ՀԱՅԱՍՏԱՆԻ ՀԱՆՐԱՊԵՏՈՒԹՅԱՆ ԿՐԹՈՒԹՅԱՆ, ԳԻՏՈՒԹՅԱՆ,
ՄՇԱԿՈՒՅԹԻ ԵՎ ՍՊՈՐՏԻ ՆԱԽԱՐԱՐՈՒԹՅՈՒՆ

ՀԱՅԱՍՏԱՆԻ ԱԶԳԱՅԻՆ ՊՈԼԻՏԵԽՆԻԿԱԿԱՆ ՀԱՄԱԼՍԱՐԱՆ

Նիկողոսյան Կարեն Հրահատի

ՀԱՅՈՑ ԼԵԶՎՈՎ ԳՐԱՎՈՐ ՏԵՔՍՏԸ ԲԱՆԱՎՈՐ ԽՈՍՔԻ ՎԵՐԱՓՈԽՄԱՆ ԱՎՏՈՄԱՏԱՑՎԱԾ ՀԱՄԱԿԱՐԳԻ ՄՇԱԿՈՒՄԸ

Ե.13.02 «Ավտոմատացման համակարգեր» մասնագիտությամբ
տեխնիկական գիտությունների թեկնածուի գիտական աստիճանի
հայցման ատենախոսության

ՍԵՂՄԱԳԻՐ

Երևան 2025

МИНИСТЕРСТВО ОБРАЗОВАНИЯ, НАУКИ, КУЛЬТУРЫ И СПОРТА
РЕСПУБЛИКИ АРМЕНИЯ

НАЦИОНАЛЬНЫЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ АРМЕНИИ

Никогосян Карен Грагатович

РАЗРАБОТКА АВТОМАТИЗИРОВАННОЙ СИСТЕМЫ ПРЕОБРАЗОВАНИЯ ПИСЬМЕННОГО ТЕКСТА НА АРМЯНСКОМ ЯЗЫКЕ В УСТНУЮ РЕЧЬ

АВТОРЕФЕРАТ

диссертации на соискание ученой степени кандидата
технических наук по специальности 05.13.02-
“Системы автоматизации”

Ереван 2025

Ատենախոսության թեման հաստատվել է Հայաստանի ազգային պոլիտեխնիկական համալսարանում (ՀԱՊՀ):

Գիտական ղեկավար՝ տ.գ.դ. Սարգիս Հովհաննեսի Սիմոնյան

Պաշտոնական ընդդիմախոսներ՝ տ.գ.դ. Արմինե Գևորգի Ավետիսյան
ֆ-մ.գ.թ. Համլետ Ցոլակի Հակոբյան

Առաջատար կազմակերպություն՝ «Երևանի կապի միջոցների գիտահետազոտական ինստիտուտ» ՓԲԸ

Ատենախոսության պաշտպանությունը կայանալու է 2025թ. հուլիսի 14-ին, ժամը 12^{00}-ին, ՀԱՊՀ-ում գործող «Կառավարման և ավտոմատացման» 032 մասնագիտական խորհրդի նիստում (հասցեն՝ 0009, Երևան, Տերյան փ., 105, 17 մասնաշենք):

Ատենախոսությանը կարելի է ծանոթանալ ՀԱՊՀ–ի գրադարանում:

Սեղմագիրն առաքված է 2025թ. հունիսի 12-ին:

032 Մասնագիտական խորհրդի գիտական քարտուղար, տ.գ.թ. Անուշ Վազգենի Մելիքյան

Тема диссертации утверждена в Национальном политехническом университете Армении (НПУА)

Научный руководитель: д.т.н. Саргис Оганесович Симонян

Официальные оппоненты: д.т.н. Армине Геворговна Аветисян
к.ф-м.н. Гамлет Цолакович Акопян

Ведущая организация: ЗАО "Ереванский научно-исследовательский институт средств связи"

Защита диссертации состоится 14-го июля 2025г. в 12^{00} ч. на заседании Специализированного совета 032 – “Управления и автоматизац”, действующего при Национальном политехническом университете Армении, по адресу: 0009, г. Ереван, ул. Теряна, 105, корпус 17.

С диссертацией можно ознакомиться в библиотеке НПУА.

Автореферат разослан 12-го июня 2025 г.

Ученый секретарь
Специализированного совета 032 к.т.н. Ануш Вазгеновна Меликян

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. В настоящее время системы преобразования текста в речь (ПТР) играют важную роль в обеспечении доступности информации и развитии человеко-машинного взаимодействия. Эти системы широко применяются в различных сферах, включая образование, библиотечное дело, государственные услуги, СМИ и обслуживание клиентов. Разработка таких систем для армянского языка позволит существенно улучшить доступность образовательных материалов для людей с нарушениями зрения, автоматизировать создание аудиокниг, обеспечить голосовой доступ к электронным услугам и создать полностью автоматизированных агентов искусственного интеллекта (ИИ) с естественной армянской речью.

Существующие решения для армянского языка ограничены, имеют коммерческий характер и недоступны для интеграции в автоматизированные системы. Особенно важно отметить отсутствие свободно доступных решений, обеспечивающих точный фонетический анализ армянского текста через графемно-фонемное преобразование (ГФП). Система ГФП является фундаментальным компонентом синтеза речи, отвечающим за правильное произношение текста. Уникальная фонетическая система армянского языка, включая правила ударения и взаимодействие звуков, требует специального подхода к разработке системы ГФП.

Разработка систем синтеза речи для встроенных устройств особенно актуальна в цифровую эпоху. Ограниченные ресурсы таких устройств требуют специальных методов оптимизации для уменьшения размера модели при соблюдении строгих требований безопасности и энергоэффективности. Для развития ПТР систем на армянском языке ключевое значение имеет создание качественного набора голосовых данных. В настоящее время не существует свободно доступного, достаточно большого, мультиголосового и высококачественного набора голосовых данных для армянского языка. Автоматизация создания такого набора данных является важной задачей, поскольку традиционный сбор данных требует значительных временных и человеческих ресурсов.

Разработка мультиголосовых ПТР-моделей представляет особое направление для армянского языка. Такие модели позволяют с помощью одной системы синтезировать речь разными голосами, сохраняя индивидуальные голосовые характеристики каждого диктора.

Интеграция в ПТР-систему модели оптического распознавания символов (ОРС) значительно расширит ее применимость, позволяя работать не только с цифровыми текстами, но и со сканированными документами, книгами и фотографированными материалами.

Объект исследования. Разработка автоматизированной системы преобразования письменного текста на армянском языке в устную речь с учетом фонетических особенностей армянского языка, специально предназначенной для использования во встроенных устройствах.

Цель работы. Разработка свободно доступных мультиголосовых и мультимодальных ПТР-систем, которые могут эффективно работать во встроенных устройствах, позволяя создавать высококачественные аудиофайлы.

Методы исследования. В ходе диссертации были использованы современные архитектуры преобразования текста в речь, в частности архитектура УАСПР (условный вариационный автоэнкодер с состязательным обучением для преобразования текста в речь), модель Conformer-CTC для преобразования графем в фонемы, методы квантизации и оптимизации моделей через формат ONNX, методы автоматизированной сегментации голосовых данных (с расчетом уровня сигнала по шкале dBFS и динамическим пороговым методом), системы ОРС для армянского языка и программные пакеты, основанные на технологиях React/Flask.

Научная новизна:

- Разработана модель ГФП с архитектурой Conformer-CTC для армянского языка, обеспечивающая высокую точность - 16.13% КСО (коэффициент словесной

ошибки) и 17.36% КФО (коэффициент фонемной ошибки), что значительно превосходит результаты существующего инструмента Phonemizer (96.60% КСО и 36.15% КФО).

- Разработана автоматизированная система сбора голосовых данных, реализующая расчет уровня сигнала по шкале dBFS и динамический пороговый метод для обнаружения оптимальных точек разделения (ОТР). Система значительно ускоряет сбор данных без требования финансовых ресурсов.
- Созданы высококачественные наборы данных - набор голосовых данных (21 час записи, 14,182 аудиофайла) и набор данных ГФП (17,862 пары слово-фонема).
- Разработана мультиголосовая ПТР-модель с архитектурой УАСПР для армянского языка, демонстрирующая превосходные результаты: средний коэффициент идентичности диктора - 0.9366, F0-корреляция - 0.8834, среднее значение ПОКР (перцептивная оценка качества речи) - 2.87 и КОВ (кратковременная объективная внятность) - 0.7248.
- Оптимизация ПТР модели с помощью формата ONNX и квантизации привела к уменьшению размера модели на 87% (с 1.0 Гб до 127.5 Мб), сокращению времени инференса примерно на 85% и обеспечению значений КРВ (коэффициент реального времени) в диапазоне 0.22…0.30, позволяя работать в 3…4 раза быстрее реального времени.
- Разработан новый метод интеграции с системой ОРС Tesseract, где использована модель, обогащенная автоматически сгенерированными синтетическими обучающими данными. Предложенный подход позволил сократить коэффициент ошибок распознавания на 56% как на словесном, так и на символьном уровнях.

