DOI: 10.33910/1992-6464-2020-198-171-177

Э. С. Израилова

ПРОЦЕСС СОЗДАНИЯ СИСТЕМЫ СИНТЕЗА ЧЕЧЕНСКОЙ РЕЧИ

В статье описан процесс моделирования системы синтеза речи, основанный на имеющихся акустических данных, которые используются для машинного обучения с целью получения модели, соответствующей естественным характеристикам речи. Описываются этапы создания фонетико-акустической базы данных, адаптированной для обучения системы автоматического синтеза речи. Представлен фонетический транскриптор, разработанный с учетом проблематики графемно-фонемных преобразований чеченского языка. Дана информация о подготовке обучающей экспериментальной базы данных, процессе машинного обучения системы, настройке параметров нейронной сети, результате эксперимента по обучению системы синтеза речи. Рассмотрена проблема устранения графической омонимии при транскрибировании чеченских текстов и пути ее решения.

Ключевые слова: синтез речи, система синтеза чеченской речи, фонетико-акустическая база данных, фонетический транскриптор, модель DCTTS, омографы.

E. Izrailova

CREATING A SYSTEM FOR SYNTHESIZING THE CHECHEN SPEECH

The article examines the process of modelling a speech synthesis system based on the available acoustic data used for machine learning in order to obtain a model that matches the natural characteristics of speech. The description includes the stages of creating a phonetic-acoustic database adapted for training a system of automatic speech synthesis. The authors present a phonetic transcriptor developed to reflect the specific features of grapheme-phonemic transformations of the Chechen language. The article reports on the preparation of an experimental training database, the process of machine training a system and configuring a neural network, and the result of an experiment on training a speech synthesis system. The authors also consider the issue of graphic homonymy which occurs when transcribing Chechen texts, and suggest possible solutions for this problem.

Keywords: speech synthesis, Chechen speech synthesis system, phonetic-acoustic database, phonetic transcriptor, DCTTS model, homographs.

В настоящее время разработка речевых технологий является одним из самых динамично развивающихся направлений в науке и информационно-коммуникационной сфере. Крупнейшие мировые научные центры и коммерческие организации уже на протяжении пяти-шести десятков лет работают над созданием программных систем в области речевых технологий. Синтез речи является одной из важнейших задач речевой обработки и имеет широкое применение в современных информационных технологиях. Синтез речи по тексту является необходимым шагом в направлении более тесного общения человека с компьютером и может потребоваться во всех случаях, когда получателем информации является человек.

Синтез речи может быть использован:

- в информационно-справочных системах, для помощи слепым и немым, для управления человеком со стороны автомата;
- при объявлениях об отправлении поездов и самолетов;

- для выдачи информации о технологических процессах: в военной и авиакосмической технике, в робототехнике, в акустическом диалоге человека с компьютером [4, с. 138].

В отделе прикладной семиотики Академии наук Чеченской Республики три года назад началась работа над проектом, связанным с созданием системы синтеза чеченской речи. Для чеченского языка это совершенно новая область как в сфере прикладных наук, так и в плане теоретического исследования. Первоочередной задачей в рамках проекта было исследование различных методов и систем синтеза речи для наиболее популярных языков мира, по которым ученым удалось достигнуть высоких результатов в области речевых технологий.

Современные информационные технологии в основном связаны с теми естественными языками, для которых доступны необходимые языковые и речевые электронные ресурсы, или же с языками, которые стали по какой-либо экономической или политической причине представлять интерес для мирового сообщества. Большая же часть языков развивающихся стран и малочисленных народов на сегодняшний день изучена недостаточно.

В Российской Федерации, помимо русского языка, наиболее распространенными по числу носителей являются татарский (свыше 5 млн носителей), чеченский, башкирский и чувашский языки (не менее 1 млн носителей). Все данные языки по международной классификации считаются малоресурсными — это понятие относится к естественным языкам с некоторыми (или всеми) из следующих свойств: недостаток своей системы письменности или устойчивой орфографии; нехватка квалифицированных лингвистов и переводчиков для данного языка; ограниченное распространение в сети Интернет; нехватка электронных ресурсов для обработки языка и речи, в том числе одноязычных корпусов, двуязычных электронных словарей, орфографических и фонетических транскрипций речи, словарей произношения и т. д. [3, с. 119].

