DOI: 10.15514/ISPRAS-2024-36(4)-1

Comparison of Voice Cloning Algorithms in Zero-shot and Few-shot Scenarios

O. Hovhannisyan, ORCID: 0009-0000-9384-7945 <olga.hovhannisyan@student.rau.am>
D. Sargsyan, ORCID: 0009-0006-0349-6031 <d.sargsyan@ispras.ru>
A. Malajyan, ORCID: 0000-0002-3566-9316 <malajyanarthur@ispras.ru>

Russian-Armenian University,
123, Hovsep Emin st., Yerevan, 0051, Armenia.

Abstract. Voice cloning technology has made significant strides in recent years, with applications ranging from personalized virtual assistants to sophisticated entertainment systems. This study compares nine voice cloning models, focusing on both zero-shot and fine-tuned approaches. Zero-shot voice cloning models have gained attention for their ability to generate high-quality synthetic voices without requiring extensive training data for each new voice and for their capability to perform real-time inference online. In contrast, non-zero- shot models typically require additional data but can offer improved fidelity in voice reproduction. The study comprises two key experiments. The first experiment evaluates the performance of zero-shot voice cloning models, analyzing their ability to reproduce target voices without prior exposure accurately. The second experiment involves fine-tuning the models on target speakers to assess improvements in voice quality and adaptability. The models are evaluated based on key metrics assessing voice quality, speaker identity preservation, and subjective and objective performance measures. The findings indicate that while zero-shot models offer greater flexibility and ease of deployment, fine-tuned models can deliver superior performance.

Keywords: voice cloning; zero-shot cloning; fine-tuning; speech synthesis; speaker adaptation.

For citation: Hovhannisyan O., Sargsyan D., Malajyan A. Comparison of voice cloning algorithms in zero-shot and few-shot scenarios. *Trudy ISP RAN/Proc. ISP RAS*, vol. 36, issue 4, 2024, pp. 7-16. DOI: 10.15514/ISPRAS-2024-36(4)-1.

Acknowledgements. This work was supported by the Science Committee of RA (Research project № 23AA-1B006).

Сравнение алгоритмов клонирования голоса в условиях нулевого и малого количества примеров

О. Оганесян, ORCID: 0009-0000-9384-7945 <olga.hovhannisyan@student.rau.am>
Д. Саргсян, ORCID: 0009-0006-0349-6031 <d.sargsyan@ispras.ru>
А. Маладжян, ORCID: 0000-0002-3566-9316 <malajyanarthur@ispras.ru>

Российско-Армянский университет,
Армения, 0051, Ереван, ул. Овсепа Эмина, 123.

Аннотация. Технология клонирования голоса сделала значительные шаги вперед в последние годы, с применением от персонализированных виртуальных ассистентов до сложных развлекательных систем. В данном исследовании проводится сравнение девяти моделей клонирования голоса, сосредотачиваясь на подходах нулевого и тонкой настройки. Модели клонирования голоса с нулевым обучением привлекают внимание своей способностью генерировать высококачественные синтетические голоса без необходимости в больших объемах обучающих данных для каждого нового голоса, а также возможностью осуществлять онлайн выводы в режиме реального времени. В отличие от них, модели, не относящиеся к нулевому обучению, обычно требуют дополнительных данных, но могут обеспечить улучшенную точность воспроизведения голоса. Исследование включает два ключевых эксперимента. Первый эксперимент оценивает эффективность моделей клонирования голоса с нулевым обучением, анализируя их способность точно воспроизводить целевые голоса без предварительного ознакомления. Второй эксперимент включает тонкую настройку моделей на целевых спикеров для оценки улучшений в качестве голоса и адаптивности. Модели оцениваются на основе ключевых показателей, оценивающих качество голоса, сохранение идентичности спикера, а также субъективные и объективные показатели производительности. Результаты показывают, что, хотя модели с нулевым обучением предлагают большую гибкость и простоту использования, модели с тонкой настройкой могут обеспечить более высокую производительность.

Ключевые слова: клонирование голоса; клонирование с нулевым обучением; тонкая настройка; синтез речи; адаптация говорящего.