Практическая ценность работы. Разработанная в диссертации система ПТР для армянского языка реализована как для серверного, так и для встроенного применения. Система обеспечивает высококачественный синтез речи с коэффициентом идентичности диктора 0.9366 и показателем КОВ 0.7248. Оптимизация для встроенных устройств с помощью формата ONNX и квантизации позволила уменьшить размер модели на 87% (с 1.0 Гб до 127.5 Мб), сократить время инференса на 85% и обеспечить значения КРВ-0.22…0.30, позволяя работать в 3…4 раза быстрее реального времени. Низкое потребление памяти оптимизированными моделями (всего несколько мегабайт для текстов средней длины) позволяет разместить систему на устройствах с ограниченными ресурсами, обеспечивая стабильную работу (стандартное отклонение значений КРВ-0.012…0.056). Интеграция компонента ОРС, который благодаря синтетическим обучающим данным сократил коэффициент ошибок распознавания на 56%, значительно расширяет применимость системы. Графический пользовательский интерфейс, основанный на технологиях React/Flask, делает систему доступной как для специалистов, так и для обычных пользователей.

На защиту выносятся следующие научные положения:

- Мультиголосовая ПТР ИИ-модель, учитывающая фонетические особенности армянского языка.
- Система ГФП, разработанная для армянского языка.
- Методы и инструментарий создания и обработки набора голосовых данных, а также собранные с их помощью высококачественные наборы данных.
- Мультиголосовые ПТР-модели, оптимизированные для встроенных устройств.
- Решения по интеграции с системой ОРС.
- Инструментарий для пользователя, основанный на современных ВЕБ-технологиях.

Достоверность научных положений. Достоверность научных положений подтверждается результатами программной реализации моделей ГФП и ПТР, представленных в диссертации, комплексными экспериментальными исследованиями и

оценками, полученными с помощью международно признанных метрик, а также математическими обоснованиями.

Внедрение. Результаты диссертации, в частности, созданные наборы данных, математические модели, модели искусственного интеллекта, методы квантования и оптимизации, автоматизированные системы создания наборов данных внедрены в теоретические и практические занятия курсов «Обработка естественного языка» и «Методы логического анализа естественного языка» кафедры «Информационные технологии и автоматизация» НПУА.

Апробация работы. Основные научные и практические результаты диссертации докладывались на:

- Международном симпозиуме “IEEE East-West Design & Test Symposium (EWDTS)” (Ереван, Армения, 2024 г.);
- научных семинарах кафедры “Информационные технологии и автоматизация” НПУА (Ереван, Армения, 2022–2024 гг.).

Публикации. Основные положения, представленные в диссертации, обобщены в девяти научных статьях, две из которых без соавторов, а одна - в научной базе данных “SCOPUS” (“СКОПУС”). Список статей приведен в конце автореферата.

Структура и объём диссертации. Работа состоит из введения, четырёх глав, заключения, списка литературы, включающего 110 наименования, и двух приложений. В первом приложении представлен акт внедрения, во втором - детально представлены параметры архитектуры УАСПР. Основной объём диссертации составляет 145 страниц, а вместе с приложениями - 150 страницы. Диссертация написана на армянском языке.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность разработки систем преобразования текста в речь для армянского языка, сформулированы цель и основные задачи исследования, описаны используемые методы, представлены научная новизна разработанных моделей, их практическая ценность и основные научные положения, выносимые на защиту.

В первой главе представлены общие положения разработки автоматизированных систем преобразования письменного текста в устную речь. Обоснована необходимость разработки системы ПТР для армянского языка. Отмечается, что за последнее десятилетие благодаря развитию методов глубокого обучения и вычислительных ресурсов технологии синтеза речи достигли значительного прогресса. Современные ПТР системы способны генерировать почти естественно звучащую речь, что привело к их широкому применению в различных областях. Однако разработка таких систем для армянского языка остаётся сложной задачей не только из-за фонетических особенностей языка, но и из-за отсутствия высококачественных наборов голосовых данных.

Представлена общая структура ПТР системы, состоящая из трёх основных компонентов: кодировщик речи, преобразователь голоса и вокодер (рис. 1). Процесс преобразования текста в речь математически сформулирован как задача последовательного отображения, где входная последовательность представлена в виде дискретных символов, а выходная - в виде временного ряда голосовых характеристик.

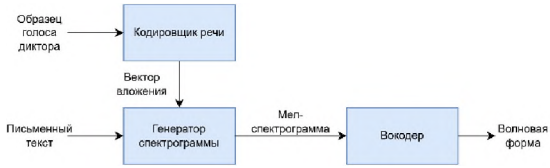

Рис. 1. Основные компоненты ПТР системы

Проведен системный анализ архитектур глубокого обучения, применяемых в ПТР системах. Рассмотрены архитектуры последовательность-в-последовательность (seq2seq) и архитектуры сквозного обучения (end-to-end). Выполнен сравнительный анализ современных архитектур преобразования текста в речь, включая Tacotron-2, DeepVoice-3, FastSpeech и FastSpeech-2.

Таблица 1

Сравнение эффективности различных ПТР-моделей с использованием набора данных LJSpeech

Модель	MOS	Тип архитектуры	Выходные данные
Tacotron-2	4.5±0.06	Авторегрессивная	Акустические характеристики
DeepVoice-3	3.44±0.32	Авторегрессивная	Акустические характеристики
FastSpeech	3.84±0.08	Неавторегрессивная	Акустические характеристики
FastSpeech-2	3.83±0.08	Неавторегрессивная	Акустические характеристики

Результаты тестирования на наборе данных LJSpeech показали, что Tacotron-2 превосходит другие модели с показателем MOS 4.5±0.06 (табл. 1).

Проанализированы современные подходы к ГФП, которые классифицированы на три основные группы: основанные на правилах, основанные на машинном обучении и гибридные. Особое внимание уделено фонетическим особенностям армянского языка, включая трехрядную систему согласных (глухие, звонкие и придыхательные глухие) и позиционные изменения гласных «և» и «ո».

Проведена оценка эффективности существующей для армянского языка ГФП-системы - библиотеки Phonemizer. Тестирование на 500 словесных единицах показало крайне низкую точность - только в 17 случаях (3.40%) система выполнила точное преобразование. КФО составил 36.15%, а КСО - 96.60%. Представлены примеры преобразований, демонстрирующие типичные ошибки системы (табл. 2).

Таблица 2

Примеры преобразования ГФП-системы

Слово	Ожидаемое	Полученное	КФО	КФО (%)	КСО	КСО (%)
կիսատ	kiˈsat	kisat	0.3333	33.33	1	100.00
սեպտեմբեր	septemˈber	september	0.2	20.00	1	100.00
ատոմային	atomaˈjin	atomaɪin	0.4444	44.44	1	100.00
քիթ	kʰitʰ	kʰɨtʰ	0	0.00	0	0.00
լրել	ləˈrel	lərel	0.1667	16.67	1	100.00
կրակահերթ	kərakaˈhertʰ	kəɹakahertʰ	0.4167	41.67	1	100.00

Проанализированы существующие наборы голосовых данных для армянского языка, включая Mozilla Common Voice и проект ArmSpeech. Отмечено, что набор Mozilla Common Voice имеет ряд качественных проблем, так как значительная часть записей выполнена непрофессиональным оборудованием и в неконтролируемой среде. Проект ArmSpeech предоставляет более качественные данные, но ограничен по объему.

На основе проведенного анализа сформулированы основные проблемы существующих ПТР-систем для армянского языка. Выявлено, что для успешной разработки многоголосовой ПТР-системы необходимо решить следующие задачи: выбор и адаптация архитектуры, соответствующей особенностям армянского языка; создание высококачественного набора голосовых данных; разработка точной ГФП модели; оптимизация модели для встроенных устройств.

Целью исследования является разработка полной многоголосовой системы преобразования текста в речь для армянского языка, которая будет включать модель ГФП, учитывающую фонетические особенности армянского языка, и будет специально предназначена для использования во встроенных устройствах.