Так как чеченский язык относится к малоресурсным языкам и мало обеспечен различными электронными базами, словарями и лингвистическими ресурсами, наиболее оптимальным методом синтеза речи должен быть тот, который использует для реализации минимальные ресурсы. Реализация такого «экономичного» метода оказалась возможна благодаря сравнительно недавно разработанным системам нейронных сетей сквозного тестирования (end-to-end) для автоматического синтеза речи. Такие системы обладают рядом преимуществ по сравнению с системами, так как объединяют в себе сразу все модули стандартных систем, что сокращает время обработки и объем требуемой памяти.

Глубокое обучение таких систем объединяет внутренние строительные блоки в единую модель и непосредственно связывает вход и выход (end-to-end). Это является основным преимуществом моделей сквозного тестирования по сравнению с классическими системами синтеза речи, обычно состоящими из различных доменных модулей (текстовый анализатор, генератор F0, генератор спектра, вокодер).

На сегодняшний день можно выделить самые актуальные архитектуры нейронных сетей для синтеза речи:

- WaveNet, разработанная компанией Google [11];
- SampleRNN, основанная на иерархии нескольких слоев нейронной сети [8, с. 1725];
- Tacotron модель с модулем «внимания» [9, с. 1];
- Deep Voice, состоящая из глубоких нейронных сетей [7, с. 196];
- Deep Convolutional Text To Speech (DCTTS) [10, с. 4785].

Для моделирования системы синтеза чеченской речи было решено остановиться на архитектуре DCTTS как наиболее оптимальной по соотношению время обучения / качество синтеза. Модель DCTTS демонстрирует

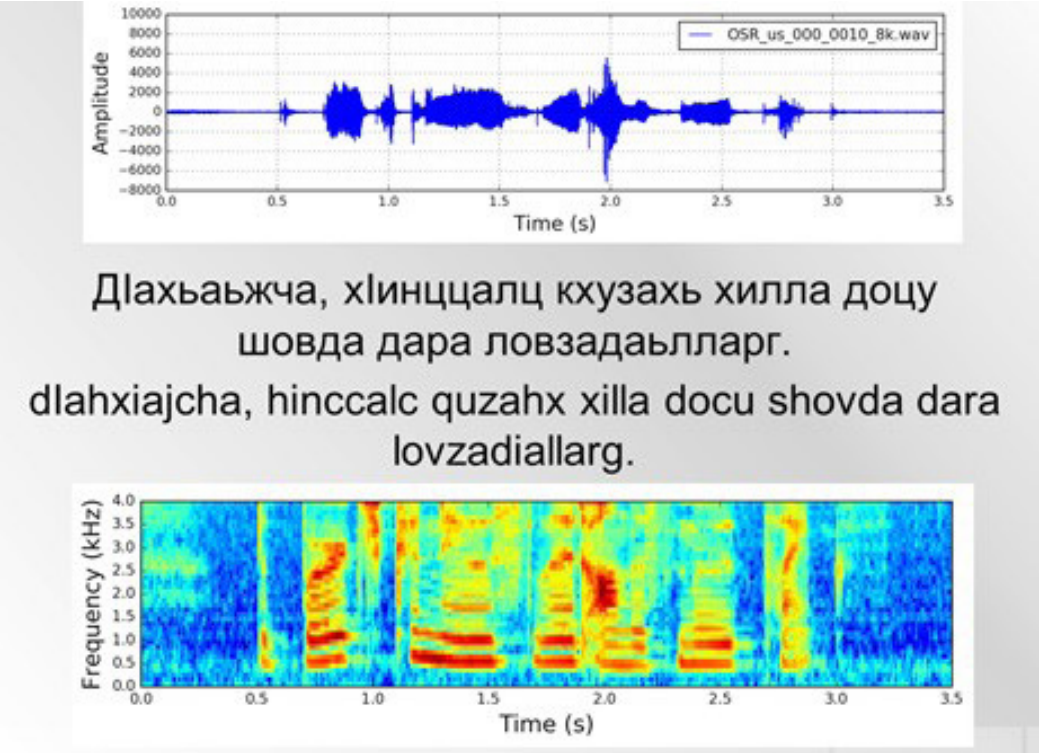

Рис. 1. Образец предложения из обучающей базы

высокую производительность и скорость обучения и имеет относительно ограниченные требования к вычислительной мощности компьютера.

Разработка систем синтеза речи ведется с использованием фонетико-акустических, текстовых, речевых баз данных. Для обеспечения высокого качества синтезированной речи эти базы данных должны содержать достаточно полный набор фонетических, просодических и акустических элементов речи. Чеченский язык является языком с богатой морфологией, что приводит к существенному увеличению размера словаря и модели языка по сравнению, например, с английским языком.

