УДК 004.032.26 **DOI 10.52167/1609-1817-2024-131-2-456-466**

Б. Медетов[1], А. Нурланкызы[2,3] ✉, А. Кулакаева[4], А. Жетписбаева[1], Т. Намазбаев[5]
[1]Казахский агротехнический исследовательский университет имени С.Сейфуллина, Астана, Казахстан
[2]Satbayev University, Алматы, Казахстан
[3]Energo University, Алматы, Казахстан
[4] Международный университет информационных технологий, Алматы, Казахстан
[5]Казахский национальный университет им. аль-Фараби, Алматы, Казахстан
E-mail: nurlankyzyaigulya@gmail.com

ОЦЕНКА ВЛИЯНИЯ ЯЗЫКА НА ТОЧНОСТЬ РАСПОЗНАВАНИЯ ЧЕЛОВЕЧЕСКОГО ГОЛОСА С ПОМОЩЬЮ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ

Аннотация. Данная работа посвящена оценке влияния языка на точность распознавания человеческого голоса с помощью искусственных нейронных сетей. Так, традиционные VAD системы работают анализом энергии и энтропии сигнала, что представляет собой алгоритмический метод. Однако в реальной жизни практически невозможно точно описать параметры человеческого голоса с помощью алгоритмов. В связи с этим, в современных технологиях распознавания речи используются искусственные нейронные сети. Так как методы, основанные на искусственных нейронных сетях, достигают впечатляющих результатов в области распознавания человеческого голоса.

Результаты исследования в данной работе указывают на важность языковых особенностей при обучении и применении нейронных сетей для распознавания речи. Дальнейшие исследования в этой области могут сфокусироваться на разработке методов, которые улучшат универсальность и обобщающую способность нейронных сетей в распознавании речи на различных языках. Эти результаты имеют важное значение для развития технологий распознавания речи и могут быть использованы в различных областях, включая разработку многоязычных систем распознавания речи.

Также в рамках данного исследования обнародованы интересные фонетические особенности. Несмотря на родственные связи казахского языка с другими тюркскими языками, наблюдалось более успешное распознавание русского языка. Эти результаты могут быть полезны для изучения фонетических сходств и различий между языками, а также для разработки эффективных методов обучения нейронных сетей для распознавания речи на разных языках.

Ключевые слова. Детекторы голосовой активности, искусственные нейронные сети, многослойный персептрон (MLP), рекуррентная нейронная сеть (RNN) и сверточная нейронноя сеть (CNN).

Введение.

Технология Voice Activity Detector (VAD) представляет собой важный инструмент в системах мобильной связи, предназначенный для определения активности голоса во время разговора. Применение VAD в мобильных сетях позволяет достичь оптимизации использования пропускной способности, улучшения эффективности кодирования речи и повышения качества связи. Уникальные особенности VAD, такие как его адаптивность, способность обнаруживать голос на фоне шума и возможность интеграции с другими технологиями, делают эту технологию важной и эффективной. Роль VAD в системах

мобильной связи заключается в обеспечении оптимизации пропускной способности, улучшении качества связи и эффективном использовании аудио ресурсов. Однако использование нейронных сетей в VAD имеет определенные преимущества по сравнению с традиционными методами обнаружения активности голоса. Нейронные сети способны автоматически извлекать признаки из аудиосигналов и обучаться на больших объемах данных, что позволяет им обобщать свои знания и быть более адаптивными к различным типам речи и окружающему шуму. Кроме того, нейронные сети имеют способность выявлять сложные шаблоны и зависимости во входных данных, что позволяет им более точно определять активность голоса.

Однако несмотря на эффективность и перспективы использования нейронных сетей для распознавания голоса, существуют нерешенные проблемы и вопросы, связанные с обеспечением достаточного объема обучающих данных. В большинстве случаев требуются большие объемы данных, что может потребовать значительных человеческих, временных и вычислительных ресурсов. Для обучения нейронной сети по распознаванию человеческого голоса необходимо использовать разнообразные голосовые данные, включая разные языки и различные возрастные и гендерные группы. Однако, это представляет значительные трудности, и возникает вопрос, можно ли ограничить этот объем данных, не снижая точность распознавания. Следовательно, данному вопросу посвящена данная работа.

Цель исследования заключалась в выяснении, как количество различных дикторов и языков влияет на точность нейронной сети при распознавании человеческого голоса. В частности, рассматривался вопрос о возможности обучения нейронной сети на одном языке и ее использования для распознавания голоса на другом языке.