Во второй главе представлена разработка наборов данных, необходимых для создания автоматизированной системы преобразования текста в речь на армянском языке. Исследованы методы сбора и обработки речевых данных, а также создания набора данных соответствий между графемами и фонемами.

Сбор, обработка и анализ набора речевых данных

Для решения проблемы отсутствия качественных речевых данных разработана автоматизированная система сегментации аудиофайлов. Система базируется на модульной архитектуре и включает механизмы интеллектуальной сегментации, сопоставления текста и автоматической оценки качества. Процесс сегментации основан на многоэтапном алгоритме обнаружения пауз, использующем уровень сигнала в шкале dBFS, который рассчитывается по формуле

$$\mathrm{dBFS}(w) = 20\log_{10}\left(\frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}}{A_{max}}\right), \quad (1)$$

где N - количество отсчетов в окне; x_i - отдельные отсчеты; A_{max} - максимально возможная амплитуда в цифровой системе.

ОТР определяется как центральная точка самого длинного участка тишины:

$$t_{\mathrm{OTP}} = \frac{t_s + t_e}{2}, \quad (2)$$

где t_s и t_e — начало и конец самого длинного тихого участка.

С использованием разработанной системы создан речевой корпус, основные количественные характеристики которого приведены в табл. 3. Корпус содержит 75,597.79 секунд (около 21 часа) аудиозаписей, 14,182 аудиофайла и 14,078 уникальных предложений.

Таблица 3

Основные количественные характеристики речевого корпуса

Критерий	Значение
Общая длительность (с)	75597.79
Минимальная длительность аудиофайла (с)	1.54
Максимальная длительность аудиофайла (с)	16.47
Средняя длительность (с)	5.33
Общее количество	14182
Количество уникальных предложений	14078
Общее количество символов	869097
Минимальное количество букв в образцах	12
Максимальное количество символов в образцах	347
Среднее количество символов в образце	61.28
Общее количество слов	137716
Уникальные слова	30466
Минимальное количество слов в образцах	4
Максимальное количество слов в образцах	56
Среднее количество слов в образце	9.71
Количество дикторов	2

Большинство файлов имеют продолжительность от 2 до 6 секунд, что соответствует естественной длительности произношения одного предложения. Корпус включает записи двух дикторов мужского пола. Анализ фонетического состава корпуса подтвердил равномерное представление всех фонем армянского языка в соответствии с их естественным распределением в языке.

Сбор, обработка и анализ набора данных для преобразования графем в фонемы

Для решения задачи ГФП разработана система автоматизированного сбора данных, использующая Викисловарь в качестве основного источника. Система применяет

многослойный анализ структуры словарных статей для извлечения фонетической транскрипции в формате МФА (Международный фонетический алфавит).

Созданный набор данных содержит 17,862 пары слово-фонема и обеспечивает полное представление фонетической системы армянского языка. Анализ соответствия между графемами и фонемами выявил сложные взаимосвязи, характерные для армянского языка. Некоторые графемы демонстрируют устойчивое соответствие конкретным фонемам, в то время как другие, например «u» или «n», могут иметь различные фонетические реализации в зависимости от позиции и контекста.

Анализ частотности фонем показал соответствие закону Зипфа, что подтверждает естественность собранных данных и может быть математически выражено формулой:

$$P(r) = \frac{A}{r^{\alpha}}, \tag{3}$$

где P(r) - частота фонемы с рангом r; r - ранг фонемы по частоте; α - близко к 1; а A - нормализующая константа.

Распределение длины слов в наборе данных показало, что преобладают слова, содержащие 6…8 символов, что соответствует естественной структуре словаря армянского языка. Средняя длина слова составляет 7.62 символа.

Созданные наборы данных являются первыми, большими общедоступными ресурсами для разработки систем преобразования текста в речь на армянском языке и представляют собой ценную основу для дальнейших исследований в области обработки естественного языка.

В третьей главе представлено детальное описание разработанной автоматизированной системы преобразования письменного армянского текста в устную речь. Рассматриваются математические модели, архитектурные решения и процессы обучения компонентов системы, а также оптимизация моделей и создание программного комплекса для практического применения.

Разработка математической модели ИИ и архитектуры для преобразования графем в фонемы

Для решения задачи преобразования армянских графем в фонемы разработана модель Conformer-CTC, объединяющая преимущества сверточных нейронных сетей и трансформеров. Данная модель обеспечивает высокоточное отображение входной графемной последовательности $x = (x_1, \ldots, x_n)$ на соответствующую фонемную последовательность $y = (y_1, \ldots, y_m)$, что выражается вероятностным представлением:

$$p(y|x) = \sum_{\pi \in B^{-1}(y)} p(\pi|x), \tag{4}$$

где $B^{-1}(y)$ представляет множество всех возможных путей CTC (Connectionist Temporal Classification), соответствующих фонемной последовательности y.

Архитектура модели, представленная на рис. 2, включает слой вложений, преобразующий входной текст в 300-мерное векторное пространство, и последовательность блоков Conformer, каждый из которых содержит четыре основных компонента: модуль многоголовного самовнимания, сверточный модуль, линейную сеть и слой нормализации.

Ключевым элементом архитектуры является механизм относительного позиционного самовнимания, описываемый формулой

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + R\right)V, \tag{5}$$

где R - матрица относительного позиционного кодирования, а d_k - размерность механизма внимания. Использование 4 голов внимания позволяет параллельно обрабатывать различные лингвистические характеристики текста.

Функция потерь CTC, применяемая для обучения модели, позволяет осуществлять обучение без прямой разметки соответствия графемы-фонемы

$$\mathcal{L}_{CTC} = -\log p(y|x) = -\log \sum_{\pi \in B^{-1}(y)} \prod_{t=1}^{T} p(\pi_t|x), \quad (6)$$

где T - количество временных шагов, а $p(\pi_t|x)$ - вероятность конкретной фонемы на данном временном шаге.

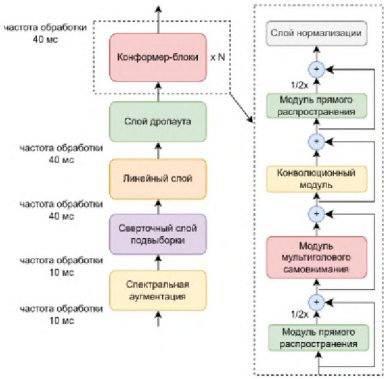

Рис. 2. Архитектура модели G2P-Conformer-CTC, включающая общую структуру модели и внутреннюю структуру блока Conformer

Обучение модели проводилось на специально подготовленном наборе данных, включающем 17,862 армянских пары слово-фонема. Использовался алгоритм оптимизации AdamW с параметрами β_1=0.9 и β_2=0.98, а скорость обучения регулировалась планировщиком Noam. Динамика обучения (рис. 3) демонстрирует стабильное снижение функции потерь как на обучающем, так и на валидационном наборах данных, достигая значений 0.4 и 0.55 соответственно после примерно 1000 шагов обучения.

Рис. 3. Динамика потерь обучения и валидации в процессе обучения

На валидационном наборе модель достигает уровня ошибки на уровне слов в 16.13% (рис. 4а), что соответствует точности предсказания 83.87% слов. Средний уровень ошибки на уровне фонем составляет 17.36% (рис. 4б), что является высоким показателем для армянского языка с его сложной системой фонетических изменений.

Спектральный анализ результатов показывает высокую эффективность модели при работе с простыми соответствиями графемы-фонемы, регулярными фонетическими изменениями и распространенными лексическими единицами. Вместе с тем отмечаются трудности при обработке редких слов, сложных фонетических изменений и отдельных случаев заимствованных слов.

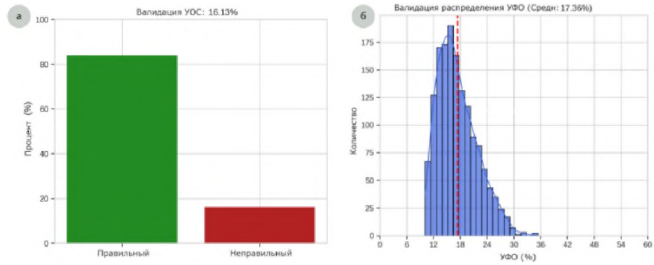

Рис. 4. Распределение КСО(а) и КФО(б) на валидационном наборе

Разработка математической модели УАСПР для преобразования письменного текста в устную речь

Для синтеза армянской речи высокого качества разработана и адаптирована модель условного вариационного автоэнкодера с состязательным обучением. Сложная фонетическая структура и богатая просодическая система армянского языка потребовали создания гибкой и мощной модели, способной сохранять все существенные фонетические и интонационные особенности речи.