Для цитирования: Оганесян О., Саргсян Д., Маладжян А. Сравнение алгоритмов клонирования голоса в условиях нулевого и малого количества примеров. Труды ИСП РАН, том 36, вып. 4, 2024 г., стр. 7–16 (на английском языке). DOI: 10.15514/ISPRAS–2024–36(4)–1.

Благодарности. Работа выполнена при поддержке Комитета по науке Республики Армения (исследовательский проект № 23AA-1B006).

1. Introduction

Voice cloning technology has advanced rapidly, enabling the creation of synthetic voices that closely mimic human speech. This technology has significant applications in personalized virtual assistants, entertainment, and communication. The core challenge in voice cloning is to produce synthetic voices that are indistinguishable from human voices while preserving the unique characteristics of the target speaker.

Two primary methodologies have emerged in voice cloning: zero-shot cloning and fine-tuning. Zero-shot models, such as XTTS 2 [1], StyleTTS [2], YourTTS [3], OpenVoice [4], VoiceCraft [5], Vall-E-X [6], and Natural Speech 3 [7], can generate voices without extensive speaker-specific training data, offering flexibility and scalability. However, maintaining voice quality and identity without prior exposure to the target speaker remains challenging. Fine-tuning models, including VITS [8] and RVC [9], improve voice fidelity by adapting pre-trained models with additional data from the target speaker, although they require more data and computational resources.

This study focuses on models that excel in flexibility, scalability, and efficiency in zero-shot and few-shot scenarios. Older models like WaveNet [10], Deep Voice [11], SV2TTS [12], TortoiseTTS [13], Tacotron [14], and Glow-TTS [15] are excluded due to their high computational demands, extensive data requirements, and complexity.

This paper evaluates nine voice cloning models through two key experiments. The first experiment tests the zero-shot capabilities of the models, assessing their performance in replicating voices without prior exposure. The second experiment involves fine-tuning the models on target speakers to evaluate enhancements in voice quality and adaptability. Evaluation metrics include speaker embedding cosine similarity (SECS [16]) for identity preservation, Mel cepstral distortion (MCD[1] [17]) for spectral similarity, F0 mean absolute error (F0 MAE [17]) for pitch accuracy, F0 Pearson correlation coefficient (F0-PCC [18]) for pitch contour correlation, and universal target mean opinion score (UTMOS2[2] [1]) for subjective quality.

The paper is organized as follows: Section 2 provides a detailed description of each voice cloning model. Section 3 outlines the experimental setup and presents the experiments and results. Only original manuscripts that have not been previously published nor in other editions, neither in the Internet, are accepted for publication in Proceedings of ISP RAS. The authors of the articles can be ISP RAS staff or representatives of other organizations. Only manuscripts in Russian or English are allowed to be published. As a rule, the volume of published articles should not be less than 8-9 pages, and shouldn't exceed 20 pages.

2. Overview of Voice Cloning Models

In recent years, various voice cloning models have emerged, each offering unique approaches and capabilities. This section provides an overview of the voice cloning models discussed in the introduction, highlighting their key features and processes.

Variational Inference Text-to-Speech (VITS): The VITS model is designed to generate speech directly from text. It incorporates a stochastic duration predictor to capture natural speech rhythms, enabling the production of authentic and fluid voice waveforms. This end-to-end approach supports high-quality voice cloning by accurately translating text into speech with precise timing and natural intonation. However, VITS is not a zero-shot model and requires training on target voices beforehand.

Retrieval-based Voice Conversion (RVC): The RVC model is a system for converting one speaker's voice into another's. It leverages a retrieval-based approach to map and synthesize voice characteristics from a database of target voices. This method enables high-quality voice transformation by accurately capturing and replicating speaker-specific traits. Unlike zero-shot models, RVC needs to be trained on a set of target voices to effectively perform voice conversion.

OpenVoice: OpenVoice uses simply a short audio clip from the reference speaker to generate speech in multiple languages. It provides flexible control over voice styles and enables zero-shot cross-lingual cloning, though it requires a TTS model trained for the target language.

StyleTTS 2: StyleTTS 2 generates high-quality, natural-sounding speech using style diffusion techniques. It models voice styles as latent variables, enabling it to clone voices with no need for specific reference recordings. By leveraging large pre-trained speech language models and innovative training methods, StyleTTS 2 excels in producing expressive and accurate voice clones, including effective zero-shot speaker adaptation.

