УДК 004.522

Обзор созданных речевых корпусов и программных средств для синтеза татарской речи

***Хусаинов Айдар Фаилович*[1],**
***Сулейманов Джавдет Шевкетович*[2]**
[1,2] *Институт прикладной семиотики Академии наук Республики Татарстан, Казань, Россия*
[1,2] *Казанский федеральный университет, Казань, Россия*

Аннотация

В статье приводится обзор основных этапов создания систем синтеза татарской речи. Обзор охватывает исследования института прикладной семиотики Академии наук Республики Татарстан, начиная с первой системы конкатенативного синтеза на основе дифонов, созданной в 1990-х годах, и до последней end-to-end нейронной системы, построенной в 2019 г. Несмотря на существенное различие в технологиях, необходимость создания корпусов звучащей речи остаётся неизменным условием построения синтезатора. Мы представляем несколько однодикторных корпусов, записанных в звукозаписывающей студии, каждый корпус был создан для определённой технологии синтеза. Этот аспект привёл к разнице в продолжительности и способе аннотации корпусов. Предварительные эксперименты показали, что синтезатор речи наилучшего качества может быть построен с использованием нескольких нейронных подходов, но единственный метод из реализованных нами, который на данный момент позволил также обеспечить синтез речи в реальном времени, основан на архитектуре Tacotron2 с нейронным вокодером WaveGlow.

Ключевые слова: синтез речи, речевые корпуса, татарский язык.

ВВЕДЕНИЕ

Задача синтеза речи состоит в генерации речевого сигнала на основе произвольного текстового фрагмента. По мере внедрения речевого интерфейса взаимодействия с различными устройствами повысилась важность синтезаторов речи, способных эффективно генерировать естественно звучащую и разборчивую речь. Несмотря на значительные успехи последних пяти лет, синтезаторы речи не созданы для подавляющего большинства мировых языков.

В данной статье мы представляем краткий обзор созданных в институте прикладной семиотики АН РТ речевых корпусов и программных средств синтеза татарской речи.

В разделе 2 данной статьи представлен обзор основных подходов к синтезу речи, раздел 3 описывает созданные обучающие корпуса, раздел 4 содержит описание созданных систем синтеза татарской речи.

1. Обзор основных подходов к синтезу речи

Среди требований, предъявляемых к системам синтеза речи, основными являются качество синтезированного сигнала и скорость работы систем. Скорость работы позволяет использовать синтезатор в приложениях реального времени, таких как голосовые ассистенты, при автоматическом озвучивании перевода. Оценка качества автоматического синтеза речи, в свою очередь, также представляет собой сложную задачу. Применяются автоматические методы, в том числе использующие оценки вероятностей, полученные с помощью систем распознавания речи, однако такие характеристики могут лишь косвенно свидетельствовать о качестве речевого сигнала. Общепринятыми на данный момент методиками оценки являются экспертные оценки согласно методикам MOS, MUSHRA [1]. Среди критериев, по которым происходит оценка, основными являются разборчивость (легкость восприятия произнесенного текста) и естественность звучания синтезированной речи.

Среди подходов, разработанных для решения задач синтеза речи, можно выделить несколько основных классов:

- конкатенативные подходы (дифонный синтез [2], Unit selection [3]). Исходной информацией в данных подходах служат выделенные из речевых записей акустические единицы;
- параметрические подходы, в котором базовыми элементами являются статистические модели звуков языка;
- гибридные модели, совмещающие два предыдущих класса.

На сегодняшний день наилучшие результаты достигаются с помощью моделей, основанных на параметрическом нейросетевом подходе.

Основной сложностью при создании качественных нейросетевых моделей синтеза речи является необходимость учитывать длительные зависимости, существующие в речевых сигналах. Введение в систему большого количества параметров, способных помочь в решении задач, одновременно с этим способно серьёзно замедлить процессы как моделирования, так и работы итоговой системы синтеза речи. При этом также следует учитывать, что усложнение системы может приводить к замедлению её работы, что в большинстве практических приложений является недопустимым. Для человеческого уха снижение частоты дискретизации ниже 16 кГц является заметным на слух, что предъявляет к синтезаторам, работающим в режиме реального времени, требования по генерации не менее 16 тысяч отсчётов аудиосигнала в секунду.