Математическая основа модели УАСПР представлена условным вероятностным распределением:

$$p_\theta(y, o, z|x) = p_\theta(y|z, o) \cdot p_\theta(z|x) \cdot p_\theta(o|x, z), \tag{7}$$

где z - латентная переменная, кодирующая просодические характеристики речи; o - вектор длительностей фонем; θ - набор параметров модели; y - аудиосигнал, а x - фонемная последовательность. Данная вероятностная модель позволяет разделить три основных компонента генерации речи:

- моделирование длительности фонем: $p_\theta(o|x, z)$;
- моделирование просодии: $p_\theta(z|x)$;
- генерация волновой формы: $p_\theta(y|z, o)$.

Общая архитектура модели УАСПР и её функциональная схема представлены на рис. 5, где отображены процедуры обучения и инференса.

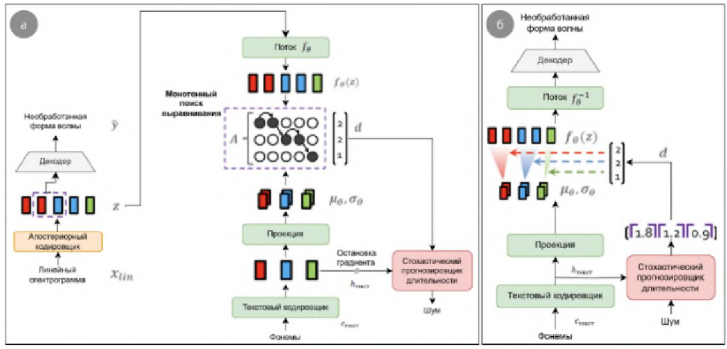

Рис. 5. Архитектура модели УАСПР, показывающая: а-процедуру обучения и б-процедуру инференса

Вариационный подход модели опирается на нижнюю границу доказательства (Evidence Lower Bound)

$$\log p(x|c) \geq E_{q(z|x,c)}[\log p(x|z, c)] - D_{KL}(q(z|x, c)|p(z|c)), \tag{8}$$

где c - входная фонемная последовательность; z - вектор латентного пространства; x - выходные речевые данные, а D_{KL} – расхождение Кульбака-Лейблера.

Модель прогнозирования длительности фонем, критически важная для армянского языка, основана на логнормальном распределении:

$$\mathrm{p}(o|x,z) = \prod_{i=1}^{N} \mathrm{r}LogNormal\left(o_i; \mu_i, \sigma_i^2\right), \tag{9}$$

где параметры μ_i и σ_i^2 оцениваются специальной сверточной нейронной сетью.

Для моделирования просодической вариативности применяется сложная структура потокового декодера с нормализующими потоками. Для каждого k-го потокового преобразования вычисляется логарифмическая вероятность:

$$\log p\left(z_k\right) = \log p\left(z_{k-1}\right) + \log\left|\det\left(\frac{\partial f_k}{\partial z_{k-1}}\right)\right|, \tag{10}$$

где f_k - потоковое преобразование, а второе слагаемое - логарифм определителя якобиана.

Задний кодировщик формирует условное распределение латентных переменных с помощью гауссова распределения:

$$q_\phi(z|y) = \mathcal{N}\left(z; \mu_\phi(y), \sigma_\phi^2(y)\right), \tag{11}$$

где ϕ - параметры заднего кодировщика, а $\mu_\phi(y)$ и $\sigma_\phi^2(y)$ - среднее значение и дисперсия.

Волновой декодер преобразует мел-спектрограмму в аудиосигнал, используя иерархическое генеративное состязательное обучение:

$$\min_G \max_D V(G,D) = \mathbb{E}_{y\sim p_{data}}[\log D(y)] + \mathbb{E}_{x\sim p_{data}}\left[\log\left(1 - D(G(x))\right)\right], \tag{12}$$

где G - генератор (волновой декодер), а D - дискриминатор, различающий реальную и синтезированную речь.

Все компоненты регулируются общей функцией потерь:

$$\mathcal{L}_{total} = \lambda_{mel}\mathcal{L}_{mel} + \lambda_{kl}\mathcal{L}_{kl} + \lambda_{dur}\mathcal{L}_{dur} + \lambda_{adv}\sum_{p\in P}\mathcal{L}_{adv}^{p} + \lambda_{fm}\mathcal{L}_{fm}, \tag{13}$$

Анализ процесса обучения продемонстрировал стабильное снижение и последующую стабилизацию всех компонентов функции потерь. Потери мел-спектрограммы (рис. 6) стабилизируются на уровне 15.0467 для обучающего набора и 15.6338 для валидационного набора. Расхождение Кульбака-Лейблера стабилизируется в диапазоне 1.5…2.0. Потери дискриминатора достигают стабильного состояния около 0.45 для обучающего набора и 0.42 для валидационного набора.

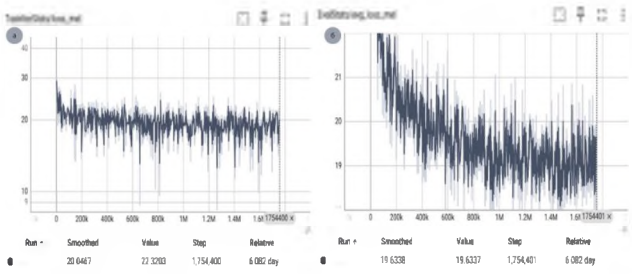

Рис. 6. Динамика потерь мел-спектрограммы. а-потери обучающей фазы, б-потери оценочной фазы

Сравнительный анализ мел-спектрограмм (рис. 7) и волновых форм (рис. 8) оригинальной и синтезированной речи подтверждает высокую точность воспроизведения как основной частотной структуры, так и просодических нюансов армянской речи.

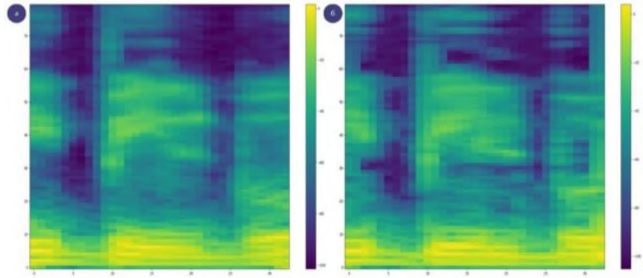

Рис. 7. Мел-спектрограмма: а-эталонной речи, б-синтезированной речи

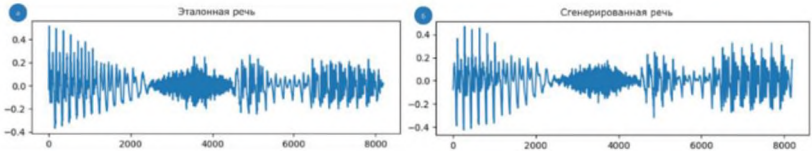

Рис. 8. Сравнение звуковых волновых форм: а-эталонной речи, б-синтезированной речи

Описание оптимизации и квантизации модели ИИ для ПТР

Для повышения эффективности использования разработанной модели синтеза речи на устройствах с ограниченными вычислительными ресурсами проведены оптимизация и квантизация модели. Ключевым этапом оптимизации стало преобразование модели в формат ONNX (Open Neural Network Exchange), обеспечивающий кроссплатформенную совместимость между различными фреймворками машинного обучения.

Формат ONNX оптимизирует структуру модели, удаляя компоненты, необходимые при обучении, но избыточные при последующем использовании. Математически модель машинного обучения представляется как функция $f(x;\theta)$, где x - входные данные, а θ - параметры модели. Эта функция состоит из последовательности операций, представленных в виде вычислительного графа G, определяющего порядок обработки данных.

При преобразовании в формат ONNX каждая операция o_i исходной модели отображается на эквивалентную операцию o_i' в формате ONNX:

$$o_i(T_{in}) \rightarrow o_i'(T_{in}). \tag{14}$$

К полученному графу применяются оптимизации: слияние операций, устранение мертвого кода и константные свертки.

После преобразования в формат ONNX проведена дополнительная квантизация, заменяющая числа с плавающей точкой высокой точности (32-бит) целыми числами меньшей точности (8-бит). Математически процесс квантизации описывается формулой

$$Q(x) = \text{round}\left(\frac{x}{S}\right) + Z, \tag{15}$$

где x - исходное значение с плавающей точкой; S - масштабный коэффициент; Z - нулевая точка, а $Q(x)$ - квантованное целое число.

Процесс квантизации модели ONNX состоял из двух основных этапов: подготовки и собственно квантизации. Для квантизации выбраны основные операторы: MatMul, Gemm,

Attention, LSTM и Gather, содержащие большое количество параметров и ответственные за значительную часть вычислительных ресурсов.

