http://dx.doi.org/10.35596/1729-7648-2021-19-6-51-58

Оригинальная статья
Original paper

УДК 004.352.243+004.934.5

АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ И ПРЕДСТАВЛЕНИЕ ТЕКСТА В ВИДЕ АУДИОПОТОКА

Л.В. СЕРЕБРЯНАЯ, И.Е. ЛАСЫЙ

Белорусский государственный университет информатики и радиоэлектроники
(г. Минск, Республика Беларусь)

Поступила в редакцию 3 мая 2021

Аннотация. Рассмотрена задача автоматической генерации речи из текстового файла. Выполнен анализ программных средств, предназначенных для распознавания текстов и преобразования их в аудиопоток. Оценены их преимущества и недостатки, на основании чего сделан вывод об актуальности разработки программного средства автоматической генерации аудиопотока из текста на русском языке. Проанализированы модели на основе искусственных нейронных сетей, которые используются для синтеза речи, после чего построена математическая модель создаваемого программного средства. Она состоит из трех компонентов: сверточного кодировщика, сверточного декодировщика и преобразователя. Спроектирована архитектура программного средства, в которую входят графический интерфейс, сервер приложения и система синтеза речи. Разработан ряд алгоритмов: предварительной обработки текста перед загрузкой в программное средство, преобразования аудиофайлов обучающей выборки и обучения сети, генерации речи на основе произвольных текстовых файлов. Создано программное средство, представляющее собой одностраничное приложение и имеющее веб-интерфейс для взаимодействия с пользователем. Для оценки качества работы программного средства использована метрика, представляющая среднюю оценку разных мнений. В результате агрегации разных мнений метрика получила достаточно высокое значение, на основании чего можно считать, что все поставленные задачи были решены.

Ключевые слова: модель искусственной нейронной сети, аудиопоток, кодировщик и декодировщик, генерация речи, спектрограмма.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Для цитирования. Серебряная Л.В., Ласый И.Е. Автоматическое распознавание и представление текста в виде аудиопотока. Доклады БГУИР. 2021; 19(6): 51-58.

AUTOMATIC RECOGNITION AND REPRESENTATION OF TEXT IN THE FORM OF AUDIO STREAM

LIYA V. SEREBRYANAYA, ILYA E. LASY

Belarusian State University of Informatics and Radioelectronics (Minsk, Republic of Belarus)

Submitted 3 May 2021

Abstract. The problem of automatic speech generation from a text file is considered. An analytical review of the software has been completed. They are designed to recognize texts and convert them to an audio stream. The advantages and disadvantages of software products are estimated. Based on this, a conclusion was drawn about the relevance of developing a software for automatic generation of an audio stream from a text in Russian. Models based on artificial neural networks, which are used for speech synthesis, are analyzed. After that, a mathematical model of the created software is built. It consists of three components: a convolutional encoder, a convolutional decoder, and a transformer. The architecture of the software is designed. It includes a graphical interface, an application server, and a speech synthesis system. A number of algorithms have been developed: preprocessing text before loading it into a software, converting audio files of a training sample and training a network, generating speech based on arbitrary text files. A software has been created, which is a single-page application and has a web interface for interacting with the user. To assess the quality of the software, a metric was used that represents the average score of different opinions. As a result of the aggregation of different opinions, the metric received a sufficiently high value, on the basis of which it can be assumed that all the tasks have been solved.

Keywords: artificial neural network model, audio stream, encoder and decoder, speech generation, spectrogram.

Conflict of interests. The authors declare no conflict of interests.

For citation. Serebryanaya L.V., Lasy I.E. Automatic recognition and representation of text in the form of audio stream. Doklady BGUIR. 2021; 19(6): 51-58.

Введение

В современных условиях, когда за короткое время требуется воспринять и обработать большие объемы информации, возрастает значимость аудиоинформации. Ее восприятие может осуществляться параллельно с другими занятиями, не требует такой концентрации внимания, как чтение, снижает нагрузку на зрение, а также может выполняться слабовидящими людьми или детьми, не умеющими читать. Поэтому в последнее время возросла популярность аудиокниг и подкастов. Однако не каждый текст имеет аналог в аудиоформате. В таком случае аудиоинформацию можно получить путем генерации аудиопотока из интересующего текста [1]. Этим занимаются специальные TTS (text-to-speech)-системы, получающие на вход файлы с текстом, а выдающие – аудиофайлы. И хотя подобные системы существуют уже довольно давно, сгенерированная ими речь до сих пор уступает по качеству естественной речи.

