ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

УДК 004.522

Владимир Алексеевич Нечаев
ФГБОУ ВО «Ивановский государственный энергетический университет имени В.И. Ленина», аспирант, Россия, Иваново, e-mail: nechaev@gapps.ispu.ru

Сергей Витальевич Косяков
ФГБОУ ВО «Ивановский государственный энергетический университет имени В.И. Ленина», доктор технических наук, профессор, заведующий кафедрой программного обеспечения компьютерных систем, Россия, Иваново, e-mail: ksv@ispu.ru

Метод разработки моделей распознавания речи для использования в информационных системах энергетики

Авторское резюме

Состояние вопроса. В настоящее время при разработке моделей автоматического распознавания речи для специализированных предметных областей, в частности для объектов энергетики, используются архитектуры глубоких нейронных сетей, которые требуют большого объема обучающих данных. При этом модели часто оказываются слабо пригодными для эксплуатации в конкретных информационных системах из-за некачественного распознавания специализированной предметной лексики. Дополнительное обучение моделей в части улучшения их качества в конкретном контексте распознавания наталкивается на сложности получения достаточного объема данных и трудоемкость их разметки. В связи с этим актуальной задачей является создание методов, позволяющих снизить трудоемкость построения прикладных моделей распознавания речи и улучшить их качество при использовании в предметных областях, в частности в области энергетики.
Материалы и методы. Применены методы тематического моделирования текста на основе языковых моделей для адаптации открытых данных. В качестве предобученной модели распознавания речи использована глубокая нейронная сеть. Для обучения использованы наборы данных из открытых источников.
Результаты. Разработан метод создания моделей автоматического распознавания речи для специализированных предметных областей, который включает этап промежуточного обучения лексике предметной области на данных из открытых источников, отобранных с использованием тематического семплирования. На основе метода создана и исследована модель автоматического распознавания речи для объектов энергетики, которая показала более высокие результаты распознавания, чем модели, полученные традиционными способами.
Выводы. Апробация предложенного метода подтвердила его эффективность. Разработанная на основе метода прикладная нейросетевая модель продемонстрировала возможность работы в информационных системах объектов энергетики на русском и английском языках без дополнительного обучения на закрытых данных.

Ключевые слова: модели распознавания речи, машинное обучение, методы тематического моделирования текста, нейронная сеть, языковая модель

Вестник ИГЭУ, 2023, вып. 4, с. 94–100.

Vladimir Alekseevich Nechaev
Ivanovo State Power Engineering University, Postgraduate Student, Russia, Ivanovo, e-mail: nechaev@gapps.ispu.ru

Sergey Vitalyevich Kosyakov
Ivanovo State Power Engineering University, Doctor of Engineering Sciences, Professor, Head of Computer Systems Software Department, Russia, Ivanovo, e-mail: ksv@ispu.ru

Development of automatic speech recognition model for energy facilities

Abstract

Background. Currently, when developing automatic speech recognition models for specialized subject areas, in particular for energy facilities, deep neural network architectures are used, which require a large amount of training data. At the same time, models often turn out to be poorly suitable for use in specific information systems due to poor-quality recognition of highly specialized subject vocabulary. Additional training of models to improve their quality in a specific context of recognition encounters the difficulty to obtain a sufficient amount of data and the laboriousness of their markup. Thus, an urgent task is to create methods that allow reducing the complexity of developing applied speech recognition models and improving their quality when used in subject areas, in particular, in the field of energy.
Materials and methods. Methods of thematic text modeling based on language models for adapting open data are applied. A deep neural network is used as a pretrained speech recognition model. For training, open-source datasets are used.
Results. A method to develop automatic speech recognition models for specialized subject areas has been developed. It includes the stage of intermediate learning of subject area vocabulary based on open-source data selected using thematic sampling. Based on the method, the authors have developed and studied a model of automatic speech recognition for energy facilities. It has showed higher recognition results than models obtained by traditional methods.
Conclusions. Approbation of the proposed method has confirmed its effectiveness. The applied neural network model developed on the method has demonstrated the possibility to work in the information systems of energy facilities in Russian and English without additional training on proprietary data.

