8A9hM?usp=sharing.

Этап 7. Подведение итогов.

По результатам исследования было определено, что для использования чужого голоса не требуется чрезвычайно много ресурсов и времени. Действительно, менее чем за 24 часа при наличии исходного материала (Dataset), можно с помощью нейросети симитировать голос любого человека.

С уверенностью можно сказать, что развитие технологий RVC открывает широкие возможности

для применения нейросетей в области преобразования голоса, предоставляя инновационные инструменты для создания реалистичных и точных голосовых моделей, которые можно применять во многих сферах.

Однако для полноценной реализации потенциала таких технологий необходимо решить ряд значительных этических, технических и правовых проблем. Действительно, описанная технология может использоваться мошенниками для создания преступных схем, в которых голос может быть скопирован и использован в корыстных целях, например для имитации речи популярной личности и обмана людей, которые ей доверяют. Следовательно, необходима разработка четких законодательных норм и правил для регулирования использования нейронных голосовых моделей. При этом технически сложной задачей является сохранение и распознавание естественности и уникальности оригинального голоса.

Дальнейшее совершенствование алгоритмов и более детальное изучение различных аспектов RVC позволит минимизировать риски и максимально использовать возможности данной технологии. Разработчики и исследователи должны продолжать сотрудничество в направлении создания более эффективных и безопасных решений, обеспечивая надлежащую защиту данных и этическое использование нейронных голосовых моделей.

Литература

1. Меркулов, П. Д. Аналитическое сравнение методов синтеза звука / П. Д. Меркулов // Молодой ученый. – 2022. – № 16 (411). – С. 31–34. – URL: <https://moluch.ru/archive/411/90535/> (дата обращения: 27.05.2024). – Текст : электронный.
2. Yu, C., Fazekas, G. Singing Voice Synthesis Using Differentiable LPC and Glottal-Flow-Inspired Wavetables / Chin-Yun Yu, & György Fazekas // Proceedings of the 24th International Society for Music Information Retrieval Conference. – 2023 – pp. 667–675. – URL: <https://doi.org/10.5281/zenodo.10265377/> (дата обращения: 27.05.2024). – Текст : электронный.
3. Ye, Z., Xue, W., Tan, X., Chen, J., Liu, Q., & Guo, Y. (2023). CoMoSpeech: One-Step Speech and Singing Voice Synthesis via Consistency Model / Z. Ye, W. Xue, X. Tan, J. Chen, Q. Liu, Y. Guo // In MM 2023 – Proceedings of the 31st ACM International Conference on Multimedia (1st ed). – 2023. – pp. 1831–1839. – URL: <https://doi.org/10.1145/3581783.3612061> (дата обращения: 27.05.2024). – Текст : электронный.
4. Speechify Academy : официальный сайт – URL: <https://studio.speechify.com/dashboard> (дата обращения: 27.05.2024).
5. Данилов, Я. Революция в мире музыки с RVC v2 (Нейросеть) / Ярослав Данилов. – URL: <https://tenchat.ru/media/1502557-revolyutsiya-v-mire-muzyki-s-rvc-v2-neyroset> (дата обращения: 27.05.2024). – Текст : электронный.
6. Hugging Face : официальный сайт – URL: <https://huggingface.co/RVC-RU/RVC-Nyamura-RU> (дата обращения: 27.05.2024).
7. Программа Voice Changer – URL: https://github.com/w-okada/voice-changer/blob/master/README_en.md (дата обращения: 27.05.2024).
8. Ultimate Vocal Remover : официальный сайт - <https://ultimatevocalremover.com/> (дата обращения: 19.03.2024).
9. Create more content in less time – URL: <https://www.kapwing.com/> (дата обращения: 19.03.2024).

G.A. Gorshkov, V.M. Zamshev, T.V. Yashchun
Vologda State University

VOICE IMITATION USING RVC V2 VOICE CONVERSION TECHNOLOGY

The article offers the results of the analysis of modern voice conversion tools using neural networks, a description of the problems of using these technologies, and the results of a study of one of the methods for simulating voices based on RVC v2. The results of estimating the time and computational costs of the voice imitation process using non-professional computer technology are presented. The results of the experiment confirm the assumption and the wide possibilities of modern neural networks for simulating someone else's voice and raise a number of relevant ethical and legal problems.

Neural networks, artificial intelligence, voice conversion, AI voices, retrieval-based voice conversion, ethical and legal issues.