Работу всех нейросетевых синтезаторов речи можно поделить на два этапа: подготовку необходимого набора характеристик, выровненных по времени (например, мел-спектрограммы, частота основного тона, различные лингвистические характеристики); преобразование подготовленных характеристик в итоговый аудиофайл. При этом количество нейросетей, входящих в систему, и их архитектура могут отличаться в различных подходах. Единым для всех подходов остается требование по наличию обучающего корпуса аудиоданных. Разрабатываемые модели для английского языка часто тестируются на речевых базах данных, среди которых наибольшей популярностью среди разработчиков пользуются корпуса The LJ Speech Dataset [4], VCTK Corpus [5], CMU_ARCTIC [6], а также множество аудиокниг.

2. Лингвистические ресурсы для обучения

Независимо от выбора архитектуры системы синтеза татарской речи, для её построения необходимо наличие корпуса речи. Источником для его создания могут служить аудиокниги, многодикторные корпуса. Однако небольшое количество имеющихся на татарском языке аудиокниг, а также желание получить максимально качественный синтезированный голос, привели нас к необходимости создания собственных речевых баз данных.

По состоянию на ноябрь 2019 г. институт прикладной семиотики подготовил четыре речевых корпуса для построения синтезатора татарской речи, корпус, собранный из аудиокниг, ещё находится в процессе подготовки.

Все записи для монодикторных корпусов осуществлялись в профессиональных звукозаписывающих студиях, в качестве дикторов приглашались актеры татарского академического театра, также имеющие опыт в озвучивании книг и ведении теле- и радиопередач.

Основные характеристики корпусов представлены в табл. 1.

Таблица 1.

Продолжительность корпусов для задачи синтеза татарской речи

Диктор	Исходный формат записи	Продолжительность записей	Итоговая продолжительность корпуса	Аннотация
Тавзих	22.05 кГц 16 бит моно	Хранится 1009 записей дифонов		
Алмаз	44.1кГц 16 бит стерео	12:08:22	5:30:02	+
Алсу	48 кГц 32 бит моно	7:30:32	6:17:31	+
Рамиль	44.1 кГц 24 бит стерео	22:58:48	16:53:59	–
Аудиокниги (31 диктор)	44.1 кГц 24 бит стерео	114:28:10	–	–

Корпус «Тавзих» создавался для конкатенативного дифонного синтеза, поэтому его структура значительно отличается от других корпусов [7]. В качестве озвучиваемого текста выступал набор шаблонных фраз, состоящих из трех ритмических групп. Средняя ритмическая группа, в которой озвучивался целевой дифон, представляла трёх- или четырехсложное синтетическое слово, начальная и конечная ритмические группы фразы при этом оставались неизменными. Часть дифонной базы была размечена по периодам основного тона. Разметка проводилась отдельно для обеих полуфонем, входящих в дифон с указанием границы между фонемами. Для размеченных дифонов числовые отсчёты приведены в соответствующих текстовых файлах. Каждое значение представляет собой удвоенное число байтов от начала звукового файла.

В корпусах «Алмаз» и «Алсу» имеются дополнительные уровни аннотации: в аудиофайлах экспертами были вручную размечены все интонационные группы, отмечены заимствованные и акцентные слова, после чего для всего прочитанного текста была построена фонетическая транскрипция и автоматическими средствами определены границы произнесения каждой фонемы. Была реализована модель разметки корпуса, основные элементы которой можно представить следующим образом.

1. Уровень фонем: текущая фонема, две предшествующие, две последующие фонемы.

2. Уровень слогов: тип слога (V, VC, CV, CVC, VCC, CVCC); позиция фонемы в слоге; количество фонем в предыдущем, текущем, последующем слоге; номер текущего слога в слове; гласная в текущем слоге.

3. Уровень слов: часть речи, количество слогов для предыдущего, текущего, следующего слова; количество предшествующих и последующих слов во фразе.

4. Уровень фразы: количество слов/слогов в предыдущей, текущей, последующей фразе.

После аннотирования корпус хранится в формате массива троек файлов: wav (аудио-фрагмент) – txt (текстовая аннотация аудиофрагмента) – lab (пофонемное аннотация аудиофрагмента). Содержимое txt-файла представляет собой текстовое представление озвученной фразы, а также следующие специальные символы: * – для обозначения интонации перечисления; |, – для обозначения незавершенной по интонации синтагмы; . – для обозначения завершённой по интонации синтагмы; ! – для восклицательной интонации; ? – для вопросительной интонации; ~ – для обозначения заимствованных слов.