VoiceCraft: VoiceCraft excels in speech editing and zero-shot text-to-speech generating. It uses a Transformer decoder and an innovative token rearrangement method to generate high-quality, natural-sounding speech by efficiently reconstructing and infilling speech tokens. It analyzes and replicates the vocal characteristics of a target speaker, capturing emotional tone and subtle vocal nuances to produce realistic and engaging speech.

YourTTS: YourTTS builds on the VITS model, excelling in zero-shot voice cloning and multi-speaker text-to-speech with minimal data. It performs well across various languages and can adapt

[1] https://pypi.org/project/pymcd/

[2] https://github.com/sarulab-speech/UTMOSv2

to new voices with less than one minute of audio. However, it may occasionally face issues with speech duration and mispronunciations.

VALL-E-X: VALL-E X is a cross-lingual neural codec model that excels in zero-shot text-to-speech and speech-to-speech translation. It generates high-quality speech in a target language from a single utterance in a source language, preserving the speaker's voice and emotion. The model avoids the need for paired cross-lingual data and effectively addresses foreign accent issues, making it suitable for diverse multilingual applications.

XTTS-2: XTTS 2 is a multilingual zero-shot text-to-speech (TTS) model trained in 16 languages. Building on the Tortoise model, XTTS 2 enhances voice cloning, speed, and multilingual capabilities. It achieves high-quality results in prosody and style mimicking, including whispering, with minimal fine-tuning data. XTTS 2 is notably faster than previous models like VALL-E.

Natural Speech 3: NaturalSpeech 3 generates high-quality, natural-sounding speech by separating and controlling speech attributes like content, prosody, and timbre. Its novel factorized diffusion approach allows for detailed and accurate speech synthesis, achieving superior performance and human-level quality on diverse datasets.

WaveNet: WaveNet generates raw audio waveforms using an autoregressive model, achieving high naturalness. However, it requires substantial computational resources and has slow inference times.

Deep Voice: Deep Voice uses a modular pipeline to produce human-like speech but requires extensive speaker-specific training data. Modern models overcome this limitation by utilizing less data, enabling more flexible and scalable voice cloning.

SV2TTS: SV2TTS employs a three-stage pipeline for voice cloning but struggles with voice quality and identity preservation without extensive fine-tuning.

Tortoise TTS: Tortoise TTS excels in expressive speech synthesis but demands significant computational resources and data for adaptation. Its complexity and inefficiency make it impractical for zero-shot applications requiring minimal data.

Tacotron: Tacotron generates speech from text with high naturalness but relies on the Griffin-Lim [19] algorithm, which can introduce artifacts. It requires substantial training data, limiting its effectiveness in zero-shot learning scenarios that require rapid adaptation.

Glow-TTS: Glow-TTS offers efficient parallel synthesis but lacks built-in support for speaker adaptation, necessitating additional modifications and data. Its focus on general TTS tasks rather than speaker-specific scenarios reduces its suitability for robust zero-shot application.

Table 1 provides an overview of the voice cloning models discussed, highlighting their zero-shot capabilities and the number of parameters.

3. Experimental setup and results

The experiments aim to evaluate the performance of nine voice cloning models – XTTS 2, StyleTTS 2, YourTTS, VITS, OpenVoice, RVC, VoiceCraft, Vall-E-X, and Natural Speech 3 – using both zero-shot and fine-tuning approaches. The goal is to assess each model's ability to reproduce target voices with high quality and fidelity.

3.1 Experimental Setup

The experiments utilize the VCTK corpus [20], which includes speech data from 109 English speakers with various accents. The experiment is structured as follows:

- **Zero-Shot Experiment:** We select 30 speakers from the VCTK dataset, using 5 audio samples per speaker, each ranging in duration from 7 to 10 seconds. Models requiring text input are evaluated with additional sentences from the remaining speakers of the same dataset, consisting of 6-15 words on average. Audio-to-audio models are tested using audio samples instead sentences from the same speakers' set, containing audios in the range of 5–10 seconds.

- **Fine-Tuning Experiment:** Each model is fine-tuned on the same 30 speakers using 10 minutes of audio per speaker (excluding test samples). The fine-tuning process involves training for 100 epochs on a single NVIDIA RTX 3060 12GB GPU. After fine-tuning, the models are tested on the same data from the zero-shot experiment.