На первом этапе исследования была собрана речевая и текстовая база данных на основе аудиоматериалов и книг, находящихся в свободном доступе в сети Интернет. Для системы DCTTS размер минимальной обучающей базы составляет 5 часов. Соответственно, был подготовлен массив объемом пять часов речи в виде 3500 тысяч предложений в формате .wav и текстовый документ, в котором каждому звуковому файлу соответствует предложение. Далее был проведен анализ полученных речевых данных, реверберация, очистка от шума. После этого были созданы спектрограммы звуковых файлов с целью сокращения количества данных для обучения нейронной сети (рис. 1).

Модель DCTTS использует в качестве символов алфавита лишь графемы, то есть задача транскрибирования текста по правилам «графема к фонеме» отпадает, что является еще одним плюсом данной системы. Однако для чеченского языка такой подход не мог быть реализован, так как существующие проблемы с орфографией и ее фонетическим соответствием не позволяли непосредственно использовать графемную запись в качестве входных данных для обучения нейронной сети. Исходя из этого было решено использовать транскрибированные чеченские тексты, а в качестве входного

алфавита — латинские буквы и символы из фонетического алфавита AZBAT, разработанного нами в качестве основы будущей системы синтеза чеченской речи [1, с. 36]. При разработке алфавита AZBAT учитывались особенности произношения и графики, правила сочетаемости и вариативность фонем, описанные в работах чеченских филологов. За основу алфавита AZBAT была взята система чеченской фонетической транскрипции из книги «Грамматика чеченского языка» [6, с. 172–189], а общепринятым обозначениям фонем международного фонетического алфавита (MFA) соответствуют английские строчные и заглавные графемы. Такое упрощенное обозначение обусловлено необходимостью дальнейшей работы с программным кодом, а различные диакритические знаки могут послужить препятствием для адекватной работы программы.

Алфавит AZBAT состоит из 63 фонем (27 гласных и 36 согласных). В нем каждая фонема представлена одной, двумя или тремя буквами (табл. 1). Две буквы используются для обозначения дифтонгов и долготы гласных звуков **[aa, ee, ii, uu, oo, ie, uo, oa]**.

Умлаутированные гласные фонемы **аь, оь, уь** обозначены двухбуквенными сочетаниями **[ia, io, iu]**, долгие дифтонги — трехбуквенными сочетаниями **[uoo, iee, iuu, ioo]**. Символом **@** обозначена гортанная смычка **(ъ)**; **I** — обозначение фарингального звонкого спиранта; **II** — ларингальный абпруптивный звонкий согласный [1, с. 36].

Для реализации поставленной задачи была разработана программа автоматического транскрибирования чеченских текстов, разработанная на языке программирования Python с соблюдением основных правил чеченской орфоэпии.

В результате подготовленные предложения из текстовой базы были транскрибированы (табл. 1) и запущено обучение нейросетевого модуля системы, состоящего из двух сетей: Text2Mel, которая синтезирует мел-спектрограмму из входного текста, и Spectrogram Super-resolution Network (SSRN), которая преобразует мел-спектрограмму звукового сигнала в амплитудную спектрограмму STFT, учитывая пропуски кадров и восстанавливая частоту дискретизации. Данный метод зависит от спектральных представлений аудиосигналов, а сверхточная нейронная сеть CNN используется для построения спектрограммы, из которой простой вокодер может синтезировать сигнал. Для кодирования-декодирования символов входного алфавита в последовательность мел-спектров используется техника, позволяющая преобразовывать аналоговый речевой сигнал в вектор фиксированной длины. Эта проблема была решена с помощью механизма, называемого «внимание», и теперь такой механизм стал стандартной идеей в методах обучения seq2seq.

Машинное обучение системы строилось на основе библиотек и модулей языка программирования Python (Tensorflow, matplotlib, numpy, scipy). Процесс обучения системы синтеза речи продолжался 27 часов, было выполнено 500 тысяч итераций для нейронной сети Text2Mel и 270 тысяч итераций для сети SSRN. Результат эксперимента по машинному обучению и моделированию системы синтеза чеченской речи оказался успешным, удалось синтезировать речевой сигнал. По нашим предварительным оценкам, достигнуто среднее качество синтезируемой речи, разборчивость речи составляет 80 %. Некоторые слова система читает еще плохо, не всегда правильно озвучиваются долгие и краткие гласные, голос системы — роботизированный, но на данном первоначальном этапе, при использовании минимальной обучающей базы, полученный результат является скорее положительным: система синтеза чеченской речи создана, остается улучшить ее качество [2, с. 33].