В работе [1] представлено приложение для смартфонов с использованием сверточной нейронной сети для определения голосовой активности в режиме реального времени с низкой задержкой звука. Архитектура сверточной нейронной сети была оптимизирована таким образом, чтобы обеспечить обработку аудиокадров в режиме реального времени без пропуска каких-либо кадров при сохранении высокой точности определения голосовой активности. Для обучения и оценки разработанного CNN VAD были использованы речевые файлы корпуса PN/NC версии 1.0. Этот корпус состоит из 20 носителей (10 мужчин, 10 женщин) из двух диалектных регионов американского английского (Тихоокеанский Северо-запад и Северные города), произносящих 180 предложений набора IEEE “Harvard”. В общей сложности он состоит из 3600 аудиофайлов. Однако в данной работе не рассмотрены задачи для оценки производительности VAD с применением наборов данных других языков, а также нет исследований, посвященных анализу необходимого количества дикторов для получения высокоэффективной VAD системы.

В работе [2] представлен метод VAD, который использует статистические функции на основе линейного спектра и частоты. Эксперименты проводились на базе данных объемом более 350 часов, состоящей из данных различных источников, таких как Youtube. Точность системы составила 99,43%. Тем не менее, также не проводились исследования, связанные с изучением количества необходимых дикторов.

В работе [3] предлагается быстрый и эффективный неконтролируемый метод VAD с использованием оценки фрактальной размерности. Для оценки эффективности предлагаемого метода используются две базы данных на разных языках. Первая — это база данных Массачусетского технологического института Texas Instruments (TIMIT), а вторая — база данных арабской речи Университета короля Сауда (KSU). Язык речевой базы данных TIMIT — английский, представляет собой 630 носителей мужского и женского пола восьми диалектных регионов. Язык речевой базы данных KSU — арабский, состоит из 328 говорящих мужчин и женщин, записавших как заранее написанный текст,

так и спонтанный текст. Однако, обучение и тестирование VAD системы в данной работе для каждого набора данных проводилось по отдельности.

В работе [4] предложена функция улучшения речи VAD на основе вариационного автокодировщика. В данной работе использовали чистые высказывания из базы данных Aurora4 [14], которая содержит 7138 непрерывных речевых высказываний для обучения и 330 высказываний для тестирования. Для построения 35-часового обучающего набора были использованы все 7138 высказываний из чистого обучающего набора. Высказывания в корпусе Aurora4 короткие, и около 80% из них – речь.

В работе [5] функция обнаружения голосовой активности (VAD) объединяется со сквозным автоматическим распознаванием речи с онлайн-речевым интерфейсом и расшифровкой очень длинных аудиозаписей. Для первой оценки был использован корпус японского языка CSJ. Он содержит около 650 часов данных о спонтанной японской речи. Для второй оценки был использован корпус TED-LIUMV2, представляющий собой набор выступлений TED на английском языке с транскрипциями. Он содержит около 200 часов речевых данных. Были исследованы два типа структуры сети: 6-слойный двунаправленный LSTM-кодер с 1-слойным LSTM-декодером и 6-слойный однонаправленный LSTM-кодер с 2-слойным LSTM-декодером. Экспериментальные результаты на несегментированных данных показывают, что предложенный метод превзошел базовые методы, использующие традиционные методы VAD, основанные на энергии и нейронных сетях. Однако, как во многих работах, не проводились исследования, посвященные анализу количества дикторов.

В работе [6] была обучена модель сквозной сегментации, которая выполняет комбинацию комбинация трех подзадач: обнаружение голосовой активности, обнаружение смены говорящего и обнаружение перекрывающейся речи. Для проведения исследования был выбран речевая корпус DIHARD3, разработанный Консорциумом лингвистических данных и содержащий около 34 часов речевых данных на английском и китайском языках. В работе данная база данных была разделена на две части: 192 файла, используемые в качестве обучающего набора, и оставшиеся 62 файла, используемые для проведения тестирования. Эксперименты с наборами данных диаризации нескольких говорящих пришли к выводу, что эту модель можно с большим успехом использовать как для обнаружения речевой активности, так и для обнаружения перекрывающейся речи.