Table 1. Comparison of Voice Cloning Models with Zero-Shot Capability and Parameters.

Model	Zero-shot	Fine-tune ability	Params	Text Ref + Audio Ref	Promt + Text Ref + Audio Ref	Audio Orig + Audio Ref	Audio Ref + Speaker ID
XTTS 2	✓	✓	518M	✓			
StyleTTS 2	✓	✓	218M	✓			
YourTTS	✓	✓	94M	✓			
VITS	×	✓	39M				✓
OpenVoice	✓	×	32M	✓			
RVC	×	✓	27M				✓
VoiceCraft	✓	✓	830M		✓		
Vall-E-X	✓	✓	300M		✓		
Natural Speech 3	✓	×	1B			✓	

3.2 Evaluation Metrics

Models are evaluated using:

- **Speaker Embedding Cosine Similarity (SECS):** Measures the retention of the speaker's identity. Values close to 1 are better**,** as they indicate a greater similarity of the speaker's identity.
- **Mel Cepstral Distortion (MCD):** Assesses spectral similarity between synthesized and reference voices. Lower values are better**,** suggesting a closer match to the reference voice and thus better spectral quality.
- **F0 Mean Absolute Error (F0 MAE):** Measures the accuracy of pitch reproduction. Lower values are better**,** as they indicate a more accurate pitch reproduction.
- **F0 Pearson Correlation Coefficient (F0-PCC):** Assesses correlation between generated and reference pitch. Higher values are better, with a value of 1 indicating a perfect correlation, demonstrating that the model effectively captures and replicates the pitch dynamics of the original voice.
- **Universal Target Mean Opinion Score (UTMOS2):** Evaluates subjective voice quality.

Values closer to 4 are better, indicating good voice quality and naturalness.

3.3 Results

In this section, we present the results of our two experiments:

- **Zero-Shot Voice Cloning:** Table 2 shows the performance of 7 out of 9 models capable of zero-shot voice cloning.
- **Fine-Tuning:** Table 3 details the results for 7 out of 9 models fine-tuned with 10 minutes of audio per speaker, comparing their performance to the zero-shot results (OpenVoice and Natural Speech 3 do not have available implementations for fine-tuning).

We analyze the results for male (M) and female (F) voices separately, as presented in the tables. In the zero-shot experiments, the SECS scores indicate that most models effectively preserve speaker identity, with scores ranging from 0.75 to 0.78 for male voices and 0.72 to 0.8 for female voices. Natural Speech 3 and XTTS 2 perform particularly well in this regard. However, spectral fidelity, as measured by MCD (Mel Cepstral Distortion), shows significant disparities. Natural Speech 3 achieves the lowest MCD (9.7 dB for males), indicating better spectral accuracy, while VoiceCraft exhibits much higher distortion, with MCD values reaching 24 dB for males and 23.7 dB for females. Female voices generally suffer from higher MCD values across all models, indicating greater spectral distortion and less natural-sounding results compared to male voices. Pitch accuracy, reflected by F0, varies widely among the models. VoiceCraft shows the highest pitch at 196.4 Hz for males, indicating a significant difference of pitch, while Natural Speech 3 the lowest pitch at 55.9 Hz. For female voices, F0 values also vary, with XTTS 2 producing the highest pitch at 113.8 Hz and Natural Speech 3 the lowest at 72.5 Hz. The F0-PCC (F0 Pearson Correlation Coefficient) scores, which measure pitch contour accuracy, are moderate across the board, with values around 0.3 to 0.4 for both genders. It suggests that while some pitch dynamics are captured, the models struggle with accurate pitch reproduction. UTMOS2 scores reflect these trends, with Natural Speech 3, StyleTTS 2, and OpenVoice achieving the highest perceived quality for male voices (up to 3.6), while female voices generally score lower, reaching the highest result of 3.6 for StyleTTS 2.

