УДК 004.021
DOI 10.17513/snt.39693

ВОПРОСЫ РЕЧЕВОЙ ОБРАБОТКИ ДАННЫХ В ПРОГРАММНОМ ОБЕСПЕЧЕНИИ

Крюков Д.А., Оруджов С.С.

ФГБОУ ВО «МИРЭА – Российский технологический университет», Москва, e-mail: dk@memfis.su, sekhran99@mail.ru

Данная статья посвящена вопросам применения технологии речевой обработки данных в программном обеспечении. В статье описан процесс распознавания речи диктора с применением принципа скрытой марковской модели и нейронных сетей, задачами которых являются распознавание фонем, слов и анализ их смыслового содержания в предложении. Рассмотрена скрытая марковская модель в дискретных системах. Описан процесс обучения нейронной сети. Проведён сравнительный анализ совместного и раздельного применения скрытой марковской модели и нейронной сети. Предложен метод подсчёта количества ключевых слов в речи диктора и проведены испытания для оценки эффективности речевой обработки данных. Показаны графики зависимости времени подсчёта ключевых слов в видеоматериалах до и после преобразования аудиодорожек в текст при использовании библиотеки Speech Recognition. Результаты проведённых испытаний показали необходимость предварительной обработки аудиодорожек всех видео для быстрого поиска ключевых слов в речи. Практическая значимость разработки программного обеспечения с применением речевой обработки данных заключается в том, что данная технология может быть результативно использована, например, для поиска видеоматериалов, содержащих ключевые слова. Опыт разработки и тестирования программного обеспечения с реализацией технологии речевой обработки данных может быть также адаптирован для решения иных прикладных задач, связанных с распознаванием речи.

Ключевые слова: речевая обработка данных, тестирование, программное обеспечение, распознавание речи, скрытая марковская модель, нейронная сеть

QUESTIONS OF SPEECH DATA PROCESSING IN SOFTWARE

Kryukov D.A., Orudjov S.S.

Federal State Educational Institution of Higher Education MIREA – Russian Technological University, Moscow, e-mail: dk@memfis.su, sekhran99@mail.ru

This article is devoted to the application of speech data processing technology in software. The article describes the process of recognizing the speaker's speech using the principle of a hidden Markov model and neural networks, whose tasks are to recognize phonemes, words and analyze their semantic content in a sentence. The hidden Markov model in discrete systems is considered. The process of training a neural network is described. A comparative analysis of the joint and separate application of a hidden Markov model and a neural network has been carried out. A method for counting the number of keywords in the speaker's speech is proposed and tests are carried out to evaluate the effectiveness of speech data processing. Graphs of the dependence of the time of counting keywords in video materials before and after converting audio tracks to text using the speech recognition library are shown. The results of the tests showed the need for pre-processing the audio tracks of all videos to quickly search for keywords in speech. The practical significance of software development using speech data processing lies in the fact that this technology can be effectively used, for example, to search for video materials containing keywords. The experience of developing and testing software with the implementation of speech data processing technology can also be adapted to solve other applied problems related to speech recognition.

Keywords: speech data processing, testing, software, speech recognition, hidden Markov model, neural network

Технология речевой обработки данных всё чаще применяется при разработке программного обеспечения. Программная реализация алгоритмов, задействовавших данную технологию, имеет свои преимущества, равно как и недостатки. В связи с чем для оценки эффективности распознавания речи диктора необходимо проведение соответствующих испытаний.

Существует несколько систем преобразования речи в текст, использующих различные подходы для достижения высокой точности распознавания речи. Одной из таких систем является система, основанная на скрытых марковских моделях (далее – СММ), использующих математическую модель, определяющую вероятность перехода между фонемами в речевом сигнале. Данная система по сравнению с более новыми системами имеет более низкую точность распознавания речи. Также это может быть система, основанная на нейронных сетях: используются алгоритмы обучения для обработки звуковых сигналов и преобразования их в текст. Они показывают более высокую точность распознавания речи по сравнению с СММ. Используется также гибридная система – применение СММ в комбинации с нейронной сетью для достижения высокой точности распознава-

ния речи. А также применяются облачные сервисы и библиотеки с готовыми решениями, которые могут быть интегрированы в различные приложения и платформы, что является удобным способом использования систем распознавания речи без необходимости разработки своих собственных моделей.