В работе [7] разработана нейронная сеть, объединяющая обучаемые фильтры и рекуррентные слои для обнаружения голосовой активности непосредственно по форме сигнала. Эксперименты со сложным набором данных DIHARD показывают, что предлагаемая сквозная модель достигает самых современных показателей и превосходит вариант, в котором обучаемые фильтры заменяются стандартными кепстральными коэффициентами. Один и тот же набор данных DIHARD, взятый из 11 различных областей, используется для оценки по двум сценариям. В внутридоменном сценарии, когда наборы обучения и тестирования охватывают одни и те же домены, где показывают, что доменно-состязательный подход не снижает производительность предлагаемой сквозной модели. В сценарии вне домена, где тестовый домен отличается от обучающего, это дает относительное улучшение более чем на 10%. Однако, обучение и тестирование проводилось с использованием только одного единственного набора данных.

В работе [8] рассматривается задача обнаружения речевой активности на основе сверточной нейронной сети. Экспериментальные исследования проводились на двух речевых наборах данных: AMI - мультимодальный набор данных, состоящий из 100 часов записей встреч на английском языке, а также набор данных CHiME-6. Корпус CHiME-6 содержит более 60 часов записей, организованных в 20 сеансов.

В работе [9] предлагается комплексная многозадачная модель для VAD с улучшением речи. В качестве источника чистой речи использовался набор данных

английского языка Wall Street Journal (WSJ0). Он содержит 12776 высказываний от 101 выступающего для обучения, 1206 высказываний от 10 выступающих для проверки и 651 высказывание от 8 выступающих для оценки. Результаты экспериментов показывают, что многозадачный метод значительно превосходит свой однозадачный аналог VAD.

В работе [10] предлагается сверточная нейронная сеть (CNN) с несколькими входами и одним выходом, которая использует новую комбинацию функций для оценки VAD. Для обучения и валидации были использованы голоса 168 дикторов из набора данных TIMIT, включая как мужчин, так и женщин. Экспериментальные результаты для сценария с одним говорящим показывают, что предлагаемая CNN способна отличать речь от блоков неречевых сигналов, тем самым превосходя базовую CNN. Кроме того, результаты показывают, что предлагаемый метод способен адаптироваться к различным невидимым акустическим условиям и фоновым шумам

Таким образом, изучение и анализ исследований, посвященных данной теме, показало, что в настоящее время недостаточное количество исследований, посвященных языконезависимому обнаружению голосовой активности, а также анализу необходимого количества дикторов для создания эффективной языконезависимой системы распознавания голоса.

Материалы и методы.

В качестве объектов исследования в данной работе были использованы различные виды искусственных нейронных сетей, используемых для распознавания человеческого голоса, такие как обычный многослойный персептрон (MLP), сверточная нейронная сеть (CNN) и рекуррентная нейронная сеть (RNN) и их способности эффективно распознавать голос в разных языках после обучения на небольшом количестве дикторов.Следовательно, данное исследование фокусируется на использовании различных искусственных нейронных сетей для распознавания человеческого голоса. Основная гипотеза заключается в том, что нейронные сети могут эффективно распознавать человеческий голос независимо от языка, даже при обучении на ограниченном числе дикторов.

Данное исследование сосредоточено на обучении нейронных сетей на данных, представляющих высказывания дикторов на казахском языке. Предполагается, что даже при этнических и лингвистических различиях фонетики различных языков, нейронные сети, обученные на одном языке, должны быть способны распознавать голоса на других языках.

Для обучения и тестирования нейронных сетей были использованы наборы данных Института умных систем и искусственного интеллекта (ISSAI) Назарбаев Университета, включая корпус казахской речи [11], корпус русской речи [12], корпус турецкого языка [13], и корпус узбекского языка [14]. Кроме того, для обучения использовался Common Voice Dataset [15], включая корпус киргизского языка, английского языка и французского языка. Из каждого набора данных были отобраны по 15 мужских и 15 женских голосов с различными характеристиками, такие как интонация, высота тона,возраст.

В работе в ходе исследования производилась ручная разметка аудиофайлов корпуса казахской речи [11] при помощи программного обеспечения Audacity 3.4.2. Присутствие звука в аудиофайле было отмечено как «1», а его отсутствие - как «0». Затем каждый блок аудиоданных был разделен на фрагменты продолжительностью 20 миллисекунд. Для каждого фрагмента были вычислены мел-кепстральные коэффициенты (Mel-Frequency Cepstral Coefficients-MFCC), а также дельта и дельта-дельта коэффициенты.