Key words: automatic speech recognition models, machine learning, thematic modelling methods, neural network, language model

DOI: 10.17588/2072-2672.2023.4.094-100

Введение. Автоматическое распознавание речи (англ. *Automatic Speech Recognition* – ASR) является одной из ключевых технологий человеко-машинного взаимодействия, позволяющей сократить время от команды, данной человеком, до выполнения ее машиной. Задача таких систем состоит в преобразовании речи в текст, который далее обрабатывается с помощью моделей для естественного языка (англ. *Natural Language Processing*). В настоящее время наилучшее качество распознавания показывают глубокие нейросетевые модели машинного обучения (англ. *Deep Neural Networks*) на основе архитектуры Transformer и механизма внимания [1, 2], которые напрямую переводят речь в последовательность токенов («сквозные» End-to-End модели) [3].

Технология распознавания речи может применяться на объектах энергетического сектора в САПР и на производстве [4], при распознавании телефонных разговоров в ситуационных центрах и службах поддержки пользователей [5]. Применение автоматического распознавания речи в энергетическом секторе позволяет оптимизировать множество процессов, улучшить обслуживание клиентов и ускорить реакцию на происшествия. С помощью технологии распознавания речи можно вводить данные голосом для выполнения команд, что упрощает работу операторов и позволяет ускорить поиск необходимой справочной информации в базах данных. В службах поддержки пользователей и ситуационных центрах распознавание речи может использоваться для автоматизации и оптимизации процессов обработки запросов, что позволяет повысить эффективность работы человека-оператора.

Особенность современных архитектур распознавания речи заключается в том, что для тренировки и тестирования требуется большой объем качественных данных, размеченных человеком вручную [6]. Сбор данных представляет собой запись речи и ее транскрипцию. При этом данные для тренировки должны быть диверсифицированными. Они должны содержать специфичную для данной предметной области лексику. Это условие составляет существенную проблему: на практике, например на объектах энергетического сектора, таких данных обычно недостаточно либо доступ к ним ограничен на основании политик информационной безопасности и конфиденциальности. Часто размеченная речь отсутствует вовсе. Как следствие, при использовании систем распознавания речи страдает качество распознавания узкоспециализированной лексики и терминов предметной области, что часто делает модель практически непригодной для использования.

В настоящее время для решения этой проблемы применяются различные комбинации

методов, расширяющих возможности использования общеязыковых (не ориентированных на конкретную предметную область) моделей распознавания речи. Примерами таких подходов являются постобработка распознанного текста и генерация данных для дополнительного обучения с использованием синтезаторов речи [7].

Применение методов постобработки распознанного текста ограничивает возможность использования моделей в реальном времени, так как снижается общая производительность системы и увеличиваются накладные расходы на вычисления. Также зачастую постобработка не может быть применена из-за большой фонетической вариативности при распознавании лексики предметной области.

Использование методов генерации данных путем синтеза речи для увеличения объема тренировочных данных пока не позволяет получать устойчивые модели, а стоимость доступа к более продвинутым технологиям генерации речи превышает стоимость сбора и разметки данных классическим способом, т. е. с помощью человека.

Целью настоящего исследования является разработка метода, позволяющего снизить затраты на получение размеченных тематических данных для обучения узкоспециализированной лексике, и изучение возможности применения этого метода при создании модели распознавания речи в предметных областях, связанных с энергетикой.

Методы исследования. Традиционный подход к созданию тематических моделей распознавания речи предполагает адаптацию общеязыковой модели к особенностям лексики предметной области путем ее дополнительного обучения на данных предметной области. В открытом доступе находится большое количество общих языковых данных для распознавания речи общей продолжительностью более тридцати тысяч часов. Если применять эти наборы для обучения моделей распознавания напрямую без адаптации, то возможно получить только общую модель, не адаптированную к предметной области.

Как уже отмечалось, именно получение достаточного объема качественных размеченных данных для обучения узкоспециализированной лексике составляет существенную трудность и требует высоких затрат. Для решения проблемы сложности этапа обучения модели на тематических данных при создании прикладной информационной системы предложено разделить процесс дополнительного обучения языковой модели на два этапа.

1. На первом этапе необходимо провести анализ и обработку доступных открытых данных, размеченных для обучения. Выбрать из них данные, содержащие лексику предметной области, и провести обучение общеязыковой модели на этих данных.

2. На втором этапе оценивается качество полученной модели. В случае необходимости создается ограниченный набор дополнительных данных вручную и проводится дообучение модели на этих данных.