In the fine-tuning experiments, SECS scores remain high, showing continued voice resemblance. XTTS 2 and YourTTS maintain good scores, but there is no significant improvement over the zero-shot scenario. Fine-tuning does lead to notable improvements in MCD for some models, especially XTTS 2, which reduces MCD from 16.6 dB to 9.1 dB for males, indicating better spectral fidelity. However, not all models benefit equally; VoiceCraft's MCD remains high, particularly for females, and Vall-E-X experiences a drastic increase in MCD for females, rising from 12.8 dB to 43.9 dB, indicating worsened spectral accuracy. Pitch accuracy shows mixed results post fine-tuning. XTTS 2 improves pitch consistency, reducing F0 from 134.2 Hz to 87.7 Hz for males, aligning better with typical pitch ranges. However, Vall-E-X exhibits worsened F0 accuracy, particularly for females. F0-PCC values remain stable, indicating little improvement in pitch contour accuracy, and UTMOS2 scores show minor gains, with XTTS 2 and YourTTS performing slightly better.

When comparing models across both experiments, XTTS 2 shows significant improvements in spectral fidelity, with MCD reducing by 7.5 dB (from 16.6 dB to 9.1 dB for males), and in pitch accuracy, with F0 improving by 46.5 Hz (from 134.2 Hz to 87.7 Hz for males). YourTTS exhibits moderate gains, reducing MCD by 6.9 dB (from 17.3 dB to 10.4 dB for males) and showing slight improvements in UTMOS2 scores, increasing by 0.1 points for males (from 2.98 to 3.08). After fine-tuning UTMOS2 decreased to 3.43 for males (down from 3.56) and 3.22 for females (down from 3.58) for StyleTTS 2. VoiceCraft performs worse post fine-tuning, with MCD reducing by 7.7 dB for males (from 24 dB to 16.3 dB) but no significant improvements in F0 accuracy, which changes by only 0.3 Hz for males (from 196.4 Hz to 196.1 Hz). UTMOS2 scores for VoiceCraft increase by 0.15 points for males (from 2.81 to 2.96), but there is still room for improvement. Vall- E-X experiences a drastic worsening in spectral fidelity for female voices, with MCD increasing by 31.1 dB (from 12.8 dB to 43.9 dB), and F0 accuracy declining by 25.8 Hz (from 85.3 Hz to 59.5 Hz).

Table 2. Results of voice cloning for zero-shot models.

Model	SECS(⬆)	MCD(⬇)	F0(⬇)	F0-PCC(⬆)	UTMOS2(⬆)
XTTS 2	**M**: 0.78 **F**: 0.77	**M:** 16.6 db **F:** 21 db	**M:** 134.2 Hz **F:** 113.8 Hz	**M:** 0.4 **F:** 0.4	**M**: 3.07 **F**: 2.76
StyleTTS 2	**M**: 0.75 **F**: 0.75	**M:** 10.7 db **F:** 12.9 db	**M:** 91.4 Hz **F:** 92.2 Hz	**M:** 0.34 **F:** 0.3	**M**: 3.56 **F**: 3.58
YourTTS	**M**: 0.77 **F**: 0.76	**M:** 17.3 db **F:** 17.8 db	**M:** 126.6 Hz **F:** 97.6 Hz	**M:** 0.4 **F:** 0.4	**M**: 2.98 **F**:2.56
OpenVoice	**M**: 0.75 **F**: 0.72	**M:** 19.1 db **F:** 15.4 db	**M:** 111.2 Hz **F:** 110.3 Hz	**M:** 0.4 **F:** 0.3	**M**: 3.51 **F**: 3.23
VoiceCraft	**M:** 0.77 **F**: 0.75	**M:** 24 db **F:** 23.7 db	**M:** 196.4 Hz **F:** 99.4 Hz	**M:** 0.3 **F:** 0.4	**M:** 2.81 **F:** 2.6
Vall-E-X	**M**: 0.75 **F**: 0.76	**M:** 16.2 db **F:** 12.8 db	**M:** 84.8 Hz **F:** 85.3 Hz	**M:** 0.3 **F:** 0.3	**M**: 3.06 **F**: 2.89
Natural Speech 3	**M**: 0.78 **F**: 0.8	**M:** 9.7 db **F:** 9.4 db	**M:** 55.9 Hz **F:** 72.5 Hz	**M:** 0.3 **F:** 0.3	**M**: 3.58 **F**: 3.41

Table 3. Results of voice cloning for fine-tuned models.