На примере разработки программного обеспечения будет рассмотрено применение, а также тестирование алгоритмов обработки речи диктора при помощи библиотеки Speech Recognition, поскольку обучение нейронной сети совместно с СММ является более сложным и трудоёмким процессом, требующим значительного количества вычислительного времени. Совместное применение СММ и нейронной сети требует мощной конфигурации ПК с достаточным объёмом памяти и вычислительной мощности для поддержки выполнения поставленных задач – данная гибридная система может быть рассмотрена в будущем.

Целью тестирования и разработки программного обеспечения является определение преимуществ и недостатков при использовании библиотеки Speech Recognition для распознавания речи.

Процесс распознавания речи диктора

Рассмотрим и определим, каким образом речь человека преобразовывается в понимаемую цифровым устройством совокупность информации.

Звук, издаваемый диктором в микрофон, является аналоговым. Микрофон содержит в себе магнит, намотанный на катушку из проволоки, который во время вибрации при помощи электромагнитной индукции создаёт электрический ток в проволоке. Далее амплитуды преобразуются в напряжение, которое может быть распознано цифровым устройством, а отдельные частоты изолируются друг от друга. Результат представляется на спектрограмме.

Каждый язык обладает фонетической библиотекой, содержащей в себе звуки, которые используются в разговорной речи. Слова составляются из фонем, которые могут быть распознаны на спектрограмме. Фонема является наименьшей воспринимаемой звуковой единицей, которую можно различить среди слов определённого языка [1]. На ранних стадиях многие распознавания фонем проводились статистически. Например, в скрытой марковской модели предполагается, что система представляет собой марковский процесс с неизвестными параметрами, т.е. фонемами, и задача состоит в том, чтобы определить скрытые фонемы из наблюдаемых фонем [2]. Также СММ применяется в распознавании рукописных текстов и в качестве метода для распознавания электрических сигналов в биомедицине [3]. Скрытая марковская модель состоит из скрытых марковских цепей и наблюдаемых переменных.

На основе СММ используется широкий спектр преобразований и шагов выделения признаков для извлечения и нормализации информации, содержащейся в речевом сигнале на входе, с максимально возможной эффективностью [4]. Поскольку каждая фонема должна быть предварительно определена, а речь диктора всегда разная ввиду акцентов и/или неправильных произношений слов, то данный подход не может быть применён к большому количеству вариаций фонем. В связи с этим была представлена альтернатива – нейронная сеть.

Нейронные сети способны постоянно улучшать себя при помощи вводимых данных. Состоит нейронная сеть из взаимосвязанных узлов, где каждая связь между ними имеет собственный вес, который определяет силу сигнала между данными узлами. Узлы, в свою очередь, обычно состоят в так называемых слоях ввода и вывода вдобавок к скрытым слоям, которые выполняют преобразование данных. Каждый раз при получении обратной связи нейронная сеть изучает её и настраивает вес связи между узлами. В отличие от СММ, нейронная сеть требует большого количества данных, которые необходимо настроить для неё. Скрытая марковская модель применяется в комбинации с нейронной сетью при распознавании речи диктора, поскольку каждый из подходов имеет свои недостатки.

Далее рассмотрим на рисунке 1 алгоритм преобразования речи диктора из аудио в текст.

После определения звуков система проводит анализ слов. Найденные фонемы далее обрабатываются так называемым модулем языка, который анализирует, насколько вероятно появление одной фонемы после другой. Далее модуль определяет, содержатся ли выходные данные в словаре языка. Если комбинация фонем довольно маловероятна в любом из стандартов, то выбирается другое слово с похожими фонемами. После того как определяются слова в предложении, языковой модуль проводит синтаксический анализ фразы, чтобы определить, подходит ли она по смыслу в данном контексте: проводится исследование порядка слов в соответствии с правилами грамматики языка, где используется так называемый parse tree – граф, который называется синтаксическим деревом [5].