При таком подходе первый этап может быть автоматизирован и не требует больших затрат на реализацию. При этом он обеспечивает возможность распознавания основной части тематической лексики и позволяет повысить качество распознавания за счет возможности получения достаточного объема и качества обучающих выборок. Реализация второго этапа в этом случае требует значительно меньших затрат данных. В итоге общая продолжительность и сложность процесса получения специализированной модели снижается, а ее качество увеличивается.

Для реализации предложенного метода необходимо разработать средства обработки открытых наборов данных и извлечения из них тематических данных. Открытые наборы данных представляют собой коллекции одноканальных аудиофайлов длительностью до 30 секунд, каждому из которых соответствует транскрипция – текстовый файл, написанный человеком вручную. Каждый аудио и текстовый файл из открытых наборов данных – это одно или несколько коротких предложений, являющихся частью диалога, озвученного объявления, аудиокниги, новости, подкаста и т.д. Используемые модели распознавания речи учитывают контекст и корректируют вывод только на промежутках до 30 секунд, поэтому важно отбирать данные на уровне фрагментов соответствующей продолжительности.

В рамках проведенного исследования для создания модели на русском языке был использован корпус данных Russian Open Speech To Text [8] общей продолжительностью 20000 часов. Для английского языка за основу взяты следующие наборы данных общей продолжительностью 15000 часов: Gigaspeech [9], LibriSpeech [10], Common Voice [11], Multilingual LibriSpeech [12], M-AILABS [13], mTEDx [14], VoxPopuli [15], VoxForge [16].

Процесс построения тематической модели выполнялся на текстовой части всех данных для каждого языка следующим образом:

1. С помощью языковой модели MPNet [17] были получены векторные представления каждого текстового файла в многомерном пространстве размерности 768.

2. С помощью метода равномерной аппроксимации многообразия и проекции для уменьшения размерности UMAP [18] векторные представления документов были переведены в пространство пониженной 10-мерной размерности (минимальное количество соседних точек

выборки – 15; в качестве метрики расстояния выбрана косинусная мера).

3. Используя метод кластеризации на основе иерархической плотности HDBSCAN [19], было выполнено преобразование:

$$T = f_T\left(f_r\left(f_v\left(D\right)\right)\right), \quad (1)$$

где T – множество тематических кластеров (категорий) текстовых документов; f_T – функция кластеризации; f_r – функция снижения размерности; f_v – функция извлечения семантического вектора (англ. embedding); D – множество текстовых файлов с ручной разметкой (документов).

В результате было получено 520 тематических кластеров (категорий) для русского языка и 543 аналогичных кластера для английского языка. Каждая категория описывается набором из 10 наиболее характерных слов и словосочетаний (токенов), которые определяются с помощью TF-IDF меры на основе классов [20]. Далее вручную были отобраны категории, которые относятся к энергетике, транспорту и информационным технологиям. Формально это можно описать выражением

$$W_{\text{отб}} \subseteq f_{c-TF-IDF}\left(T, D\right) \Rightarrow D_{\text{отб}} \subseteq D, \quad (2)$$

где $W_{\text{отб}}$ – множество отобранных кластеров; $f_{c-TF-IDF}$ – функция TF-IDF на основе классов от множества тематических кластеров T и текстовых файлов D; $D_{\text{отб}}$ – множество отобранных документов.

Примеры описаний отобранных кластеров содержат следующие токены (на английском языке): energy, power, solar, battery, charger, plant, oil, carbon, technology, digital, data, computer и др. Для русского языка отобранные кластеры описываются аналогичными токенами. В результате для английского языка отобрано: 31 кластер, что соответствует 250 тысячам аудиофайлов с транскрипциями из открытых наборов данных общей продолжительностью 320 часов. Для русского языка: 36 кластеров, 280 тысяч аудиофайлов продолжительностью 358 часов. Данные разделены на тренировочную и тестовые подвыборки в соотношении 9 к 1 соответственно.

4. Используя библиотеку с открытым исходным кодом NVIDIA NeMo [21] и данные, полученные с помощью описанного способа тематического семплирования, была произведена тонкая настройка (англ. fine-tuning) моделей распознавания речи. В качестве архитектур выбраны глубокие нейронные сети Citrinet [22] и Conformer-CTC [23]. Инициализация произведена с весов общего назначения указанных архитектур для русского и английского языков. В процессе тонкой настройки применялись разные скорости обучения (англ. learning rate) от 0,01 до 0,1. Обучение производилось на сервере с четырьмя графическими ускорителями (GPU) NVIDIA Tesla V100 в течение 24 часов для архитектуры Citrinet и 48 часов для архитектуры Conformer-CTC для каждого языка.