Model	SECS(⬆)	MCD(⬇)	F0(⬇)	F0-PCC(⬆)	UTMOS2(⬆)
XTTS 2	**M**: 0.77 **F**: 0.76	**M:** 9.1 db **F:** 12.2 db	**M:** 87.7 Hz **F:** 92.6 Hz	**M:** 0.3 **F:** 0.4	**M**: 3.31 **F**: 2.87
StyleTTS 2	**M**: 0.67 **F**: 0.7	**M:** 15.4 db **F:** 11.8 db	**M:** 34.1 **F:** 79.3	**M:** 0.4 **F:** 0.4	**M**: 3.43 **F**: 3.22
YourTTS	**M:** 0.77 **F:** 0.74	**M:** 10.4 db **F:** 14.1 db	**M:** 88.1 Hz **F:** 95.4 Hz	**M:** 0.3 **F:** 0.5	**M**: 3.08 **F**: 2.88
VITS	**M:** 0.53 **F:** 0.58	**M:** 25.8 db **F:** 21.8 db	**M:** 111.9 Hz **F:** 154.8 Hz	**M:** 0.4 **F:** 0.4	**M**: 3.02 **F**: 3.27
VoiceCraft	**M:** 0.76 **F:** 0.73	**M:** 16.3 db **F:** 13.5 db	**M:** 76.8 Hz **F:** 101.4 Hz	**M:** 0.3 **F:** 0.4	**M**: 2.96 **F**: 2.7
Vall-E-X	**M:** 0.67 **F:** 0.75	**M:** 33.2 db **F:** 43.9 db	**M:** 26.8 Hz **F:** 60 Hz	**M**: 0.3 **F:** 0.3	**M:** 2.26 **F:** 2.65
RVC	**M:** 0.72 **F:** 0.71	**M:** 9.9 db **F:** 10.7 db	**M:** 58.9 Hz **F:** 114.2 Hz	**M:** 0.3 **F:** 0.3	**M**: 2.87 **F**: 2.68

4. Conclusion

This study presents a comparison of nine voice cloning algorithms across zero-shot and fine-tuning scenarios. Zero-shot models demonstrate flexibility and satisfactory performance without the need for extensive data, making them highly suitable for rapid deployment. However, these models face challenges in maintaining spectral accuracy, as evidenced by elevated MCD values, particularly for female voices.

Fine-tuning introduces significant improvements in spectral fidelity and pitch accuracy for some models, notably XTTS 2 and YourTTS. XTTS 2 shows a reduction in MCD and an improvement in F0 for males, while YourTTS reduces MCD and slightly improves UTMOS2 scores. However, the impact of fine-tuning is mixed for other models. For instance, StyleTTS 2 experiences a mixed effect on perceived quality with a UTMOS2 increase for males but a slight decrease for females. Meanwhile, VoiceCraft and Vall-E-X exhibit worsened spectral fidelity and pitch accuracy post fine-tuning, especially for female voices.

Overall, fine-tuning successfully enhances certain aspects of voice cloning for specific models and presents opportunities for further refinement to extend these improvements to other models as well.

References

[1]. Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, Julian Weber. XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model. arXiv:2406.04904, 2024.

[2]. Yinghao Aaron Li, Cong Han, Vinay S. Raghavan, Gavin Mischler, Nima Mesgarani. StyleTTS 2: Towards Human-Level Text-to-Speech through Style Diffusion and Adversarial Training with Large Speech Language Models. arXiv preprint arXiv:2306.07691, 2023

[3]. Edresson Casanova, Julian Weber, Christopher Shulby, Arnaldo Candido Junior, Eren Gölge, Moacir Antonelli Ponti. YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for everyone. arXiv preprint arXiv:2112.02418, 2023.

[4]. Zengyi Qin, Wenliang Zhao, Xumin Yu, Xin Sun. OpenVoice: Versatile Instant Voice Cloning. arXiv preprint arXiv:2312.01479v5, 2024.

[5]. Puyuan Peng, Po-Yao Huang, Daniel Li, Abdelrahman Mohamed, David Harwath. VoiceCraft: Zero-Shot Speech Editing and Text-to-Speech in the Wild. arXiv preprint arXiv:2403.16973v1, 2024.

[6]. Ziqiang Zhang, Long Zhou, Chengyi Wang, Sanyuan Chen, Yu Wu, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, Furu Wei. Speak Foreign Languages with Your Own Voice: Cross-Lingual Neural Codec Language Modeling. arXiv preprint arXiv:2303.03926v1, 2023.