Структура lab-файла построена на основе адаптированного под татарский язык файла аннотации, изначально разработанного для синтеза английской речи на базе скрытых Марковских моделей [8]. Итоговый формат файла аннотации:

t1 t2 p1^p2-p3+p4=p5@p6_p7 /A:a3 /B:b3@b4-b5&b6-b7|b16 /C:c3 /D:d1_d2 /E:e1+e2@e3+e4 /F: f1_f2 /H:h1=h2|h5 /J:j1,

где t1, t2 – временные отсчёты начала и окончания звучания каждой фонемы, полученные автоматически с помощью систем распознавания речи, специально построенных для каждого из голосов.

Описание остальных использованных обозначений приводится в табл.2.

Таблица 2

Обозначения параметров, использованных в lab-файлах аннотации

Параметр	Описание	Параметр	Описание
p1	Фонема за две до текущей	c3	Число фонем в следующем слоге
p2	Предыдущая фонема	d1	Часть речи предыдущего слова
p3	Текущая фонема	d2	Число слогов в предыдущем слове
p4	Следующая фонема	e1	Часть речи текущего слова
p5	Фонема через одну от текущей	e2	Число слогов в текущем слове
p6	Позиция текущей фонемы в текущем слоге	e3	Позиция текущего слова в текущей фразе
p7	Позиция текущей фонемы в текущем слоге, считая с конца слога	e4	Позиция текущего слова в текущей фразе, начиная с конца фразы
a3	Количество фонем в предыдущем слоге	f1	Часть речи следующего слова
b3	Количество фонем в текущем слоге	f2	Число слогов в следующем слове
b4	Позиция текущего слога в текущем слове	h1	Число слогов в следующей фразе
b5	Позиция текущего слога в текущем слове, начиная с конца слова	h2	Число слов в следующей фразе
b6	Позиция текущего слога в текущей фразе	h5	Тип синтагмы
b7	Позиция текущего слога в текущей фразе, начиная с конца фразы	j1	Заимствованное слово или нет
b16	Гласная в текущем слоге		

Корпуса «Алмаз» и «Алсу» для использования при обучении были приведены к формату аудио 16 кГц 16 бит моно, корпус «Рамиль» – 22.05 кГц 16 бит моно. Записи корпуса «Рамиль» также прошли дополнительную предобработку, позволившую улучшить качество итоговой системы синтеза, обработка включала следующие этапы:

- были отфильтрованы короткие (<1 секунды) и длинные (>11 секунд) записи;

- все аудио были нормализованы по громкости;
- были удалены начальные и конечные фрагменты тишины.

Различия в аннотации корпусов также включали формат представления соответствующих текстовых фрагментов: «Алмаз» и «Алсу» содержали фонетическую транскрипцию, «Рамиль» – транслитерацию на латиницу; во всех корпусах были отфильтрованы все знаки пунктуации, кроме знаков точки, запятой, восклицательного и вопросительного знаков.

Коллекция аудиокниг на данный момент не размечена. После аннотации планируется использование корпуса для тестирования алгоритмов многодикторного синтеза речи и технологий построения синтезатора речи для нового голоса по небольшому количеству записей.

3. Архитектура систем синтеза речи

Системы синтеза татарской речи были построены на базе следующих подходов:

- конкатенативный дифонный синтез [7];
- HTS-параметрический синтез [9];
- нейросетевой синтез на основе Merlin [10];
- нейросетевой синтез на основе DCTTS [11];
- нейросетевой синтез на основе Tacotron2 [12] / WaveGlow [13].

Система Merlin отличается от двух других нейросетевых подходов тем, что требует пофонемно аннотированного корпуса для обучения нейросетей: первая нейросеть обучается моделированию длительностей произнесения для каждой фонемы, вторая – акустических характеристик каждой из фонем. Заключительным этапом работы системы является использование вокодера, преобразующего акустические характеристики и данные о длительностях фонем в результирующий речевой сигнал. В качестве вокодера для системы синтеза татарской речи был использован вокодер WORLD [14].

