УДК 004.8

САМОФОРМИРУЕМЫЙ ЛОКАЛЬНЫЙ ДИАЛОГОВЫЙ ВИРТУАЛЬНЫЙ ДВОЙНИК ЧЕЛОВЕКА

Кузнецов Э. С.[1]
Научный руководитель – канд. техн. наук, доцент Ефимова В. А.[1]
[1]Университет ИТМО
ernestkuznecov8@gmail.com

Введение

Актуальной задачей является развитие технологий цифрового сохранения личности, позволяющих воспроизводить не только факты о человеке, но и его стиль общения, голос и визуальный образ в формате интерактивного диалога. Практический интерес к таким системам связан с потребностью в более «живом» цифровом представлении человека по сравнению со статичными архивными материалами (текстами, аудио- и видеозаписями).

Существующие сервисы и решения, как правило, реализуют лишь отдельные компоненты этой задачи. Часть систем обеспечивает клонирование голоса и/или визуального образа, но внутри содержат базовые instruct-LLM (инструктивно настроенные большие языковые модели), ориентированные на выполнение стандартных запросов, без функции копирования и воспроизведения личности конкретного человека. Другие подходы позволяют добавлять фактическую информацию о человеке (например, через документы и retrieval-механизмы), однако не обеспечивают одновременное воспроизведение индивидуального стиля речи, голоса и визуального представления в едином автоматизированном пайплайне. Кроме того, многие пользователи не готовы передавать личную информацию сторонним сервисам.

Основная часть

Работа направлена на разработку самоформируемого локального диалогового виртуального двойника человека — мультимодальной системы, формируемой по аудио-/видеодиалогам и визуальным данным личности и объединяющей ASR (автоматическое распознавание речи) [1, 2], LLM (большую языковую модель) [3, 4] с RAG (извлечением релевантного контекста), TTS (синтез речи) [5] и Avatar Generator (генератор аватара). Архитектура системы разделена на два этапа: обучение (формирование профиля) и инференс.

На этапе обучения система автоматически выполняет предобработку аудио- и видеоматериалов, транскрибацию, диаризацию и постобработку транскрипций, формирование диалоговых пар «реплика — ответ» и retrieval-корпуса, инструктивную настройку и/или дообучение LLM по корпусу диалогов, выделение сегментов речи копируемого человека с последующей фильтрацией и ранжированием на основе прокси-метрики качества, подготовку артефактов для zero-shot синтеза речи, а также подготовку персонализированного аватара по визуальным данным профиля.

На этапе инференса реализуется диалог в реальном времени: запись реплики пользователя, распознавание речи, поиск релевантных фрагментов в RAG-корпусе, генерация ответа LLM, синтез ответа голосом клона и визуализация ответа сгенерированным аватаром. Для поддержки контекстности диалога сохраняется состояние взаимодействия между сессиями. Все данные, модели, артефакты и память диалога хранятся и обрабатываются локально, без передачи во внешние сервисы.

Выводы

Разрабатываемое решение ориентировано на задачу цифрового сохранения личности и направлено на воспроизведение стиля общения, голоса и визуального образа человека при локальной обработке данных без передачи во внешние сервисы.

Оценка результата предполагается по сходству синтезированной речи с голосом оригинала, по стилевому и содержательному сходству ответов LLM с реальными ответами человека, а также по задержке ответа и ресурсоемкости локального запуска.

Практическое применение решения возможно в семейных, образовательных, культурных и мемориальных сценариях, где требуется интерактивное цифровое представление человека.

Литература

1. Bain M., Huh J., Han T., Zisserman A. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/2303.00747 (дата обращения: 01.02.2025).
2. Bredin H., Yin R., Coria J. M., Gelly G., Korshunov P., Lavechin M., Fustes D., Titeux H., Bouaziz W., Gill M.-P. pyannote.audio: neural building blocks for speaker diarization [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/1911.01255 (дата обращения: 01.02.2025).
3. Yang A., Li A., Yang B. et al. Qwen3 Technical Report [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/2505.09388 (дата обращения: 01.02.2025).
4. Lin J., Tang J., Tang H., Yang S., Chen W.-M., Wang W.-C., Xiao G., Dang X., Gan C., Han S. AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/2306.00978 (дата обращения: 01.02.2025).
5. Casanova E., Davis K., Gölge E., Göknar G., Gulea I., Hart L., Aljafari A., Meyer J., Morais R., Olayemi S., Weber J. XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/2406.04904 (дата обращения: 01.02.2025).