Наши коллеги из Академии наук Татарстана провели эксперимент по машинному обучению нейросетевой модели Tacotron2+WaveGlow, аналогичной системе сквозного тестирования DCTTS на той же самой 5-часовой базе чеченской речи. Но эта попытка оказалась неудачной, хотя обучение

Таблица 1

Примеры транскрибированных предложений

• Айдамирова Машар. Сирла шовда.	aydamirova mashar. sirla shovda.
• Цхьаммо аьлла боху-кх: «Ирсан цIа дан дезаш хилча, цуьнан уггаре а йоккха чоь — ладегIаран чоь хир яра».	chxammo ialla boxuq irsan cIa dan dyezash xilcha, ciunan uggarye a yokqa chio — ladyegIaran chio xir yara.
• ДIахьаьжча, хIинццалц кхузахь хилла доцу шовда дара ловзадаьлларг.	dIahxiajcha, hinccalc quzahx xilla docu shovda dara lovzadiallarg.
• Олуш ду-кх: «ЙоI дуьненчу яьлча — лаьттахь керла шовда гучудолу, кIант дуьненчу ваьлча — стиглахь седа лета».	olush duq yoI diunyenchu yialcha — liattahx kyerla shovda guchudolu, kIant diunyenchu vialcha — stiglahx syeda lyeta.
• Гу тIера охьахьаьжча гуш болу Iам, хIокхуьнга гIийла бIаьрг бетташ Iара.	gu tIyera ohxahxiajcha gush bolu Iam, hoqiunga gIiyla bIiarg byettash Iara.

модели на татарском речевом корпусе, состоящем из 17 часов речи, показало положительный результат — синтез татарской речи высокого качества и разборчивости. Данный факт свидетельствует о том, что для обучения текущей конфигурации нейронных сетей требуется более длительный и качественный речевой корпус.

В рамках проекта автоматической проверки правописания в отделе прикладной семиотики была собрана полная лексическая база чеченского языка, состоящая из 3,8 миллионов слов и словоформ. На основе созданной базы были проведены статистические исследования и выявлены данные, согласно которым из всех слов и словоформ чеченской лексической базы фактически в речи используется только 270 000 слов. Обучающий речевой корпус должен содержать все фактически используемые слова — это и есть первоочередная задача оптимизации созданной системы синтеза речи. Размер обучающего корпуса чеченской речи должен быть не менее 20 часов речи, озвученной профессиональным диктором в студии звукозаписи.

Как нам кажется, разрабатываемая система синтеза чеченской речи должна удовлетворять следующим требованиям:

- автоматический синтез любого произвольного текста и подготовленных текстовых файлов с созданием соответствующих звуковых файлов в формате .wav;
- программный модуль предварительной обработки должен генерировать речевой сигнал первого предложения исходного текста в среднем за 0,8 секунд, задержка при озвучивании следующих предложений должна быть не более 0,5 секунд, так как вокодер продолжит работу параллельно остальным модулям системы;
- в дальнейшем планируется совершенствование системы синтеза речи со-

гласно требованиям стандарта ETSI TS101329 и рекомендациям Международного консультационного комитета по телефонии и телеграфии ITU-T, в которых определены допустимые задержки при передаче речи от 150 до 400 мс (для коммерческого применения);

- качество синтезируемой речи должно быть максимально приближено к естественной (MOS от 4 до 5) (оценка качества речи производится носителями синтезируемого языка, это так называемая MOS-оценка, производимая по пятибалльной шкале по нескольким категориям: общее впечатление, слуховое усилие, естественность, понимание смысла сообщения, темп, разборчивость, приятность голоса);
- разборчивость синтезируемой речи должна быть не менее 85 %.

Следующей задачей улучшения качества синтезируемой речи является устранение графической омонимии при транскрибировании чеченских текстов.

Приведем основные особенности чеченской орфографии, воздействие которых на автоматически синтезируемую речь представляется наиболее значимым:

1) Омографы, которые различимы только по контексту в процессе чтения всего предложения.
2) Дифтонги и трифтонги, обозначающиеся одной/двумя буквами, и отсутствие четко прописанных правил для их чтения.
3) Долгота гласных, которая никак не обозначается на письме.

Особенно сложный класс признаков — омографы, многозначные слова, произносимые по-разному в зависимости от контекста.

В чеченском языке омографы различаются по долготе/краткости гласных фонем. Долгота гласных фонем является фонематичной, а вот ударение в чеченском языке фиксированное, оно падает на первый слог в слове. То есть ударение не является фонематичным, в отличие от русского языка, в котором главный отличительный признак омографов — это ударение.