Развитие технологий машинного и глубокого обучения, в частности искусственных нейронных сетей (ИНС), позволило создать методы и алгоритмы, на основе которых синтезируются аудиопотоки, максимально приближенные к естественной речи [2, 3]. В основном это аудиоинформация на английском языке. Цель данной работы – создание программного средства (ПС), генерирующего из текстовой информации аудиопоток на русском языке.

Архитектура и принципы функционирования сверточных и рекуррентных ИНС способствовали тому, что данные сети чаще других используются в системах синтеза речи. Для оценки качества синтезируемой речи применяется метрика – средняя оценка мнений, она получается в результате статистической обработки большого числа мнений слушателей-экспертов. Эта же метрика использовалась для оценки качества передачи звука в телефонных сетях.

Анализ ПС по теме исследования, таких как Speechify, Voicedream Reader, NaturalReader, позволил оценить их преимущества и недостатки, а также сделать вывод о том, что создание программного приложения, предназначенного для автоматической генерации аудиопотока из текста, является актуальной задачей. К основным функциям создаваемого ПС можно отнести: возможность работы с текстовыми файлами в популярных форматах; синтез и воспроизведение речи на русском языке; возможность регулировать скорость и высоту генерируемого звука; возможность прерывания и возобновления прослушивания. Кроме того, значение метрики для оценки качества звука должно быть не меньше трех баллов по принятой шкале оценок.

Математическая модель программного средства

Синтез речи из текста можно рассматривать как модель «последовательность – последовательность», где в качестве входной последовательности используется текст, а в качестве выходной – спектрограмма аудиосигнала. Подобные модели состоят из двух частей: кодировщика и декодировщика. Первая из частей переводит входной сигнал в его представление в векторном виде $h=g(x)$, а вторая восстанавливает сигнал по его коду $x=f(h)$ [1].

Данная модель может быть основана на сверточных или рекуррентых ИНС, а также на их комбинациях [4]. Существенным недостатком модели «кодировщик – декодировщик» является то, что кодировщик передает на вход декодировщика только последнее состояние после этапа кодирования. Это не позволяет декодировщику анализировать более длинные зависимости внутри входной последовательности. Для устранения названного недостатка используется механизм внимания, сообщающий сети, на что необходимо обратить большее внимание. Механизм формирует матрицу весов «важности». При обучении сети «важность» становится функцией вероятности того или иного исхода в зависимости от поступивших на вход данных.

Декодировщик получает состояния со всех этапов кодировки. При этом механизм внимания назначает оценку каждому состоянию. В ходе умножения каждого состояния на преобразованную функцией *softmax* оценку определяются более и менее важные состояния. Весь процесс работы механизма внимания описывается формулой

$$A(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V, \quad (1)$$

где Q – матрица запроса; K – матрица ключей; V – матрица значений; d – взвешивающий множитель; T – длина последовательности; *softmax* – логистическая функция.

В результате механизм внимания значительно повышает производительность работы сети как во время ее обучения, так и в ходе основной работы. На рис. 1 приведена модель сети, используемая в работе.

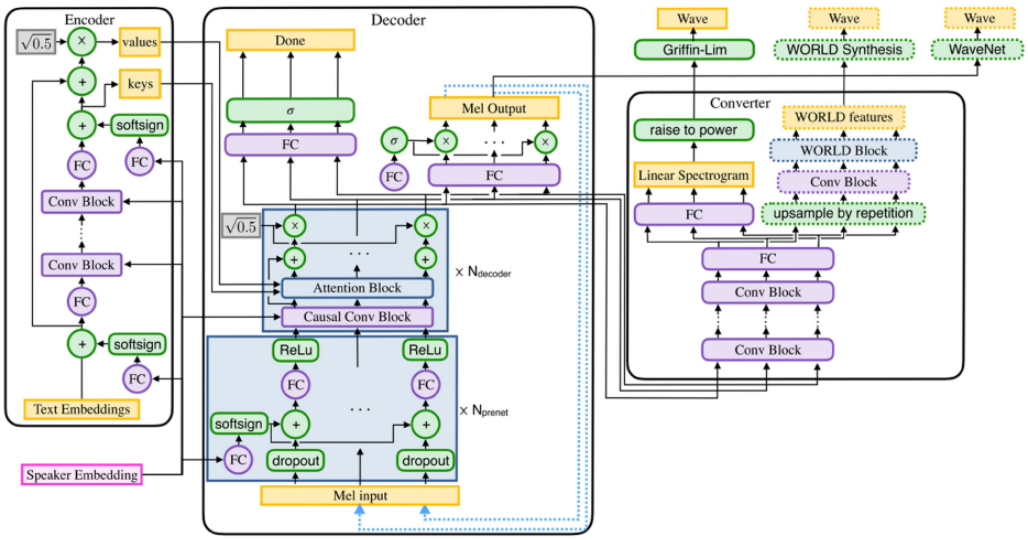

Рис. 1. Модель сети для синтеза речи
Fig. 1. Network model for speech synthesis

Модель сети состоит из трех компонентов:

– сверточного кодировщика, конвертирующего признаки текста во внутреннее векторное представление;

– сверточного декодировщика с механизмом внимания, преобразующего внутреннее представление в мел-спектограмму;

– преобразователя, предсказывающего параметры вокодера, основываясь на состояниях декодировшика.

Кодировщик начинается со слоя вложения, который преобразует символы входного текста в векторное представление h_e. Для извлечения из текстовой информации имеющихся в ней зависимостей вложения сначала проходят через полносвязный слой, а затем через ряд сверточных блоков. Они состоят из сверточного фильтра, вентильной линейной единицы, остаточной связи с входными данными и взвешивающим множителем, равным $\sqrt{0,5}$.

На последнем шаге кодирования векторы проецируются на размерность вложений, чтобы создать векторы ключей h_k для механизма внимания. Векторы значений h_v рассчитываются на основании текстовых сложений и векторов ключей по формуле

$$h_v = \sqrt{0,5}(h_k + h_e). \quad (2)$$

Декодировщик начинается с нескольких полносвязных слоев с пороговой функцией активации, чтобы предобработать входящие мел-спектрограммы обучающей выборки. Затем следуют несколько сверточных блоков. Они генерируют запросы для передачи в блоки механизма внимания. И наконец, полносвязный слой выводит очередную группу спектрограмм, а также бинарное сообщение, указывающее на то, был ли обработан последний отрезок аудио.

Проектирование приложения

Создаваемое ПС имеет модульную структуру и состоит из трех основных компонентов: графического интерфейса, сервера приложения и системы синтеза речи (рис. 2). Для их реализации был разработан ряд алгоритмов. Рассмотрим наиболее интересные из них.

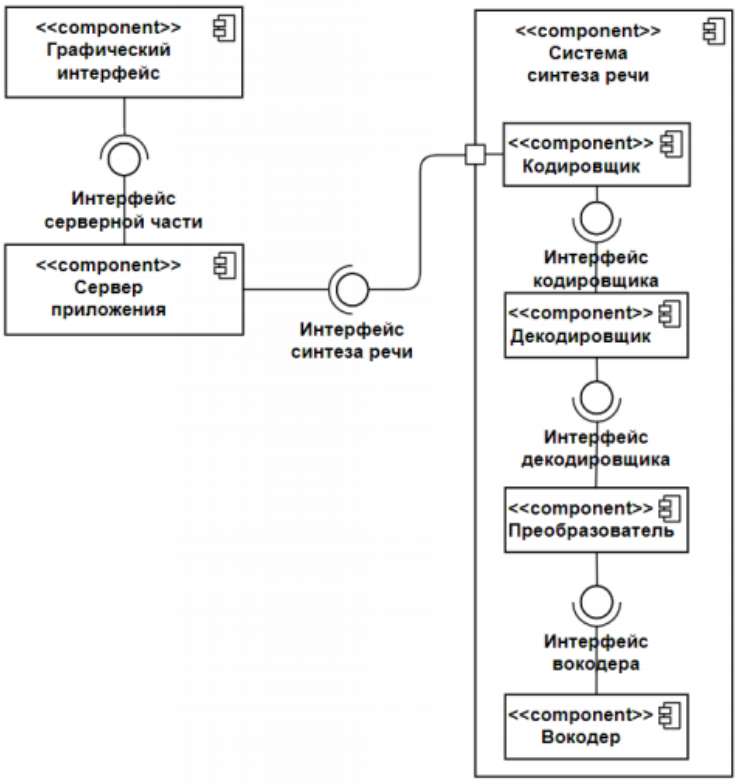

Рис. 2. Структура программного средства
Fig. 2. Software structure

Алгоритм предобработки текста необходим для предварительной обработки текста, загружаемого в ПС. На данном этапе выполняются следующие действия:

– проверка необходимости замены слов исходного текста фонемами, что может улучшить результат работы модели генерации речи;

– удаление знаков препинания, поскольку они являются непроизносимыми, и их наличие будет мешать работе модели генерации речи;

– расшифровка общеупотребляемых аббревиатур русского языка, например, «и т. д.», «т. е.», «и др.», а также замена всех специальных символов и чисел их текстовыми аналогами;

– удаление всех символов, не входящих в список разрешенных символов, остаются только буквы русского алфавита, знаки «точка» и «вопросительный знак»;

– добавление специального символа «~», обозначающего конец предложения, что необходимо модели генерации речи для создания пауз между высказываниями;

– замена всех символов их индексами в списке разрешенных символов, поскольку модель генерации речи может принимать на вход только числовые значения.

Алгоритм преобразования аудиофайлов выполняет предварительную обработку файлов обучающей выборки и содержит выполнение следующих действий:

– удаление «тишины» в начале и в конце аудиофайла, порогом «тишины» является значение, равное 15 дБ;

– при необходимости выполняется масштабирование аудиофайла, когда все значения амплитуд аудиопотока делятся на максимальное по модулю значение амплитуды аудиопотока;

– получение спектров аудиопотока и сохранение их в файл.

Для получения спектрограмм необходимо выполнить ряд действий, приведенных ниже. Сначала слабые и более высокие частоты сигнала усиливаются с целью увеличения отношения сигнала к шуму, минимизируя такие эффекты, как амплитудное искажение и клиппинг. Затем выполняется преобразование Фурье для перехода от представления аудио во временной области в частотную. При этом применяется оконная функция Хэмминга для решения проблемы «растекания спектра». Далее создается матрица мел-фильтра для перевода звука в полосу низких частот, которые лучше воспринимаются человеком. Построенная матрица перемножается со значениями исходного частотного сигнала. В результате преобразований в мел-спектрограмме содержится меньше значений, чем в исходном аудиопотоке, благодаря чему процесс обучения модели происходит быстрее. Финальным преобразованием является перевод полученных значений в децибелы для привычного описания громкости.

После того, как выполнены все подготовительные действия, можно приступать к обучению модели ИНС генерации речи. Перед запуском обучения происходит инициализация сети и, возможно, передача в нее ряда параметров: количество слоев кодировщика, коэффициент скорости обучения и т. д. В противном случае всем параметрам присваиваются значения по умолчанию. Затем в сеть загружаются предобработанные тексты из обучающей выборки и соответствующие им аудиофайлы – мел-спектрограммы.

Когда сеть обучена, она может использоваться для генерации речи на основе произвольных текстовых файлов. Работа начинается с инициализации обученной модели, затем в нее подается текст, который проходит предобработку, аналогичную текстам из обучающей выборки. Затем строится мел-спектрограмма текста, которую получает заранее обученный нейросетевой вокодер WaveNet и генерирует финальную спектрограмму. Она получается очищенной от шумов и максимально похожа на естественную речь. На завершающем этапе аудиопоток сохраняется в файл заданного формата.

Реализация программного средства

Разработанные модели и алгоритмы были положены в основу создаваемого ПС. В качестве основного языка реализации выбран Python, поэтому все последующие решения принимались исходя из этого выбора [5, 6]. В качестве каркаса ПС выбран фрэймворк Python для веб-приложений – Django [7]. ПС для распознавания и преобразования текста в аудиопоток представляет собой одностраничное приложение и имеет веб-интерфейс для взаимодействия с пользователем.

Для предварительной обработки текстов использовались стандартные библиотеки, предназначенные для программирования на языке Python. Использование платформы NLTK (Naturel Language Toolkit) позволило выполнить необходимую обработку текстов. С помощью библиотек NumPy, SciPy, librosa, lws выполнялась обработка аудиофайлов: построение спектрограмм, мел-спектрограмм, преобразования Фурье, перевод из амплитудных значений в децибелы и др.

Для работы с ИНС была выбрана библиотека глубокого обучения Tensorflow, не ограничивающая сложность сети [8, 9]. Для тренировки сети использовалась совокупность аудиокниг на русском языке. Исходные записи разбивались на временные отрезки от 1 до 10 с, суммарная длительность которых составляет приблизительно 20 ч.

В качестве функции потерь для обучения модели выбрана маскированная сумма средней абсолютной ошибки мел-спектрограмм, генерируемых декодировщиком, и средней абсолютной ошибки спектрограмм, генерируемых преобразователем. Наличие маски позволяет исключить данные с нулевыми значениями из расчетов функции потерь.