В ходе экспериментов с подходом Merlin были построены синтезаторы речи с использованием различных архитектур нейросетей:

- размер скрытых слоёв: [1024, 1024, 1024, 1024, 1024, 1024], тип скрытых слоёв: ['TANH', 'TANH', 'TANH', 'TANH', 'TANH', 'TANH'];
- размер скрытых слоёв: [512, 512, 512, 512, 512, 512], тип скрытых слоёв: ['TANH', 'TANH', 'TANH', 'TANH', 'TANH', 'TANH'];

- размер скрытых слоёв: [1024, 1024, 1024, 1024, 256], тип скрытых слоёв: ['TANH', 'TANH', 'TANH', 'TANH', LSTM'];
- размер скрытых слоёв: [1024, 1024, 1024, 1024, 384], тип скрытых слоёв: ['TANH', 'TANH', 'TANH', 'TANH', LSTM'];
- размер скрытых слоёв: [1024, 1024, 1024, 1024, 256], тип скрытых слоёв: ['TANH', 'TANH', 'TANH', 'TANH', 'BLSTM'];
- размер скрытых слоёв: [1024, 1024, 1024, 1024, 384], тип скрытых слоёв: ['TANH', 'TANH', 'TANH', 'TANH', 'BLSTM'],

где 'TANH' – слой с гиперболическим тангенсом в качестве функции активации нейронов, (B)LSTM – (двунаправленный) слой с нейронами, имеющими дополнительные блоки (gates), позволяющие запоминать (или забывать) длительный контекст.

Была произведена субъективная оценка качества звучания построенных систем синтеза речи, согласно которой в качестве итоговой была выбрана система с заключительным двунаправленным слоем 384 long-short term memory нейронов (BLSTM).

Обучение нейросетей происходило на базе подготовленных lab-файлов, входящих в состав корпусов татарской речи. Обучение системы на подкорпусе голоса «Алсу» длительностью 3 часа 41 минут заняло около 14 часов на видеокарте GTX 1070.

Большинство этапов построения систем синтеза речи на базе описанного подхода было автоматизировано. Наиболее трудозатратной является операция подготовки разметки в виде lab-файлов. Недостатком построенной системы является низкая скорость генерации речи: с учетом всех необходимых процедур предобработки текста синтез одного предложения с частотой дискретизации 22.05 кГц на видеокарте GTX 1080 занимает около 16 секунд.

Подход, предложенный в DCTTS, является попыткой ускорить обучение нейросетевых моделей синтеза речи. Проблемы, связанные с трудностью обучения систем, аналогичных Merlin, заключаются в использовании рекуррентных слоев, которые, с одной стороны, позволяют моделировать «далекие» закономерности в речи, улучшая качество звучания, но, с другой стороны, не позволяют производить распараллеливание вычислений. В DCTTS предполагается обучать две глубокие нейросети, состоящие из свёрточных слоёв (Deep Convolutional TTS), обучение которых происходит гораздо быстрее.

Основная идея данного метода заключается в обучении нейросетей задаче генерации мел-спектрограмм, на основе которых вокодер тренируется восстанавливать речевой сигнал. Отличительной особенностью является отсутствие необходимости разметки обучающего аудиокорпуса: на вход нейросети достаточно подавать массив аудиофрагментов и соответствующий им текст (или его транскрипцию). Основная сложность заключается в сложности подбора гиперпараметров, а некорректное обучение механизма внимания, входящего в состав нейросетей, может приводить к периодическим повторам или удалениям отдельных частей фраз из итогового аудиосигнала.

Заключительным на момент написания статьи подходом к синтезу речи, реализованным в институте прикладной семиотики АН РТ, является Tacotron2 + WaveGlow. Аналогично DCTTS, Tacotron2 пытается сформировать соответствия между векторами для входных символов (char embeddings) и мел-спектрограммами. Изначально мел-спектрограммы подавались на вход модели нейросетевого вокодера WaveNet. Однако мы использовали архитектуру с альтернативным вокодером WaveGlow, нейросеть которой учится восстанавливать высококачественный аудиосигнал по мел-спектрограмме. Обучение нейросети осуществляется с помощью единственной функции потерь, что делает процедуру обучения более простой и стабильной. Данная архитектура позволила сохранить качество синтеза, достигаемое с помощью нейросетей WaveNet, при значительном ускорении работы синтезатора с 0.11 кГц (для WaveNet) до 520 кГц [13] (последние реализации вокодера позволили достичь скорости 4850 кГц [15]). Процесс обучения Tacotron2 и WaveGlow может вестись независимо друг от друга. Длительность обучения зависит от выбора гиперпараметров, текущие версии нейросетей для системы татарского синтезатора были обучены суммарно за 16 дней на 8 видеокартах V100 32GB.