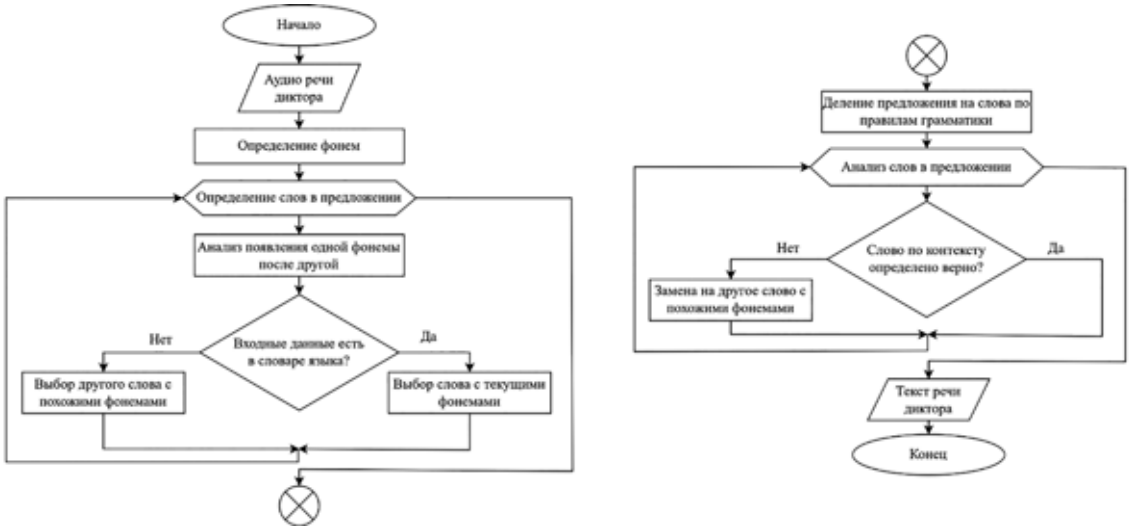

Рис. 1. Алгоритм преобразования аудио в текст

Благодаря синтаксическому дереву предложение разбивается на более мелкие части, пока не остаются одни слова в отдельности. В случае если модуль языка определяет, что слово было выбрано неверно в контексте предложения, производится возврат к фонемам, которые были распознаны ранее, и используются похожие фонемы, чтобы определить, какие слова должны быть задействованы на самом деле. В конечном итоге после доведения слов в предложении до минимальных несоответствий предоставляются выходные данные – речь диктора, преобразованная в текст.

Процесс обучения нейронной сети

Рассмотрим, каким образом проходит обучение нейронной сети для дальнейшего использования в комбинации с СММ. Процесс обучения проходит в несколько этапов. Сначала для обучения производится подготовка данных, состоящих из аудиодорожек, а также текстового представления речи. Далее для каждого аудио извлекается набор признаков, который будет использоваться в качестве входных данных для нейронной сети. Этим признаком может являться спектрограмма. На следующем этапе проходит обучение нейронной сети на подготовленных данных. Обучение заключается в определении оптимальных весов, которые связывают входные данные со знакомой системой отображения. Чаще всего в обучении применяется метод обратного распространения ошибки [6]. Далее строится скрытая марковская модель. Используются те же данные, которые были применены для обучения нейронной сети. Определяется структура СММ и вероятности перехода между фонемами. В конечном итоге СММ объединяется с нейронной сетью. При подаче звукового сигнала на вход нейронной сети производится построение его высокоуровневого представления, которое используется для выбора наиболее вероятной последовательности фонем, соответствующей входному звуковому сигналу.

Итак, сначала нейронная сеть рассчитывает выходные значения, а затем они сравниваются с целевыми значениями. Это приводит к появлению ошибок в выходных значениях, которые могут быть обратно проведены через сеть и учтены при изменении весов связей между нейронами [6]. Процесс обучения нейронной сети заключается в настройке весов связей между нейронами для минимизации ошибок предсказания фонем в аудиодорожках.

Скрытая марковская модель в дискретных системах

В своих трудах [7] Rabiner L.R. привёл основы скрытой марковской модели для дискретных систем. Рассмотрим данные СММ.

Для определения скрытой марковской модели необходимо задать:

– множество состояний СММ – $S = \{s_1,\ldots,s_N\}$, где N – количество состояний в нём;

– алфавит СММ – $V = \{v_1,\ldots,v_M\}$, где M – количество символов в нём;

– q_t – состояние системы в момент времени t, зависящее от её состояния в момент времени $t-1$ – q_{t-1};

– o_t – значение наблюдаемых параметров в момент времени t, зависящее от состояния q_t;

– $Q = \{q_1,\ldots,q_T\}$ – последовательность состояний, в которых бывает система, невидимая наблюдателю, где T – размер последовательности;

– $O = \{o_1,\ldots,o_T\}$ – последовательность состояний, видимая наблюдателю;

– $A = \{a_{ij}\}$ – вероятности переходов между состояниями СММ, где a_{ij} – вероятность перехода модели из состояния s_i в s_j;

– $B = \{b_j(k)\}$ – вероятности выпадения всех значений M в наблюдаемых параметрах СММ в каждом из состояний N, где $b_j(k)$ – вероятность выпадения k-го значения наблюдаемых параметров СММ в состоянии s_j;

– $\pi = \{\pi_i\}$ – вектор, необходимый для определения вероятности появления какого-либо начального состояния СММ, где $\{\pi_i\}$ – вероятность того, что СММ в начальный момент окажется в состоянии s_i [7].