В результате многоэтапной оптимизации размер модели значительно уменьшен - с 1,0 Гб до 127,5 Мб, что составляет сокращение на 87,7% (табл. 4) без существенной потери качества синтезированной речи.

Таблица 4

Изменения размера модели ПТР на различных этапах оптимизации

Модель	Объем памяти	Уменьшение от оригинала
Оригинал	1,0 Гб	-
ONNX	134,9 Мб	87,0%
Предобработанная	127,6 Мб	87,7%
Квантованная	127,5 Мб	87,7%

Разработка и использование программного инструмента автоматизированной системы ПТР

Для практического применения разработанной системы синтеза армянской речи создано web-приложение, позволяющее пользователям легко преобразовывать письменный армянский текст в естественную речь. Система построена по трехкомпонентному принципу с тремя основными вкладками, обеспечивающими различные функциональные возможности.

Web-приложение включает три основные вкладки: "Text Input", "Image to Text" и "History". Вкладка "Text Input" служит для непосредственного ввода армянского текста и его преобразования в аудиофайл (рис. 9). Здесь пользователь может выбрать диктора, ввести текст, запустить генерацию и прослушать или скачать результат.

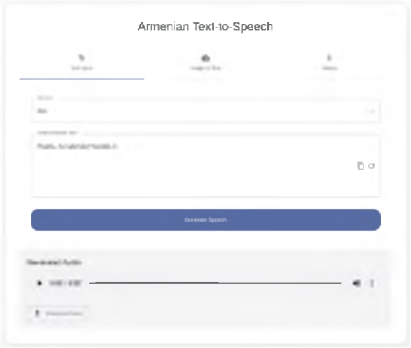

Рис. 9. Вид основной вкладки web-сайта ПТР в процессе использования

Вкладка "Image to Text" позволяет загружать изображения с армянским текстом, который автоматически распознается OPC и направляется на синтез речи (рис. 10а). Для распознавания текста используется специально адаптированная модель Tesseract, оптимизированная для армянского языка.

Вкладка "History" хранит историю предыдущих сессий пользователя, что особенно полезно при работе с повторяющимися текстами (рис. 10б). Каждая запись включает дату и время генерации, фрагмент сгенерированного текста и имя выбранного диктора, а также кнопки для просмотра, воспроизведения и скачивания.

Внутренняя структура web-приложения основана на современных технологиях и компонентной архитектуре с четким разделением на фронтенд и бэкенд части, взаимодействующие через REST API. Фронтенд разработан с использованием библиотеки React, а серверная часть реализована на языке Python с применением фреймворка Flask.

Основные API-запросы системы включают:

- /api/synthesize - для генерации аудиофайла из текста,
- /api/ocr - для распознавания текста из изображения,
- /api/history - для получения и управления записями истории.

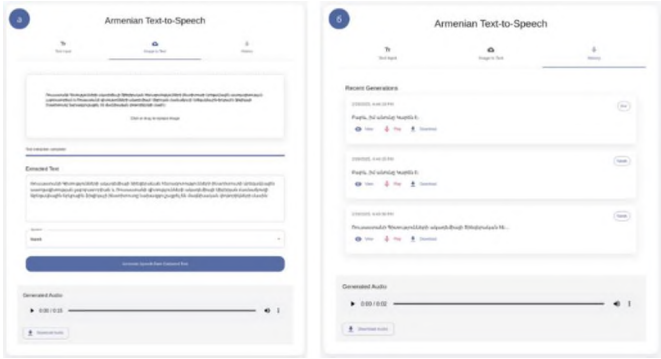

Рис. 10. Вид вкладки ПТР с изображением (а) и вкладка "History" (б) web-сайта в процессе работы

В четвертой главе представлены результаты экспериментальных исследований разработанной автоматизированной ПТР системы для армянского языка и дано обоснование её эффективности. Рассмотрены вопросы оценки качества синтезированной речи с использованием объективных метрик, а также проведен анализ производительности оптимизированных версий модели.

Анализ результатов оценки мультидикторской ПТР ИИ модели с архитектурой УАСПР для армянского языка

Оценка систем ПТР модели имеет существенное значение для определения качества и эффективности этих систем. В случае мультидикторских моделей ПТР важно оценивать не только общее качество синтезированной речи, но и степень сохранения идентичности диктора.

Мультидикторская модель ПТР для армянского языка на архитектуре УАСПР была оценена с помощью объективных метрик. Использовалась коллекция из 500 выражений от двух дикторов (Гор и Нарек). Оценка включала сбор оригинальных записей и генерацию синтезированной речи.

Для оценки системы были использованы пять основных метрик: коэффициент сходства диктора (КСД), ПОКР, F0-среднеквадратическая ошибка (F0-СКО), корреляция F0 и КОВ (STOI). Эта комбинация метрик позволяет всесторонне оценить работу системы, учитывая качество звука, разборчивость, точность интонации и сохранение идентичности диктора.

Как видно из табл. 5, система демонстрирует высокие результаты по показателям сходства диктора (0.9366) и точности интонации (корреляция F0: 0.8834), одновременно поддерживая хорошие результаты с точки зрения качества восприятия (ПОКР: 2.8700) и разборчивости (КОВ: 0.7248).

Таблица 5

Общая статистическая сводка оценки мультидикторской системы ПТР

Метрика	Среднее	Стандартное отклонение	Минимум	Максимум
Сходство	0.9366	0.0083	0.9308	0.9425
ПОКР	2.8700	0.2213	2.4500	3.1200
F0-СКО	8.4009	1.6057	6.9049	9.8970
F0-корреляция	0.8834	0.0397	0.7733	0.8994
КОВ	0.7248	0.0805	0.6392	0.8264

Для оценки сохранения идентичности диктора был использован метод косинусного сходства между векторами эмбеддинга. На графике (рис. 11) видно, что большинство коэффициентов сходства сосредоточено в диапазоне 0.93…0.95, что свидетельствует о высоком сходстве синтезированной речи с оригинальным диктором.

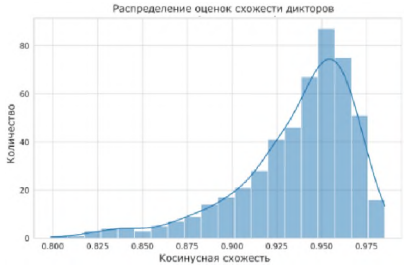

Рис. 11. Распределение коэффициентов сходства диктора

Рассчитанный для модели ИИ показатель КРО (коэффициент равной ошибки) составляет 0.10 (10%), что считается очень хорошим результатом для систем ПТР, особенно для языков с ограниченными ресурсами, таких как армянский язык. Этот показатель свидетельствует о том, что система может с 90%-й точностью различать соответствующие и несоответствующие голоса дикторов.

Для оценки качества восприятия использовалась метрика ПОКР, которая является международным стандартом и измеряет перцептивное качество речи. ПОКР рассчитывается путем сопоставления психоакустических представлений оригинального и синтезированного сигналов. Со средним значением 2.87 этот результат можно считать хорошим показателем для языка с ограниченными ресурсами, такого как армянский.

Для оценки разборчивости синтезированной речи использовалась метрика КОВ, распределение значений которой сосредоточено вокруг среднего значения 0.72. Этот результат считается хорошим для систем ПТР, поскольку значения КОВ выше 0.7 обычно указывают на хорошую разборчивость и соответствуют примерно 80% или более высокой разборчивости речи.

Для оценки точности интонации и фонетической точности используются коэффициенты F0-СКО и корреляция F0. Их средние значения 8.4009 Гц и 0.8834 соответственно свидетельствуют о способности модели точно воспроизводить интонационные характеристики.

Сравнительный анализ дикторов показал, что для диктора Гор получены лучшие результаты с точки зрения точности интонации (F0-СКО: 8.10 Гц и корреляция F0: 0.89) и разборчивости (КОВ: 0.74), в то время как диктор Нарек имеет небольшое преимущество с точки зрения качества восприятия (ПОКР: 2.9).

В целом разработанная мультидикторская модель ПТР для армянского языка с архитектурой УАСПР демонстрирует высокое качество с точки зрения сохранения идентичности диктора, точности интонации, качества восприятия и разборчивости. Полученные результаты близки к международным стандартам, что особенно важно для языка с ограниченными ресурсами, такого как армянский.

Анализ результатов оценки оптимизированной мультидикторской ПТР ИИ модели

Оценка оптимизированных моделей ПТР имеет важное значение как для уточнения их качественных характеристик, так и производительности. В данном разделе представлен сравнительный анализ трех вариантов мультидикторской модели ПТР для армянского языка, проведенный в условиях контролируемого тестирования с целью оценки их производительности в различных прикладных средах.