Так как долгота гласных на письме не обозначается, возникает проблема правильного транскрибирования гласных. Приведем примеры омографов, различаемых по долготе гласных фонем: **/бажа/** — **[baaja]** — свояк и **/бажа/** — **[baja]** — стадо; **/шу/** — **[shu]** — вы и **/шу/** — **[shuu]** — холм, возвышенность.

Исследования в области устранения графической омонимии начаты. Для решения этой проблемы существует три основных подхода:

- основанный на правилах;
- основанный на статистике;
- основанный на машинном обучении.

Мы решили остановиться на гибридном методе устранения графической омонимии, основанном на использовании машинного обучения и статистики. Для чеченского языка этот подход кажется нам наиболее оптимальным, так как при чтении чеченских текстов читателю приходится анализировать контекст, чтобы правильно произнести слова-омографы. Статистический метод направлен на анализ контекста омографов по следующим признакам:

- позиция в предложении;
- наличие послелогов;
- контекст слова справа и слева.

При помощи созданной в нашем отделе программы DoshStat [5], предназначенной для анализа и исследования частоты слов, был проведен статистический анализ с целью выявления омографов, наиболее часто встречающихся в текстах различных стилей. Эта же программа позволяет извлекать из текстов предложения, содержащие указанные омографы. Таким образом подготовлена база данных, состоящая из текстовых файлов, в которых для каждого омографа представлено более 100 предложений, содержащих эти омографы в различных контекстах. В каждом предложении определены уникальные теги — идентификаторы омографов.

Неоднозначные примеры, которые вызывали затруднения при идентификации омографов, были отброшены.

Далее мы планируем создать алгоритм и программную реализацию нейронной сети, которая после обучения на подготовленной базе, будет осуществлять классификацию омографов и их выявление в тексте.

После решения перечисленных выше задач по созданию более качественного и длительного корпуса речи, а также устранения графической омонимии при транскрибировании чеченских текстов, будет запущено повторное обучение на архитектуре DCTTS с целью оптимизации созданной системы синтеза чеченской речи.

СПИСОК ЛИТЕРАТУРЫ

1. *Израилова Э. С.* «Фонетический алфавит» чеченского языка как основа системы синтеза речи // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2018. № 2. С. 35–39.

2. *Израилова Э. С.* Особенности машинного обучения средствами CNN в рамках синтеза речи // Вестник ГГНТУ. Технические науки. 2019. Т. XV. № 2 (16). С. 29–35.

3. *Карпов А. А., Верходанова В. О.* Речевые технологии для малоресурсных языков мира // Вопросы языкознания. 2015. № 2. С. 117–135.

4. *Тягунов Д. В.* Обзор существующих методов синтеза речи по печатному тексту // Научный вестник Черновицкого университета. Физика. Электроника. 2008. Вып. 423. С. 138–142.

5. *Умархаджиев С. М., Бекаев М. Х., Бадаева А. С., Израилова Э. С., Султанов З. А.* и др. DoshStat. Свидетельство о регистрации программы для ЭВМ RUS 2018617362. Дата регистрации: 11.05.2018.

6. *Халидов А. И., Тимаев А. Д., Овхадов М. Р.* Грамматика чеченского языка. Т. 1. Введение в грамматику. Фонетика. Морфемика. Словообразование. Грозный: ФГУП ИПК «Грозненский рабочий», 2013. 848 с.

7. *Arik S. O, Chrzanowski M., Coates A. et al*. Deep voice: Real-time neural text-to-speech // Proceedings of the 34th International Conference on Machine Learning (PMLR). 2017. Vol. 70. P. 195–204.

8. *Cho K., van Merriënboer B., Gulcehre C. et al.* Learning phrase representations using RNN encoder-decoder for statistical machine translation // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha: Association for Computational Linguistics, 2014. P. 1724–1734.

9. *Sotelo J., Mehri S., Kumar K., Santos J. F., Kastner K*. et al. Char2wav: End-to-end speech synthesis // Proceedings of ICLR, 2017. [Online]. URL: https://mila.quebec/wp-content/uploads/2017/02/end-end-speech.pdf (accessed 12.09.2018).

10. *Tachibana H., Uenoyama K., Aihara S.* Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. [Online]. URL: https://arxiv.org/pdf/1710.08969.pdf (accessed 19.11.2018).

11. *van den Oord A., Dieleman S., Zen H.* WaveNet: A Generative model for raw audio. [Online]. URL: https://deepmind.com/blog/wavenet-generative-model-raw-audio/ (accessed 12.09.2018).