Определим элементы A по следующей формуле:

$$a_{ij} = p\left(q_{t+1} = s_j | q_t = s_i\right),$$

$$1 \le i \le N, 1 \le j \le N. \quad (1)$$

Вероятности перехода СММ из состояния s_i в s_j определены значениями:

$$a_{ij} \ge 0, 1 \le i \le N, 1 \le j \le N,$$

$$\sum_{j=1}^{N} a_{ij} = 1, 1 \le i \le N. \quad (2)$$

Определим элементы B по следующей формуле:

$$b_j\left(k\right) = p\left(o_t = v_k | q_t = s_j\right),$$

$$1 \le j \le N, 1 \le k \le M. \quad (3)$$

Рассмотрим ограничения элементов B:

$$b_j\left(k\right) \ge 0, 1 \le j \le N, 1 \le k \le M$$

$$\sum_{k=1}^{M} b_j\left(k\right) = 1, 1 \le j \le N. \quad (4)$$

Определим вероятность появления какого-либо начального состояния СММ по формуле:

$$\pi_i = p\left(q_1 = s_i\right), 1 \le i \le N. \quad (5)$$

Определим ограничения элементов π:

$$\pi_i \ge 0, 1 \le i \le N \sum_{i=1}^{N} \pi_i = 1, 1 \le i \le N. \quad (6)$$

Итак, для дискретных систем определим скрытую марковскую модель следующим образом:

$$\lambda = (A, B, \pi). \quad (7)$$

Таким образом, в скрытой марковской модели принято рассматривать три задачи для её правильного применения.

Первая задача заключается в том, что необходимо определить вероятность того, что $O = \{o_1,\ldots,o_T\}$ построена именно для данной СММ [7].

Второй задачей является подбор такой $Q = \{q_1,\ldots,q_T\}$, которая лучше смогла бы отразить $O = \{o_1,\ldots,o_T\}$ [7].

Третьей задачей является подбор параметров $\lambda = (A, B, \pi)$ для получения максимальной вероятности в первой задаче [7].

Структура СММ и нейронной сети

Определим структуру модели, которая состоит из СММ и нейронной сети. В контексте распознавания речи СММ используется для моделирования фонем. Каждая фонема соответствует определённой последовательности акустических признаков. Нейронная сеть используется для обработки акустических признаков и построения связи между скрытыми состояниями СММ и наблюдаемыми признаками. Входные данные для нейронной сети – это временные ряды акустических признаков. В данной модели нейронная сеть может быть представлена в виде рекуррентной нейронной сети, которая обрабатывает последовательность входных признаков. Скрытые состояния рекуррентной нейронной сети связываются со скрытыми состояниями СММ, что позволяет описать последовательность звуков и произвести их распознавание. Последний слой нейронной сети позволяет определить, какие фонемы присутствуют в речи, на основе вероятностей, вычисленных из модели СММ. После прохождения через СММ и нейронную сеть на выходе получается текст распознанной речи.

Научная новизна применения СММ в комбинации с нейронной сетью

СММ хорошо подходит для обработки временных зависимостей в речевых сигналах, поскольку моделирует распределение вероятностей наблюдаемого речевого сигнала как последовательность скрытых состояний. Однако у СММ есть ограничения в представлении сложных нелинейных отношений между входными объектами и выходными метками. С другой стороны, нейронная сеть хорошо определяет сложные нелинейные отношения между вход-

ными и выходными данными. Таким образом, можно распознавать и извлекать соответствующие функции из входного речевого сигнала. Однако нейронные сети могут быть не столь эффективны при моделировании временных зависимостей. Совместное использование СММ и нейронной сети, известное как гибридная модель, преодолевает ограничения обеих моделей. Нейронная сеть используется для оценки распределения вероятностей входных признаков с учётом текущего состояния СММ. Таким образом, модель изучает сложные нелинейные отношения между входными объектами и выходными метками, при этом учитывая временные зависимости. Научная новизна совместного использования СММ и нейронной сети заключается в способности повышать точность систем распознавания речи за счёт объединения сильных сторон обеих моделей. Данная гибридная система широко используется в современных системах распознавания речи: Siri от Apple, а также распознавание речи от Google.