Оценка проводилась с использованием стратифицированного подхода к отбору образцов, который обеспечивает комплексный охват различных длин текста и типов

дикторов. Входные тексты были классифицированы по длине символов на три категории: короткие (1 < 50 символов), средние (50 ≤ 1 ≤ 150 символов) и длинные (1 > 150 символов). Для каждой категории было выбрано 25 образцов для двух дикторов – "Гор" и "Нарек". Общее количество тестовых данных составило 150 образцов.

В качестве основных метрик оценки были выбраны время вывода, использование памяти, использование центральное процессорное устройство (ЦПУ) и КРВ. Последний имеет особое значение для систем ПТР и рассчитывался по формуле

$$\text{КРВ} = \frac{T_p}{T_a}, \tag{16}$$

где T_p представляет время обработки в секундах, а T_a – продолжительность сгенерированного аудиофайла в секундах.

Оценка проводилась в двух различных аппаратных конфигурациях: многоядерной и одноядерной. Одноядерная конфигурация позволяет оценить базовую производительность модели в условиях ограниченных ресурсов, в то время как многоядерная конфигурация позволяет понять возможности масштабирования модели.

Анализ КРВ (табл. 6) показал, что модель ONNX и квантованная модель продемонстрировали стабильно более высокую производительность, чем реальное время во всех категориях, со значениями КРВ от 0.29 для коротких текстов до 0.22 для длинных текстов. Стандартное отклонение значений КРВ уменьшалось с увеличением длины текста, что свидетельствует о более стабильной производительности при обработке длинного содержания.

Таблица 6

Результаты анализа КРВ в многоядерной системе

Модель	Категория	Диктор	Средн. длина голоса (с)	Средн. КРВ	Станд. отклонение КРВ	Мин. КРВ	Макс. КРВ
УАСПР	Короткая	Гор	3.176	1.672	0.665	0.562	3.482
УАСПР	Короткая	Нарек	4.518	1.333	0.475	0.678	2.301
УАСПР	Средная	Гор	6.644	1.213	0.419	0.576	2.189
УАСПР	Средная	Нарек	8.556	0.996	0.328	0.381	1.608
УАСПР	Длинная	Гор	13.033	0.789	0.119	0.617	1.017
УАСПР	Длинная	Нарек	13.752	0.754	0.103	0.511	0.905
ONNX	Короткая	Гор	2.669	0.296	0.056	0.202	0.396
ONNX	Короткая	Нарек	4.053	0.296	0.034	0.241	0.371
ONNX	Средная	Гор	6.129	0.261	0.021	0.238	0.319
ONNX	Средная	Нарек	7.995	0.247	0.016	0.219	0.284
ONNX	Длинная	Гор	12.468	0.226	0.015	0.193	0.255
ONNX	Длинная	Нарек	13.365	0.226	0.012	0.205	0.252
Квантованная	Короткая	Гор	2.683	0.344	0.051	0.271	0.499
Квантованная	Короткая	Нарек	3.984	0.309	0.031	0.254	0.361
Квантованная	Средная	Гор	6.141	0.284	0.025	0.241	0.337
Квантованная	Средная	Нарек	8.060	0.258	0.018	0.222	0.295
Квантованная	Длинная	Гор	12.471	0.227	0.012	0.203	0.249
Квантованная	Длинная	Нарек	13.381	0.231	0.011	0.214	0.263

Анализ использования памяти показал значительные различия между реализациями. Реализация УАСПР показала значительно более высокое потребление памяти, особенно для текстов средней длины — до 25.99 Мб для диктора «Гор» и 6.07 Мб для диктора «Нарек». Модель ONNX продемонстрировала заметно более эффективное управление памятью с почти нулевым использованием памяти для коротких и средних текстов, достигая лишь 7.5 Мб для длинных текстов с диктором «Гор».

Измерения времени вывода также показали существенные различия между моделями. Модель УАСПР зафиксировала самое высокое время вывода, варьирующееся примерно от 5 секунд для коротких текстов до более чем 10 секунд для длинных текстов. В отличие от этого, модель ONNX продемонстрировала значительно улучшенную производительность, сократив время, необходимое для вывода, примерно на 85% по сравнению с УАСПР.

В одноядерной конфигурации оптимизированные модель ONNX и квантованная модель также продемонстрировали явное преимущество над моделью УАСПР. Например, время вывода для длинных текстов составило 8.97 секунд для модели УАСПР и 6.06 секунд для модели ONNX что подтверждает, что оптимизированные модели могут эффективно работать даже в средах с ограниченными ресурсами.

Оптимизированные модели продемонстрировали более стабильную производительность с меньшими стандартными отклонениями. Например, стандартное отклонение значений КРВ для модели ONNX составило 0.012…0.056, тогда как для модели УАСПР оно составило 0.103…0.665. Это свидетельствует о том, что оптимизированные модели обеспечивают более предсказуемую производительность для различных текстов.

Учитывая характеристики производительности, модель ONNX и квантованная модель гораздо более подходят для практического применения, особенно в системах, работающих в реальном времени. Производительность квантованной модели, которая почти идентична модели ONNX, создает исключительные возможности для развертывания на устройствах с ограниченными ресурсами.

Проведенное исследование подтвердило эффективность разработанной системы ПТР для армянского языка с точки зрения как качества синтезированной речи, так и производительности. Полученные результаты демонстрируют успешное решение поставленных задач и достижение целей исследования. Оптимизированные версии модели обеспечивают возможность практического применения системы в различных сценариях, включая устройства с ограниченными ресурсами и системы, работающие в реальном времени. Разработанная система представляет собой значительный вклад в развитие технологий синтеза речи для армянского языка. Достигнутые показатели производительности и качества открывают новые возможности для внедрения данной технологии. Результаты исследования могут служить основой для дальнейшего совершенствования систем синтеза речи. Таким образом, представленная работа вносит существенный вклад в область обработки естественного языка и речевых технологий.

ОСНОВНЫЕ ВЫВОДЫ ПО ДИССЕРТАЦИОННОЙ РАБОТЕ

1. Разработана и реализована полнофункциональная система преобразования текста в речь на армянском языке, включающая модель преобразования графем в фонемы, обученную и оптимизированную с учетом фонетических особенностей армянского языка. Система успешно адаптирована для использования на встроенных устройствах посредством оптимизации модели с применением методов квантования и оптимизации вычислительного графа. [2, 3, 8, 9]
2. Создана система ПТР на армянском языке с архитектурой УАСПР, учитывающая специфику фонологической системы армянского языка, включая трёхрядную классификацию согласных (глухие, звонкие и придыхательные), позиционные изменения произношения гласных (различное произношение букв «և» и «ո» в начале и середине

слова), а также правила ударения. Система продемонстрировала выдающиеся результаты по сохранению идентичности диктора (средний коэффициент сходства - 0.9366), воспроизведению интонации (корреляция по F0 - 0.8834) и разборчивости речи (средний коэффициент ошибок слов - 0.7248). [2, 7]

3. Разработана модель ГФП на архитектуре Conformer-CTC, обеспечивающая высокую точность: коэффициент ошибок на уровне слов (КСО) - 16.13%, на уровне фонем (КФО) - 17.36%. Эти показатели значительно превосходят результаты существующих решений (например, у Phonemizer: КСО - 96.60%, КФО - 36.15%), что подтверждает эффективность предложенной модели в обработке сложной фонетической системы армянского языка. [1, 4, 5, 7]
4. Сформированы высококачественные и полноценно представленные наборы данных для армянского языка, доступные для дальнейших исследований: [1, 8, 9]
    - набор звуковых данных: 21 час (75,598 секунд) записи, 14,182 аудиофайла и 14,078 уникальных предложений, содержащих 137,716 слов (30,466 уникальных);
    - набор G2P-данных: 17,862 пары слово-фонема, обеспечивающие комплексное покрытие фонетической системы армянского языка.
5. Разработаны и внедрены автоматизированные системы сбора данных, существенно ускоряющие процесс аннотирования и пригодные для генерации аналогичных корпусов на других языках: [1, 8, 9]
    - система автоматического сбора звуковых данных, использующая шкалу dBFS для оценки уровня сигнала и метод динамического порогового значения;
    - система автоматического сбора ГФП-данных, основанная на использовании Википедии как основного источника и реализующая многоуровневую систему лингвистического анализа.
6. Модель ПТР оптимизирована в формате ONNX с применением методов квантования, что позволило значительно сократить её размер (на 87%) - с 1.0 ГБ до 127.5 МБ. Оптимизированные модели (в форматах ONNX и квантованные) показали значительно более высокую производительность, сократив время вывода примерно на 85% и обеспечив значения КРВ в диапазоне 0.22…0.30, что позволяет использовать их в системах реального времени с 3…4-кратным превышением скорости воспроизведения по сравнению с обычным временем. [5, 6]
7. Разработан интуитивно понятный и доступный для пользователя программный интерфейс, включающий модули ввода текста, оптического распознавания текста и историю взаимодействий. Четкое разделение клиентской и серверной частей системы, а также применение технологий React и Flask обеспечили стабильную, быструю и удобную графическую среду. [1, 3]
8. Результаты исследования продемонстрировали, что с применением современных методов глубокого обучения возможно эффективно решать задачу синтеза речи на армянском языке, создавая систему, способную генерировать естественно звучащую, высококачественную речь. Полученные результаты приближены к международным стандартам, что представляет особую ценность для языков с ограниченными ресурсами, таких как армянский. [1, 2, 8, 9]