В итоге были получены модели распознавания речи на английском и русском языках, которые ориентированы на применение в информационных системах энергетики.

Исследование разработанной модели. Для оценки качества полученных моделей распознавания речи были подготовлены тестовые наборы данных. Эти данные получены на производственных объектах и представляют собой записи телефонных разговоров сотрудников колл-центра с пользователями, состоят из аудиофайлов с текстовыми транскрипциями, выполненными разметчиками вручную. Тестовые данные использовались исключительно для оценки качества и не участвовали в процессе обучения. Для русского и английского языков подготовлено по 400 семплов длительностью до 20 секунд каждый – в общей сложности по 2 часа размеченных тестовых данных.

Тестовые данные подавались на вход моделей, а полученные (распознанные) в результате строки сравнивались с данными, записанными людьми. Использованы метрики качества: частота ошибок в словах (англ. word error rate, WER); частота ошибок в символах (англ. character error rate, CER). Также, по аналогии с предыдущими метриками, применена метрика частоты ошибок в терминах предметной области (англ. term error rate, TER), которая показывает, как модель справляется с распознаванием специфической лексики предметной области:

$$TER = \frac{S_t + D_t + I_t}{N_t}, \quad (3)$$

где S_t, D_t, I_t – количество замен, удалений и вставок терминов, которые необходимо выполнить, чтобы получить исходную (тестовую) строку терминов из полученной (предсказанной) в результате работы модели; N_t – общее количество терминов в тестовой строке.

Для расчета метрики TER в тестовой и предсказанной строках остаются только целевые термины, определяемые словарем специфической лексики, который составляется вручную для данной предметной области. Остальные слова удаляются.

В таблице приведены результаты тестирования моделей для русского и английского языков: Cit. и Conf. означают архитектуры Citrinet и Conformer-CTC соответственно, общ. – модель общего назначения, мод. – модель дообученная на тематически семплированных данных (модифицированная для предметной области).

Анализ данных таблицы показывает, что для всех архитектур и языков на тестовой выборке лучше показывают себя модифицированные модели, а архитектура Conformer-CTC превосходит Citrinet. На улучшения указывают сниженные значения всех используемых мет-

рик качества – частоты ошибок в словах, символах и терминах.

Результаты оценки моделей распознавания речи

Модель	WER, %	CER, %	TER, %
Cit., англ., общ.	17,3	12,2	21,1
Cit., англ., мод.	**15,5**	**10,4**	**19,2**
Cit., рус., общ.	18,8	13,4	22,6
Cit., рус., мод.	**16,3**	**11,5**	**19,7**
Conf., англ., общ.	14,5	9,5	17,8
Conf., англ., мод.	**11,8**	**7,7**	**15,2**
Conf., рус., общ.	15,1	9,9	18,6
Conf., рус., мод.	**12,2**	**8,3**	**16,4**

Модифицированные модели, дообученные на тематических данных, показывают значения метрик WER, CER и TER лучше, чем исходные модели общего назначения, в среднем на 2–3 %. Для того чтобы проверить, не находятся ли результаты оценок в пределах колебаний средних значений, был использован метод оценки *t*-критерия Стьюдента для двух связанных выборок [24]. В результате проверки для каждой пары моделей до и после модификации для каждой метрики получены значения p-критерия (p-value) ниже 0,04. Это позволяет отвергнуть нулевую гипотезу при достигаемом уровне значимости 5 % о том, что метрики моделей до и после дообучения имеют одинаковые средние значения.

Во всех экспериментах CER показывает значения меньше, чем WER. Это говорит о том, что модели на уровне символов более точны, чем модели на уровне отдельных слов, т. е. модели склонны возвращать слова так, как они слышатся (фонетически), в ущерб тому, как они должны быть записаны (орфографически). TER в экспериментах всегда выше, чем WER. Это говорит о том, что предметная лексика хуже распознается моделями, чем общая, так как специфические термины гораздо реже встречаются в обучающих данных.