Сравнительный анализ совместного и раздельного применений СММ и нейронной сети

Одним из показателей качества, который можно использовать для оценки эффективности совместного использования СММ и нейронной сети в распознавании речи, является Word Error Rate (далее – WER), который измеряет процент ошибок при переводе речи в текст [8].

Для вычисления WER необходимо определить N – общее количество слов, а также количество ошибок: S – замены слов, D – удаления слов, I – вставки слов [8].

Определяется Word Error Rate по следующей формуле:

$$WER = \frac{S + D + I}{N} \times 100. \qquad (8)$$

Для проведения сравнительного анализа эффективности совместного применения СММ и нейронной сети с отдельно используемой СММ можно использовать обучающую выборку, содержащую аудиодорожки и соответствующие им тексты, полученные из речи диктора. В первом случае необходимо обучить нейронную сеть на основе этих данных и создать скрытую марковскую модель для каждого слова в словаре. Затем СММ и нейронная сеть используются совместно для распознавания аудиодорожек и оценивается WER. Во втором случае используется только СММ и также проводится оценка Word Error Rate.

Гибридная система распознавания речи приводит к более высокому уровню точности распознавания речи, поскольку нейронная сеть может распознавать шаблоны, которые скрытая марковская модель может пропустить, в то время как СММ обеспечивает получение нейронной сетью правдоподобных и последовательных результатов.

Применение библиотеки Speech Recognition в программном обеспечении

В рамках программной реализации было принято решение использовать библиотеку для распознавания речи диктора – Speech Recognition. Данная библиотека использует комбинацию алгоритмов машинного обучения и методов обработки сигналов для анализа звуковой волны речи, а также преобразования аудиодорожки в текст.

Сначала библиотека получает на вход информацию с микрофона или аудиофайла, а затем обрабатывает данные для извлечения характеристик: частоты, амплитуды и продолжительности, которые передаются в модель машинного обучения – как правило, в глубокую нейронную сеть, предварительно обученную на большом массиве речевых данных. Модель использует эти данные для изучения закономерностей и взаимосвязей между звуковыми характеристиками и соответствующим текстом, что позволяет ей делать точные прогнозы относительно содержания речи диктора.

Библиотека Speech Recognition также использует различные методы обработки сигналов для повышения качества входных данных перед их передачей в модель машинного обучения. Например, библиотека может выполнить шумоподавление или фильтрацию для удаления любых нежелательных фоновых шумов или искажений аудиосигнала, что может повысить точность транскрибации – перевода аудио в текст.

Библиотека Speech Recognition является эффективным инструментом для распознавания речи диктора, однако, как и в любой модели машинного обучения, её точность может варьироваться в зависимости от качества входных данных, сложности содержимого в речи и т.д.

Рассмотрим применение данной библиотеки на примере поиска видеоматериалов. Во время изложения материала в речи диктора используются слова, по которым слушатели смогут быстрее найти необходимые им видеозаписи. Наличия ключевого слова в видео недостаточно, поскольку упоминание необходимого термина малым количеством раз не означает, что вся видеозапись посвящена именно этой теме. Возникает необходимость подсчёта коли-

чества ключевых слов в речи диктора: чем больше раз ключевые слова были произнесены во время видео, тем выше в списке результатов будет отображаться соответствующая видеозапись слушателям. Проведём оценку эффективности технологии речевой обработки данных при помощи тестирования с применением библиотеки Speech Recognition.

Результаты тестирования с применением библиотеки Speech Recognition

По окончании разработки программного обеспечения было принято решение провести тестирование. Определено количество найденных ключевых слов в видеозаписях, а также время, необходимое на их подсчёт. Оценка времени была произведена до и после преобразования аудиодорожек в текст. Тестирование проводится на примере 20 видеозаписей разной длительности. Рассмотрим результаты тестирования на рисунке 2, где отображается график зависимости времени обработки аудиодорожки видео от его продолжительности.

Исходя из результатов тестирования, можно сделать вывод, что, как правило, при увеличении длительности видеозаписи происходит увеличение времени обработки соответствующей аудиодорожки, которая сначала преобразовывается в текст.