Предлагается:

1. Создание полнофункциональной многоязычной ПТР - системы на армянском языке на базе архитектуры Tacotron2 с высокими качественными показателями (0.9366 - идентичность диктора, 0.8834 - F0-корреляция), учитывающей фонологические особенности языка и интегрирующей высокоточную ГФП-модель (КСО - 16.13%, КФО - 17.36%).

2. Внедрение оптимизированной и квантованной модели в формате ONNX, уменьшающей объем модели на 87% (до 127.5 Мб) и обеспечивающей работу в реальном времени (RPB 0.22…0.30) даже на устройствах с ограниченными вычислительными ресурсами, что открывает возможности интеграции в различные платформы и программные решения.
3. Применение высококачественных наборов данных на армянском языке (21 час записи, 14,182 аудиофайла и 17,862 пары слово-фонема), а также методологии их автоматизированного сбора для решения различных задач в области обработки армянской речи и создания аналогичных систем для других языков с ограниченными ресурсами.
4. Разработка удобного пользовательского интерфейса с интеграцией OPC, применимого в образовательной среде, целью обеспечения доступности информации для лиц с ограниченными возможностями, расширения телекоммуникационных сервисов и укрепления цифрового присутствия армянского языка.

Основные результаты диссертации опубликованы в следующих работах:

1. **Nikoghosyan K., Ghukasyan T.** Fine-Tuning Tesseract For More Accurate And Robust Optical Character Recognition // Bulletin of RAU: Physics-Mathematics and Natural Sciences. - 2022. - No 1. - P. 31-41, doi: 10.48200/1829-0450_pmn_2022_1_33.
2. **Nikoghosyan K.H.** An Overview Of The Existing Text-To-Speech(TTS) Algorithms Based On Deep Learning // Proceedings of NPUA: Information Technologies, Electronics, Radio engineering. - 2022. - No 2. - P. 63-71, doi: 10.53297/18293336-2022.2-63.
3. **Nikoghosyan K.H., Harutyunyan E.A., Galstyan D.M.** Improving The Image-To-Speech System Accuracy Through Integration Of Optical Character Recognition And Language Processing Techniques // Proceedings of NPUA: Information Technologies, Electronics, Radio engineering. - 2023. - No 1. - P. 44-50, doi: 10.53297/18293336-2023.1-44.
4. **Galstyan D.M., Harutyunyan E.A., Nikoghosyan K.H.** Human Action Recognition: Improving The Accuracy Of Deep Conv-LSTM Architecture Through Noise Cleaning Prior To Key Frames Selection // Proceedings of the RA NAS and NPUA. Series of Technical Sciences: ISSN:0002-306X. - 2023. - Vol. 76, № 2. - P. 202-209, doi: 10.53297/0002306X-2023.v76.2-202.
5. **Nikoghosyan K.H., Khachatryan T.B., Harutyunyan E.A., Galstyan D.M.** Acceleration Of Transformer Architectures On Jetson Xavier Using TensorRT // Proceedings of NPUA: Information Technologies, Electronics, Radio engineering. - 2023. - No 2. - P. 30-40, doi: 10.53297/18293336-2023.2-30.
6. **Nikoghosyan K.H., Khachatryan T.B., Harutyunyan E.A., Galstyan D.M.** A Comprehensive System For Detecting Deepfake Videos And AI-Generated Text // Proceedings of NPUA: Information Technologies, Electronics, Radio engineering. - 2024. - No 1. - P. 37-44, doi: 10.53297/18293336-2024.1-37.
7. **Nikoghosyan K.H., Khachatryan T.B., Harutyunyan E.A., Galstyan D.M.** Evaluating Open-Source Image Captioning Models With Multiple Metrics On The IAPR TC-12 Dataset // Bulletin of NPUA: Collection of scientific papers. - 2024. - No 1. - P. 164-172.
8. **Simonyan S., Nikoghosyan K.** Enhancing TTS Performance for Noisy and Low-Resource Languages Using Advanced Noise Removal Techniques // 2024 IEEE East-West Design & Test Symposium (EWDTS). - 2024. - P. 1-7, doi: 10.1109/EWDTS63723.2024.10873673.
9. **Nikoghosyan K.H.** Leveraging Pause Detection For Enhanced TTS Dataset Generation // Proceedings of NPUA: Information Technologies, Electronics, Radio engineering. - 2024. - No 2. - P. 45-56, doi: 10.53297/18293336-2024.2-45.

ԵԶՐԱՀԱՆԳՈՒՄ

Ատենախոսության հիմնական արդյունքները և եզրահանգումները՝

1. Մշակվել և իրագործվել է հայոց լեզվով բազմախոսնակային տեքստը խոսքի վերափոխման (ՏԽՎ) ամբողջական համակարգ, որը ներառում է գրանշանները հնչյունների վերափոխման (ԳՀՎ) մոդել և ուսուցանվել ու օպտիմալացվել է հայոց լեզվի հնչյունաբանական առանձնահատկությունները հաշվի առնելով: Համակարգը հաջողությամբ հարմարեցվել է ներկառուցված սարքերում օգտագործման համար՝ իրականացնելով մոդելի օպտիմալացում քվանտացման և հաշվարկային գրաֆի օպտիմալացման տեխնիկաների միջոցով: [2, 3, 8, 9]
2. Մշակվել է ամբողջական ՓԱՏԽ ճարտարապետությամբ հայոց լեզվով ՏԽՎ համակարգ, որը հաշվի է առնում հայոց լեզվի հնչյունական համակարգի առանձնահատկությունները, ներառյալ բաղաձայնների եռաշար համակարգը (խուլ, ձայնեղ և շնչեղ խուլ), ձայնավորների դիրքային փոփոխությունները («ե» և «ո» տառերի տարբեր արտասանությունը բառասկզբում և բառամիջում) և շեշտադրման կանոնները: Համակարգը ցուցաբերել է գերազանց արդյունքներ խոսնակի նույնականության պահպանման (0.9366 միջին նմանության գործակից), ինտոնացիայի վերարտադրման (0.8834 F0-կոռելյացիա) և հասկանալիության (0.7248 ԿՕՀ) տեսանկյունից: [2, 7]
3. Մշակվել է G2P-Conformer-CTC ճարտարապետությամբ ԳՀՎ մոդել, որն ապահովում է բարձր ճշգրտություն՝ 16.13% բառային սխալի գործակից (ԲՍԳ) և 17.36% հնչյունային սխալի գործակից (ՀՍԳ): Այս ցուցանիշները զգալիորեն գերազանցում են առկա գործիքների արդյունքները (օրինակ՝ Phonemizer-ի 96.60% ԲՍԳ և 36.15% ՀՍԳ ցուցանիշները), ապացուցելով մշակված մոդելի արդյունավետությունը հայոց լեզվի բարդ հնչյունական համակարգի մշակման հարցում: [1, 4, 5, 7]
4. Ստեղծվել են հայոց լեզվի համար բարձրորակ տվյալների համապարփակ հավաքածուներ, որոնք հասանելի են հետագա հետազոտությունների համար. [1, 8, 9]
    - Ձայնային տվյալների հավաքածու՝ 21 ժամ (75,598 վայրկյան) ձայնագրություն, 14,182 ձայնային ֆայլ և 14,078 եզակի նախադասություն, 137,716 բառ (30,466 եզակի)
    - ԳՀՎ տվյալների հավաքածու՝ 17,862 բառ-հնչյուն զույգ, որոնք ապահովում են հայոց լեզվի հնչյունային համակարգի համապարփակ ներկայացվածությունը
5. Մշակվել և իրագործվել են տվյալների հավաքագրման ավտոմատացված համակարգեր, որոնք զգալիորեն արագացնում են տվյալների հավաքման գործընթացը և կարող են օգտագործվել նաև այլ լեզուների համար նմանատիպ տվյալների հավաքածուների ստեղծման համար: [1, 8, 9]
    - Ձայնային տվյալների հավաքագրման ավտոմատացված համակարգ՝ dBFS սանդղակով ազդանշանի մակարդակի հաշվարկով և դինամիկ շեմային մեթոդով
    - ԳՀՎ տվյալների հավաքագրման ավտոմատացված համակարգ, որն օգտագործում է Վիքիբառարանը որպես հիմնական աղբյուր և ներդրում է բազմաշերտ վերլուծության մեխանիզմ
6. ՏԽՎ ԱԲ մոդելն օպտիմալացվել է ONNX ձևաչափի և քվանտացման միջոցով, ինչը հանգեցրել է մոդելի չափի զգալի նվազեցման (87%-ով)՝ 1.0 ԳԲ-ից մինչև 127.5 ՄԲ: Օպտիմալացված մոդելները (ONNX և քվանտացված) ցուցաբերել են զգալիորեն ավելի բարձր արդյունավետություն, նվազեցնելով առերեսման ժամանակը մոտավորապես 85%-ով և ապահովելով ԻԺԳ արժեքներ 0.22-0.30 միջակայքում, ինչը նշանակում է, որ դրանք կարող են աշխատել իրական ժամանակից 3-4 անգամ ավելի արագ: [5, 6]
7. Մշակվել է օգտագործողի համար հարմարավետ և մասնչելի ծրագրային գործիք, որը ներառում է տեքստային մուտքագրման, նկարից տեքստի ճանաչման (ՕՆՃ) և պատմության մոդուլները: Համակարգի ճակատային և սերվերային հատվածների