Модифицированные модели показывают значения TER ниже, чем модели общего назначения. Далее приводятся примеры словосочетаний из предметной области и варианты их ошибочного распознавания моделями общего назначения:

- Versicharge – versu charge, versus charge;
- Siemens eMobility – siemen sima bility, simons immobility, simon see mobility;
- Evocharge – eva charge, eve charge, evil charge и т.п.

На этих примерах видно, что для разных узкоспециализированных терминов и имен собственных модели общего назначения склонны выдавать различные вариации общеупотребительных слов и словосочетаний. Модифицированные же модели при распознавании терминов предметной области склонны чаще возвращать их орфографически правильную форму. Также, по эмпирическим наблюдениям, снижается вариативность распознаваний терминов, т. е. уменьшается разнообразие возвращаемых моделью словоформ, что положительно влияет на возможность использования методов постобработки распознанного текста.

Результаты исследования. Анализ применения средств распознавания речи в такой области, как энергетика, показывает необходимость специального обучения моделей распознаванию специализированной лексики. При этом традиционный способ улучшения качества моделей распознавания, предполагающий сбор сотен часов акустических данных и их ручную разметку для дальнейшего обучения моделей общего назначения, негативно влияет на стоимость и сроки внедрения информационных систем. Также имеют значение вопросы конфиденциальности, ограничивающие доступ к такой информации.

Предложенный метод заключается в использовании для обучения распознаванию тематической лексики наборов данных с открытым доступом путем поиска и выделения из них семплов, семантически близких предметной области, для последующего обучения моделей общего назначения. В процессе исследования были выбраны и испытаны программные библиотеки и технологии с открытым доступом, обеспечивающие возможность создания обучающей выборки для сферы энергетики. Результаты экспериментов показали, что получение такой выборки занимает несколько дней работы на одном типовом компьютере.

Применение метода позволяет:

1) улучшить качество распознавания терминов предметной области и снизить их вариативность;

2) использовать только общедоступные данные для улучшения качества распознавания специализированной лексики;

3) использовать полученную модель в качестве инициализирующей для дальнейшей тонкой настройки на закрытых данных;

4) уменьшить объем закрытых наборов данных, необходимых для дальнейшей тонкой настройки.

Разработанная на основе метода прикладная модель продемонстрировала высокий уровень распознавания терминов на русском и английском языках без дополнительного обучения на специально подготовленных вручную наборах данных.

Выводы. Разработанный метод создания моделей распознавания речи позволяет существенно упростить и ускорить процесс обучения распознаванию специализированной профессиональной лексики за счет использования размеченных данных из открытых источников, отобранных с применением методов те-

матического семплирования. Испытания предложенного метода при создании модели для объектов энергетики подтвердили его эффективность. Результаты исследования могут применяться при разработке информационных систем в области проектирования и эксплуатации объектов энергетики и энергетического оборудования.

Список литературы

1. **Attention** is all you need / A. Vaswani, N. Shazeer, N. Parmar, et al. // Advances in neural information processing systems. – 2017. – Vol. 30.

2. **Transformer** transducer: A streamable speech recognition model with transformer encoders and RNN-T loss / Q. Zhang, H. Lu, H. Sak, et al. // ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2020. – P. 7829–7833.

3. **Li J.** Recent advances in end-to-end automatic speech recognition // APSIPA Transactions on Signal and Information Processing. – 2022. – Vol. 11. – No. 1.

4. **Невлюдов И.Ш., Цымбал А.М., Милютина С.С.** Использование искусственной нейронной сети в подсистеме ввода голосовой информации САПР ТП роботизированного производства // Радиоэлектроника и информатика. – 2007. – № 1. – С. 56–61.

5. **Saon G., Chien J.T.** Large-vocabulary continuous speech recognition systems: A look at some recent advances // IEEE signal processing magazine. – 2012. – Vol. 29, No. 6. – P. 18–33.

6. **Deep** contextualized acoustic representations for semi-supervised speech recognition / S. Ling, Y. Liu, J. Salazar, K. Kirchhoff // ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2020. – P. 6429–6433.

7. **Audiolm:** a language modeling approach to audio generation / Z. Borsos, R. Marinier, D. Vincent, et al. // arXiv preprint arXiv:2209.03143. – 2022.