Рис. 2. График зависимости времени обработки аудиодорожки видео от его продолжительности

Рис. 3. График зависимости времени обработки текста аудиодорожки от продолжительности видео

Далее производится подсчёт ключевых слов в речи диктора. Время обработки в данном случае очень высокое, даже для видео длиной до 2-3 минут. Обратим внимание, что на фоне роста времени обработки также присутствуют случаи, в которых данное время намного меньше: некоторые аудиодорожки не смогли быть обработаны, и в них не удалось преобразовать речь диктора в текст. Далее рассмотрим результаты тестирования на рисунке 3, где отображается график зависимости времени обработки текста аудиодорожки от продолжительности видео.

По итогам тестирования предварительное преобразование всех аудиодорожек в текст позволило значительно уменьшить время ожидания для получения результатов. Подсчёт ключевых слов в речи занимает меньше 0,01 секунды. Обратим внимание на те самые аудиодорожки, которые не были преобразованы в текст: время обработки в данном случае составляет 0 секунд, поскольку речи в аудиодорожках не были распознаны данной библиотекой и не были поданы на вход программе для нахождения ключевых слов в них.

В статье рассмотрен один из методов применения речевой обработки данных в программном обеспечении. В основе распознавания речи диктора лежит принцип скрытой марковской модели совместно с нейронными сетями. Применение скрытой марковской модели неэффективно без нейронных сетей: большое количество вариаций фонем не может быть распознано корректно в связи с особенностями языков дикторов. Описана скрытая марковская модель в дискретных системах. Полученные результаты тестирования показали, что, несмотря на возможные проблемы с аудиодорожками, речевая обработка данных помогает ускорить работу пользователей. Решением задачи длительного ожидания результата является предварительное преобразование аудиодорожек в текст со стороны специалистов программной инженерии. Поиск видео по ключевым словам повышает точность результатов и занимает меньше времени для их нахождения слушателями. Графики с результатами тестирования показывают, насколько важно не только внедрить данную технологию, но и корректно определить последовательность процесса обработки аудиозаписей, поскольку от этого зависит дальнейшее время ожидания пользователей. Библиотека Speech Recognition использует предварительно созданные алгоритмы и модели для распознавания речи, в то время как гибридная система, т.е. СММ в комбинации с нейронной сетью, требует создания и обучения пользовательской модели. Данная библиотека распознавания речи представляет собой предварительно созданное решение, требующее минимальной настройки и обучения, а гибридный подход является более сложным процессом, но обеспечивает более высокую точность распознавания речи диктора. Гибридная система (СММ и нейронная сеть) может быть рассмотрена в будущем в связи с высокими требованиями к конфигурации ПК и значительным количеством вычислительного времени в процессе обучения нейронной сети.

Список литературы

1. Phoneme [Электронный ресурс]. URL: https://www.britannica.com/topic/phoneme (дата обращения: 30.06.2023).

2. Franzese M., Iuliano A. Hidden Markov Models. Encyclopedia of Bioinformatics and Computational Biology // Elsevier. 2019. Vol. 1. P. 753-762.

3. Sarmiento C., Savage J. Comparison of Two Objects Classification Techniques using Hidden Markov Models and Convolutional Neural Networks // Informatics and Automation. 2020. Vol. 19, Is. 6. P. 1222-1254. DOI: 10.15622/ia.2020.19.6.4.

4. Alam M.J., Gupta V., Kenny P., Dumouchel P. Speech recognition in reverberant and noisy environments employing multiple feature extractors and i-vector speaker adaptation // EURASIP Journal on Advances in Signal Processing. 2015. Vol. 2015, Is. 1. P. 1-13. DOI: 10.1186/s13634-015-0238-6.

5. Cooper K. D., Torczon L. Chapter 3 – Parsers. Engineering a Compiler (Third Edition) // Elsevier. 2022. P. 85-157.

6. Guo Y., Chen J., Du Q., Van Den Hengel A., Shi Q., Tan M. Multi-way backpropagation for training compact deep neural networks // Neural Networks. Elsevier. 2020. Vol. 126. P. 250-261.

7. Rabiner L.R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition // IEEE. 1989. Vol. 77. P. 257-286.

8. The Problem with Word Error Rate (WER) [Электронный ресурс]. URL: https://www.speechmatics.com/company/articles-and-news/the-problem-with-word-error-rate-wer (дата обращения: 19.06.2023).