Заключение

В данной статье были представлены результаты по созданию систем синтеза татарской речи. Полученные результаты в виде аннотированных корпусов, алгоритмов и программных систем являются первыми для татарского языка. Записанные монодикторные корпуса, а также автоматическая и экспертная разметки позволили сформировать набор обучающих данных, необходимых для обучения систем нейросетевого синтеза речи (например, Merlin). Разработанные системы автоматического синтеза татарской речи позволяют вести работы по внедрению речевого человеко-машинного интерфейса на татарском языке. Система синтеза речи уже внедрена в веб-сервис русско-татарского машинного переводчика [16] и татарского синтезатора [17]. Продолжение работ по данной тематике включает в себя проведение сравнительной оценки качества построенных синтезаторов.

Литература

1. International Telecommunication Union. Method for the subjective assessment of intermediate quality levels of coding systems. URL: https://www.itu.int/rec/R-REC-BS.1534/en.
2. *Moulines, E., Charpentier, F.* «Pitch-sinchronous waveform processing techniques for text-to-speech synthesis using diphones». In Speech Communication, 9 (5/6), 1990, p. 453–467.
3. *Sagisaka, Y.* ATR v-talk speech synthesis system. In Proc. ICSLP-92, 1992, Banff, Canada.
4. *Keith Ito.* The LJ Speech Dataset [Electronic resource]. URL: https://keithito.com/LJ-Speech-Dataset/.

5. English Multi-speaker Corpus for CSTR Voice Cloning Toolkit [Electronic resource]. URL: https://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html.
6. CMU_ARCTIC speech synthesis databases [Electronic resource]. URL: http://festvox.org/cmu_arctic/.
7. *Ибрагимов Т.И.* Синтезатор татарской речи. /*Т.И. Ибрагимов, Ф.И. Салимов, Д.Ш. Сулейманов* //Международная научно-практическая конференция «Система непрерывного образования инвалидов: опыт, проблемы, тенденции, решения». Казань: Академия управления ТИСБИ, 2006. — С. 91–97.
8. An example of context-dependent label format for HMM-based speech synthesis in English. URL: https://wiki.inf.ed.ac.uk/twiki/pub/CSTR/F0parametrisation/hts_lab_format.pdf.
9. Speech human-machine interface for the Tatar language / *A. Khusainov, A. Khusainova* // Artificial Intelligence and Natural Language Conference. (Saint-Petersburg, 10–12 November 2016). Helsinki: FRUCT Oy, 2016. P. 60–65.
10. *Zhizheng Wu, Oliver Watts, Simon King.* Merlin: An Open Source Neural Network Speech Synthesis System» in Proc. 9th ISCA Speech Synthesis Workshop (SSW9), September 2016, Sunnyvale, CA, USA.
11. *Hideyuki Tachibana, Katsuya Uenoyama, Shunsuke Aihara.* Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention. https://arxiv.org/abs/1710.08969.
12. *Jonathan Shen at al.* Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. arXiv:1712.05884v2.
13. *Ryan Prenger, Rafael Valle, Bryan Catanzaro.* WaveGlow: A Flow-based Generative Network for Speech Synthesis. arXiv:1811.00002.
14. *M. Morise, F. Yokomori, and K. Ozawa.* «WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,» IEICE transactions on information and systems, 2016.
15. NVIDIA/waveglow: A Flow-based Generative Network for Speech Synthesis. URL: https://github.com/NVIDIA/waveglow.
16. Русско-татарский переводчик TatSoft. Переводчик. URL: https://translate.tatar.
17. Синтезатор татарской речи TatSoft. Синтез. URL: https://speech.tatar.

OVERVIEW OF SPEECH CORPORA AND SOFTWARE FOR THE TATAR SPEECH SYNTHESIS

Khusainov Aidar Failovich[1], Suleymanov Dzhavdet Shevketovich[2]
[1,2] Institute of Applied Semiotics, Tatarstan Academy of Sciences,Kazan, Russia
[1,2] Kazan Federal University, Kazan, Russia
[1] khusainov.aidar@gmail.com
[2] dvdt.slt@gmail.com

Abstract

In this paper, we describe the main stages of creating systems for the synthesis of Tatar speech. This description covers our research starting from the first diphone-based concatenative synthesis system built in 1990s to the last end-to-end neural system built in 2019. Despite the significant difference in technology, the need to create a high-quality corpus of sounding speech remains an unchanged condition for the construction of a synthesizer. We present several single-speaker corpora recorded in sound recording studio, each of the corpus was created for a specific synthesis technology. This fact led to difference in total duration and annotation of corpora. Preliminary experiments showed that the best quality speech synthesizer can be built using several neural approaches, but the only method that also provides real-time synthesis uses Tacotron-2 architecture followed by neural WaveGlow-vocoder.

Keywords: speech synthesis; speech corpora, Tatar language.