[7]. Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiang-Yang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, Sheng Zhao. NaturalSpeech 3: Zero-Shot Speech Synthesis with Factorized Codec and Diffusion Models. arXiv:2403.03100, 2024

[8]. Jaehyeon Kim, Jungil Kong, Juhee Son. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. arXiv:2106.06103

[9]. Retrieval based Voice Cloning. Available at the link: https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI

[10]. Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. arXiv:1609.03499, 2016.

[11]. Sercan O. Arik, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, Shubho Sengupta, Mohammad Shoeybi. Deep Voice: Real-time Neural Text-to-Speech. arXiv:1702.07825, 2017.

[12]. Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu. Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis. arXiv:1806.04558, 2018.

[13]. James Betker. Better speech synthesis through scaling. arXiv:2305.07243, 2023.

[14]. Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, Rif A. Saurous. Tacotron: Towards End-to-End Speech Synthesis. arXiv:1703.10135, 2017.

[15]. Jaehyeon Kim, Sungwon Kim, Jungil Kong, Sungroh Yoon. Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search. arXiv:2005.11129, 2020.
[16]. L. Wan, Q. Wang, A. Papir, and I. L. Moreno, “Generalized end-to-end loss for speaker verification” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 487–488.
[17]. Robert F. Kubichek. Mel-cepstral distance measure for objective speech quality assessment. Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing, 1:125–128 vol.1, 1993.
[18]. J. Benesty, J. Chen, Y. Huang, and I. Cohen, “Pearson correlation coefficient,” in Noise reduction in speech processing. Springer, pp. 1–4, 2009.
[19]. Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. IEEE Transactions on Acoustics, Speech, and Signal Processing, 32(2):236–243, 1984.
[20]. Veaux, Christophe; Yamagishi, Junichi; MacDonald, Kirsten. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit, [sound]. University of Edinburgh. The Centre for Speech Technology Research (CSTR). https://doi.org/10.7488/ds/1994, 2017.

Информация об авторах / Information about authors

Ольга ОГАНЕСЯН – научный сотрудник Центра передовых программных технологий (CAST) и аспирант Российско-Армянского университета, специализируется на математическом и программном обеспечении вычислительных систем. Получила степень бакалавра по информатике и прикладной математике в Российско-Армянском Университете Армении (2021) и степень магистра по интеллектуальным системам и робототехнике в Российско-Армянском университете (2023). Её исследования сосредоточены на обработке речи и синтезе голоса, включая клонирование голоса, а также на развитии методов машинного обучения.

Olga HOVHANNISYAN is a researcher at the Center of Advanced Software Technologies (CAST) and a Ph.D. student at Russian-Armenian University, specializing in mathematical and software support for computing systems. She holds a B.Sc. in Informatics and Applied Mathematics from the Russian- Armenian University of Armenia (2021) and an M.Sc. in intellectual systems and robotics from Russian-Armenian University (2023). Her research centers on speech processing and voice synthesis, including voice cloning, as well as advancements in machine learning.

Давид САРГСЯН студент бакалавра в Российско-Армянского университете по направлению Прикладной Математики и Информатики. Он также является исследователем в Центре Передовых Программных Технологий (CAST). Его научные интересы включают автоматическое распознавание речи, технологии синтеза речи и большие языковые модели (LLM).

David SARGSYAN is a Bachelor’s student in Applied Mathematics and Informatics at the Russian-Armenian University. He is also a researcher at the Center of Advanced Software Technologies (CAST). His research interests include speech recognition, text-to-speech technologies, and large language models (LLMs).

Артур МАЛАДЖЯН – бакалавр в области информатики и прикладной математики в Российско-Армянском Университете Армении (2020). Магистр в области машинного обучения в Российско-Армянском Университете, Армения (2022). В настоящее время он является исследователем в Центре Передовых Программных Технологий (CAST). Сфера научных интересов: обработка естественного языка и голосовые технологии.

Artur MALAJYAN received his Bachelor’s degree in Informatics and Applied Mathematics from the Russian-Armenian University in 2020. In 2022, he earned a Master’s degree in Machine Learning from Russian-Armenian University, Armenia. He is currently a researcher at the Center of Advanced Software Technologies (CAST). His research interests include natural language processing (NLP) and voice technologies.