Результаты и их обсуждение

Для оценки результатов работы созданного ПС, выполняющего перевод текста в речь, была использована средняя оценка мнений группы людей. Им предоставлялись тексты, а также аудиозаписи, сгенерированные по ним. Каждому человеку предлагалось оценить по пятибалльной шкале качество аудиопотока. В результате агрегации полученных значений метрика оказалась равной 3,8. Таким образом, ПС имеет достаточно высокий показатель оценки, на основании чего можно считать, что все поставленные задачи успешно решены.

В ПС реализован ряд функций для синтеза речи на основе текста. Существует возможность ввода текста как с клавиатуры, так и из текстового файла. Имеется поддержка интерфейса для слабовидящих людей.

По завершении ввода текста через непродолжительное время начинается его воспроизведение. Текущее предложение выделено в окне текста, что показано на рис. 3. Во время воспроизведения аудио можно останавливать процесс, перемещаться по тексту, выбирая интересующее предложение. Предложенный подход с небольшими изменениями может быть реализован не только для русского языка, но и для других языков.

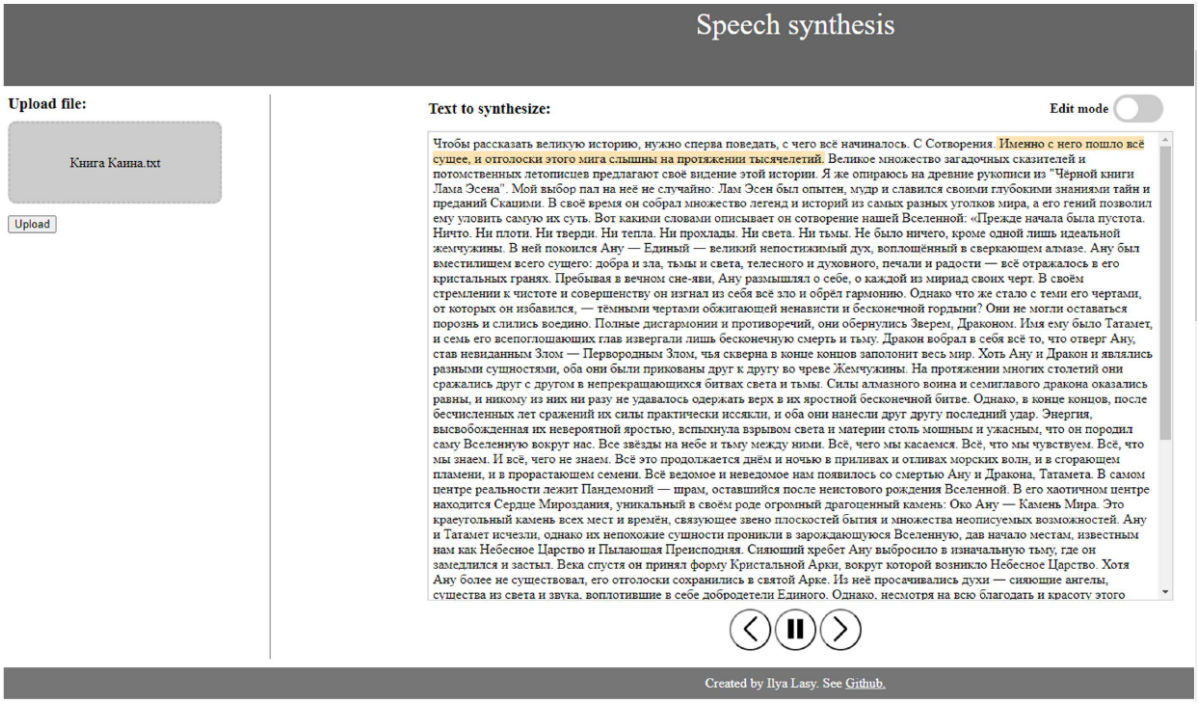

Рис. 3. Окно работы программного средства
Fig 3. Software window

Заключение

В ходе работы были исследованы различные подходы к решению задачи представления текста в виде аудиопотока. Анализ программных продуктов по теме исследования, а также методов, положенных в их основу, показал, что наиболее удачной моделью для преобразования текста в речь является искусственная нейронная сеть.