8. **Russian** open speech to text (stt/asr) dataset (2022) / A. Slizhikova, A. Veysov, D. Nurtdinova, et al. https://github.com/snakers4/open_stt/

9. **Gigaspeech:** An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio / G. Chen, S. Chai, G. Wang, et al. // arXiv preprint arXiv:2106.06909. – 2021.

10. **Librispeech:** an asr corpus based on public domain audio books / V. Panayotov, G. Chen, D. Povey, S. Khudanpur // 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). – IEEE, 2015. – С. 5206–5210.

11. **Common** voice: A massively-multilingual speech corpus / R. Ardila, M. Branson, K. Davis, et al. // arXiv preprint arXiv:1912.06670. – 2019.

12. **Mls:** A large-scale multilingual dataset for speech research / V. Pratap, Q. Xu, A. Sriram, et al. // arXiv preprint arXiv:2012.03411. – 2020.

13. **Munich** Artificial Intelligence Laboratories GmbH. The m-ailabs speech dataset. https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/, 2017.

14. **The multilingual** tedx corpus for speech recognition and translation / E. Salesky, M. Wiesner, J. Bremerman, et al. // arXiv preprint arXiv:2102.01757. – 2021.

15. **Voxpopuli:** A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation / C. Wang, M. Rivière, A. Lee, et al. // arXiv preprint arXiv:2101.00390. – 2021.

16. **VoxForge**, Free Speech Recognition. www.voxforge.org, 2022.

17. **Mpnet:** Masked and permuted pre-training for language understanding / K. Song, X. Tan, T. Qin, et al. // Advances in Neural Information Processing Systems. – 2020. – Т. 33. – С. 16857–16867.

18. **McInnes L., Healy J., Melville J.** Umap: Uniform manifold approximation and projection for dimension reduction // arXiv preprint arXiv:1802.03426. – 2018.

19. **McInnes L., Healy J.** Accelerated hierarchical density based clustering // 2017 IEEE International Conference on Data Mining Workshops (ICDMW). – IEEE, 2017. – С. 33–42.

20. **Grootendorst M.** BERTopic: Neural topic modeling with a class-based TF-IDF procedure // arXiv preprint arXiv:2203.05794. – 2022.

21. **Nemo:** a toolkit for building ai applications using neural modules / O. Kuchaiev, J. Li, H. Nguyen, et al. // arXiv preprint arXiv:1909.09577. – 2019.

22. **Citrinet:** Closing the gap between non-autoregressive and autoregressive end-to-end models for automatic speech recognition / S. Majumdar, J. Balam, O. Hrinchuk, et al. // arXiv preprint arXiv:2104.01721. – 2021.

23. **Conformer:** Convolution-augmented transformer for speech recognition / A. Gulati, J. Qin, C. Chiu, et al. // arXiv preprint arXiv:2005.08100. – 2020.

24. **Student's** t-test. Dependent t-test for paired samples. URL: https://en.wikipedia.org/wiki/T-test#Dependent_t-test_for_paired_samples (дата обращения: 01.02.2023).

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I. Attention is all you need. Advances in neural information processing systems, 2017, vol. 30.

2. Zhang, Q., Lu, H., Sak, H., Tripathi, A., McDermott, E., Koo, S., Kumar, Sh. Transformer transducer: A streamable speech recognition model with transformer encoders and RNN-T loss. ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 7829–7833.

3. Li, J. Recent advances in end-to-end automatic speech recognition. APSIPA Transactions on Signal and Information Processing, 2022, vol. 11, no. 1.

4. Nevlyudov, I.S., Cymbal, A.M., Milyutina, S.S. Ispol'zovanie iskusstvennoj nejronnoj seti v podsisteme vvoda golosovoj informacii SAPR TP robotizirovannogo proizvodstva [Application of artificial neural network in the subsystem of entering voice information of CAD TP of robotic production]. *Radioelektronika i informatika*, 2007, no. 1, pp. 56–61.

5. Saon, G., Chien, J.T. Large-vocabulary continuous speech recognition systems: A look at some recent advances. IEEE signal processing magazine, 2012, vol. 29, no. 6, pp. 18–33.

6. Ling, S., Liu, Y., Salazar, J., Kirchhoff, K. Deep contextualized acoustic representations for semi-supervised speech recognition. ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 6429–6433.