հստակ տարանջատումը և React/Flask տեխնոլոգիաների կիրառումը ապահովել են կայուն, արագ աշխատող և հեշտ կիրառելի գրաֆիկական միջավայր: [1, 3]

8. Հետազոտությունն ապացուցել է, որ խորը ուսուցման ժամանակակից մեթոդների կիրառմամբ հնարավոր է լուծել հայոց լեզվի խոսքի սինթեզի բարդ խնդիրը՝ ստեղծելով համակարգ, որն արտադրում է բնական հնչող, բարձրորակ հայերեն խոսք: Ստացված արդյունքները շատ մոտ են միջազգային չափանիշներին, ինչը խոստումնալից է հայոց լեզվի նման սահմանափակ ռեսուրսներով լեզուների համար: [1, 2, 8, 9]

Առաջարկվում է՝

1. Ամբողջական ՓԱՏԽ ճարտարապետությամբ հայոց լեզվով բազմախոսնակային ՏԽՎ համակարգ՝ բարձր որակի ցուցանիշներով (0.9366 խոսնակի նույնականություն, 0.8834 F0-կոռելյացիա), որը հաշվի է առնում հայոց լեզվի հնչյունական համակարգի առանձնահատկությունները և ինտեգրում է բարձր ճշգրտությամբ ԳՀՎ մոդել (16.13% ԲՍԳ և 17.36% ՀՍԳ):
2. Օպտիմալացված և քվանտացված մոդել (ONNX ձևաչափով), որը 87%-ով նվազեցնում է մոդելի չափը (127.5 ՄԲ) և ապահովում է իրական ժամանակում աշխատանք (ԻԺԳ 0.22-0.30)՝ նույնիսկ սահմանափակ ռեսուրսներով սարքերում, ինչը հնարավորություն է տալիս համակարգը ներդնել տարբեր հարթակներում և ծրագրային լուծումներում:
3. Բարձրորակ հայոց լեզվի տվյալների հավաքածուներ (21 ժամ ձայնագրություն, 14,182 ձայնային ֆայլ և 17,862 բառ-հնչյուն զույգ) և դրանց ավտոմատացված հավաքագրման մեթոդաբանություն, որը կիրառելի է հայոց լեզվով խոսքի մշակման տարբեր խնդիրների համար և այլ սահմանափակ ռեսուրսներով լեզուների համար նմանատիպ համակարգերի մշակման ժամանակ:
4. Օգտագործողի համար հարմարավետ ծրագրային լուծում ՕՆՃ ինտեգրմամբ, որը կիրառելի է կրթական միջավայրում, հաշմանդամություն ունեցող անձանց համար տեղեկատվության մատչելիության ապահովման, հեռահաղորդակցության ոլորտի ծառայությունների ընդլայնման և հայոց լեզվի թվային միջավայրում ներկայացվածության ամրապնդման նպատակներով:

CONCLUSION

The main results and conclusions of the dissertation:

1. A complete multi-speaker text-to-speech (TTS) system in Armenian language has been developed and implemented, which includes a grapheme-to-phoneme (G2P) model and has been trained and optimized taking into account the phonological features of the Armenian language. The system has been successfully adapted for use in embedded devices by implementing model optimization through quantization and computational graph optimization techniques. [2, 3, 8, 9]
2. A complete TTS system with VITS architecture in Armenian language has been developed, which takes into account the peculiarities of the Armenian phonetic system, including the tri-tiered consonant system (voiceless, voiced, and aspirated voiceless), positional changes of vowels (different pronunciation of the letters "ե" and "ո" at the beginning and middle of words) and stress rules. The system has demonstrated excellent results in terms of speaker identity preservation (0.9366 average similarity coefficient), intonation reproduction (0.8834 F0-correlation) and intelligibility (0.7248 STOI). [2, 7]
3. A G2P-Conformer-CTC architecture G2P model has been developed, which provides high accuracy: 16.13% word error rate (WER) and 17.36% phoneme error rate (PER). These indicators significantly exceed the results of existing tools (for example, Phonemizer's 96.60% WER and 36.15% PER), proving the effectiveness of the developed model in processing the complex phonetic system of the Armenian language. [1, 4, 5, 7]

4. Comprehensive high-quality data collections for the Armenian language have been created, which are available for further research: [1, 8, 9]
    a. Audio data collection: 21 hours (75,598 seconds) of recording, 14,182 audio files and 14,078 unique sentences, 137,716 words (30,466 unique)
    b. G2P data collection: 17,862 word-phoneme pairs, which provide comprehensive representation of the Armenian language phonetic system
5. Automated data collection systems have been developed and implemented, which significantly accelerate the data collection process and can also be used to create similar data collections for other languages: [1, 8, 9]
    a. Automated audio data collection system with signal level calculation on the dBFS scale and dynamic threshold method.
    b. Automated G2P data collection system that uses Wiktionary as the main source and implements a multi-layer analysis mechanism.
6. The TTS AI model has been optimized through ONNX format and quantization, which has led to a significant reduction in model size (by 87%): from 1.0 GB to 127.5 MB. The optimized models (ONNX and quantized) have shown significantly higher efficiency, reducing inference time by approximately 85% and providing RTF values in the range of 0.22-0.30, which means they can run 3-4 times faster than real-time. [5, 6]
7. A user-friendly and accessible software tool has been developed, which includes text input, optical character recognition (OCR), and history modules. The clear separation of the system's frontend and backend parts and the application of React/Flask technologies have provided a stable, fast running, and easy to use graphical environment. [1, 3]
8. The research has proven that with the application of modern deep learning methods, it is possible to solve the complex problem of speech synthesis for the Armenian language by creating a system that produces natural-sounding, high-quality Armenian speech. The obtained results are very close to international standards, which is promising for languages with limited resources like Armenian. [1, 2, 8, 9]

It is proposed:

1. A complete VITS architecture multi-speaker TTS system in Armenian language with high quality indicators (0.9366 speaker identity, 0.8834 F0-correlation), which takes into account the peculiarities of the Armenian phonetic system and integrates a high-accuracy G2P model (16.13% WER and 17.36% PER).
2. An optimized and quantized model (in ONNX format), which reduces the model size by 87% (127.5 MB) and provides real-time operation (RTF 0.22-0.30) even on devices with limited resources, which allows the system to be implemented on various platforms and software solutions.
3. High-quality Armenian language data collections (21 hours of recording, 14,182 audio files, and 17,862 word-phoneme pairs) and their automated collection methodology, which is applicable for various speech processing tasks in Armenian and for the development of similar systems for other languages with limited resources.
4. A user-friendly software solution with OCR integration, which is applicable in educational environments, ensuring information accessibility for people with disabilities, expanding telecommunications services, and strengthening the representation of the Armenian language in the digital environment.