Была построена математическая модель для распознавания текста и генерации аудиопотока. На ее основе спроектирована архитектура ПС, разработан ряд алгоритмов и программных модулей. Тестирование ПС показало, что оно реализует все необходимые функции. Оценка качества сгенерированного аудиопотока приближает его к уровню естественной речи. Намечены дальнейшие пути развития созданного ПС и расширения его функциональных возможностей.

Список литературы

1. Гольдберг Й. *Нейросетевые методы в обработке естественного языка*. Москва: ДМК-Пресс; 2019.
2. Гудфеллоу Я., Бенджио И., Курвилль А. *Глубокое обучение = Deep Learning*. Москва: ДМК-Пресс; 2017.
3. Николенко С.И., Кадурин А.А., Архангельская Е.О. *Глубокое обучение*. Санкт-Петербург: Питер; 2018.
4. Траск Э. *Грокаем глубокое обучение*. Санкт-Петербург: Питер; 2019.
5. Шолле Ф. *Глубокое обучение на Python*. Санкт-Петербург: Питер; 2018.
6. Элбон К. *Машинное обучение на Python*. *Сборник рецептов*. Санкт-Петербург: BHV; 2019.
7. Меле А. *Django 2 в примерах*. Москва: ДМК-Пресс; 2019.
8. Реза Б.З., Рамсундар Б. *TensorFlow для глубокого обучения*. Санкт-Петербург: BHV; 2019.
9. Ганегедара Т. *Обработка естественного языка с TensorFlow*. Москва: ДМК-Пресс; 2019.

References

1. Goldberg J. [*Neural network methods in natural language processing*]. Moscow: DMK-Press; 2019. (In Russ)
2. Gudfellou Ya., Bendzhio I., Kurvill' A. [*Glubokoye obucheniye = Deep Learning*]. Moscow: DMK-Press; 2017. (In Russ)
3. Nikolenko S.I., Kadurin A.A., Arkhangel'skaya Ye.O. [*Deep Learning*]. St. Petersburg: Piter; 2018. (In Russ)
4. Trask E. [*Grokay deep learning*]. St. Petersburg: Piter; 2019. (In Russ)
5. Scholle F. [*Deep Learning in Python*]. St. Petersburg: Piter; 2018. (In Russ)
6. Elbon K. [*Machine learning in Python. Collection of recipes*]. St. Petersburg: BHV; 2019. (In Russ)
7. Mele A. [*Django 2 in examples*]. Moscow: DMK-Press; 2019. (In Russ)
8. Reza BZ, Ramsundar B. [*TensorFlow for deep learning*]. St. Petersburg: BHV; 2019. (In Russ)
9. Ganegedara T. [*Natural Language Processing with TensorFlow*]. Moscow: DMK-Press; 2019. (In Russ)

Вклад авторов

Серебряная Л.В. сформулировала задачи, которые необходимо было решить в ходе исследований, разработала математическую модель, а также выполняла анализ и интерпретацию полученных результатов.

Ласый И.Е. разработал архитектуру и алгоритмы программного средства автоматического синтеза речи на основе текста, выполнил экспериментальную проверку полученных результатов.

Authors' contribution

Serebryanaya L.V. identified the tasks that needed to be solved during the research, developed a mathematical model and performed analysis and interpretation of the results obtained.

Lasy I.E. developed the architecture and algorithms of a software for automatic speech synthesis based on text, carried out an experimental verification of the results.

Сведения об авторах

Серебряная Л.В., к.т.н., доцент, доцент кафедры программного обеспечения информационных технологий Белорусского государственного университета информатики и радиоэлектроники.

Ласый И.Е., выпускник кафедры программного обеспечения информационных технологий Белорусского государственного университета информатики и радиоэлектроники.

Information about the authors

Serebryanaya L.V., PhD, Associate Professor, Associate Professor at the Information Technologies Software Department of the Belarusian State University of Informatics and Radioelectronics.

Lasy I.E., Graduate of the Information Technologies Software Department of the Belarusian State University of Informatics and Radioelectronics.

Адрес для корреспонденции

220013, Республика Беларусь,
г. Минск, ул. П. Бровки, 6,
Белорусский государственный университет
информатики и радиоэлектроники;
тел. +375-17-293-84-93;
e-mail: L_silver@mail.ru
Серебряная Лия Валентиновна

Address for correspondence

220013, Republic of Belarus,
Minsk, P. Brovka str., 6,
Belarusian State University
of Informatics and Radioelectronics;
tel. +375-17-293-84-93;
e-mail: L_silver@mail.ru
Serebryanaya Liya Valentinovna