7. Borsos, Z., Marinier, R., Vincent, D., Kharitonov, E., Pietquin, O., Sharifi, M., Teboul, O., Grangier, D., Tagliasacchi, M., Zeghidour, M. Audiolm: a language modeling approach to audio generation. arXiv preprint arXiv:2209.03143. 2022.

8. Slizhikova, A., Veysov, A., Nurtdinova, D., Voronin, D., Baburov, Y. Russian open speech to text (stt/asr) dataset (2022). URL: https://github.com/snakers4/open_stt/

9. Chen, G., Chai, S., Wang, G., Du, J., Zhang, W., Weng, C., Su, D., Povey, D., Trma, J., Zhang, J., Jin, M., Khudanpur, S., Watanabe, S., Zhao, S., Zou, W., Li, X., Yao, X., Wang, Y., Wang, Y., You, Z., Yan, Z. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. arXiv preprint arXiv:2106.06909. 2021.

10. Panayotov, V., Chen, G., Povey, D., Khudanpur, S. Librispeech: an asr corpus based on public domain audio books. 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2015, pp. 5206–5210.

11. Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F., Weber, G. Common voice: A massively-multilingual speech corpus. arXiv preprint arXiv:1912.06670. 2019.

12. Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., Collobert, R. Mls: A large-scale multilingual dataset for speech research. arXiv preprint arXiv:2012.03411. 2020.

13. Munich Artificial Intelligence Laboratories GmbH. The m-ailabs speech dataset. https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/, 2017.

14. Salesky, E., Wiesner, M., Bremerman, J., Cattoni, R., Negri, M., Turchi, M., Oard, D., Post, M. The multilingual tedx corpus for speech recognition and translation. arXiv preprint arXiv:2102.01757. 2021.

15. Wang, C., Rivière, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., Dupoux, E. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. arXiv preprint arXiv:2101.00390. 2021.

16. VoxForge, Free Speech Recognition. www.voxforge.org, 2022.

17. Song, K., Tan, X., Qin, T., Lu, J., Liu, T. Mpnet: Masked and permuted pre-training for language understanding. Advances in Neural Information Processing Systems, 2020, vol. 33, pp. 16857–16867.

18. McInnes, L., Healy, J., Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426. 2018.

19. McInnes, L., Healy, J. Accelerated hierarchical density based clustering. 2017 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, 2017, pp. 33–42.

20. Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794. 2022.

21. Kuchaiev, O., Li, J., Nguyen, H., Hrinchuk, O., Leary, R., Ginsburg, B., Kriman, S., Beliaev, S., Lavrukhin, V., Cook, J., Castonguay, P., Popova, M., Huang, J., Cohen, J. Nemo: a toolkit for building ai applications using neural modules. arXiv preprint arXiv:1909.09577. 2019.

22. Majumdar, S., Balam, J., Hrinchuk, O., Lavrukhin, V., Noroozi, V., Ginsburg, B. Citrinet: Closing the gap between non-autoregressive and autoregressive end-to-end models for automatic speech recognition. arXiv preprint arXiv:2104.01721. 2021.

23. Gulati, A., Qin, J., Chiu, C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., Pang, R. Conformer: Convolution-augmented transformer for speech recognition. arXiv preprint arXiv:2005.08100. 2020.

24. Student's t-test. Dependent t-test for paired samples. URL: https://en.wikipedia.org/wiki/T-test#Dependent_t-test_for_paired_samples (access date: 01.02.2023).

ВЕСТНИК ИВАНОВСКОГО ГОСУДАРСТВЕННОГО ЭНЕРГЕТИЧЕСКОГО УНИВЕРСИТЕТА

Выпуск 4

Издание зарегистрировано в Федеральной службе по надзору в сфере связи, информационных технологий и массовых коммуникаций.
Свидетельство о регистрации ПИ № ФС77-82616 от 18.01.2022 г.

Подписано в печать 08.08.2023. Выход в свет 31.08.2023. Формат 60x84 $^1/_8$.
Усл. печ. л. 11,62. Уч.-изд. л. 12,4. Тираж 100 экз. Цена свободная. Заказ

Адрес редакции и издательства: Ивановский государственный энергетический университет, 153003, Ивановская область, г. Иваново, ул. Рабфаковская, 34.
Типография ООО «ПресСто»: 153025, Ивановская обл., г. Иваново, ул. Дзержинского, 39, строение 8.