Далее каждый блок аудиоданных разбивался на фрагменты продолжительностью 20 миллисекунд. Затем для каждого фрагмента осуществлялось вычисление MFCC, а также дельта и дельта-дельта коэффициентов. Таким образом, каждый фрагмент речевого

сигнала представляется 36 коэффициентами. При обучении сетей MLP, на вход подаются 36 значений, включающих 12 коэффициентов MFCC, 12 коэффициентов дельта и 12 коэффициентов дельта-дельта. Соответственно, входной слой сетей MLP содержит 36 нейронов, а на выходе имеется только один нейрон, который может принимать либо значение 1, если текущий фрагмент является речевым, либо 0, если текущий фрагмент не является речевым.

Для сетей типа CNN используются только 12 коэффициентов MFCC, без коэффициентов delta and delta-delta. Тем не менее на входе данной сети также имеются 36 нейронов, но в отличие от сетей MLP, остальные 24 значения берутся из последующих соседних фрагментов.

Структура сетей типа RNN, использованных в данном исследовании, имеют похожую на CNN структуру. Здесь также не используем delta and delta-delta коэффициенты. При этом динамику аудиоданных задаем с помощью использования MFCC коэффициентов соседних трех фрагментов сигнала.

Результаты.

Для осуществления оценки влияния языка на точность распознавания человеческого голоса снова будем использовать установленные функции аппроксимации. На рисунке 1 показаны графики зависимости ошибки распознавания человеческого голоса по высказываниям на различных языках от количества дикторов для нейронной сети MLP.

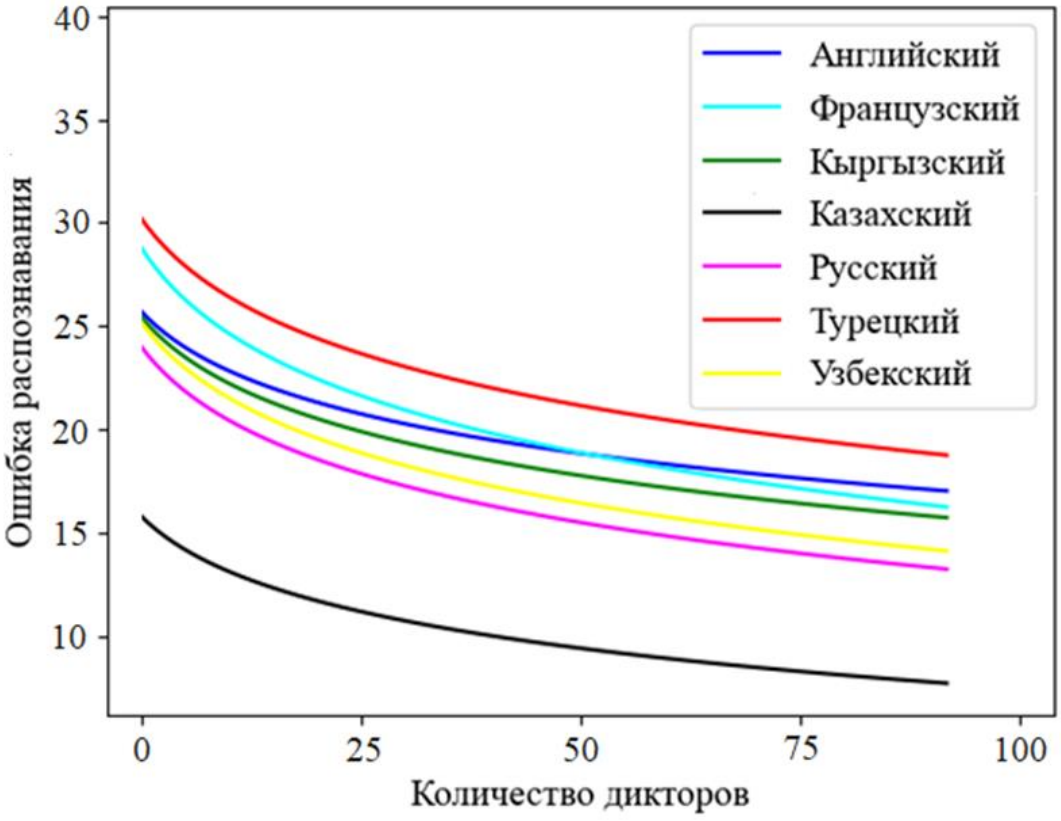

Рисунок 1 - Графики зависимости ошибки распознавания человеческого голоса по высказываниям на различных языках от количества дикторов для нейронной сети MLP

А на рисунке 2 приведены те же результаты, но уже для нейронной сети CNN.

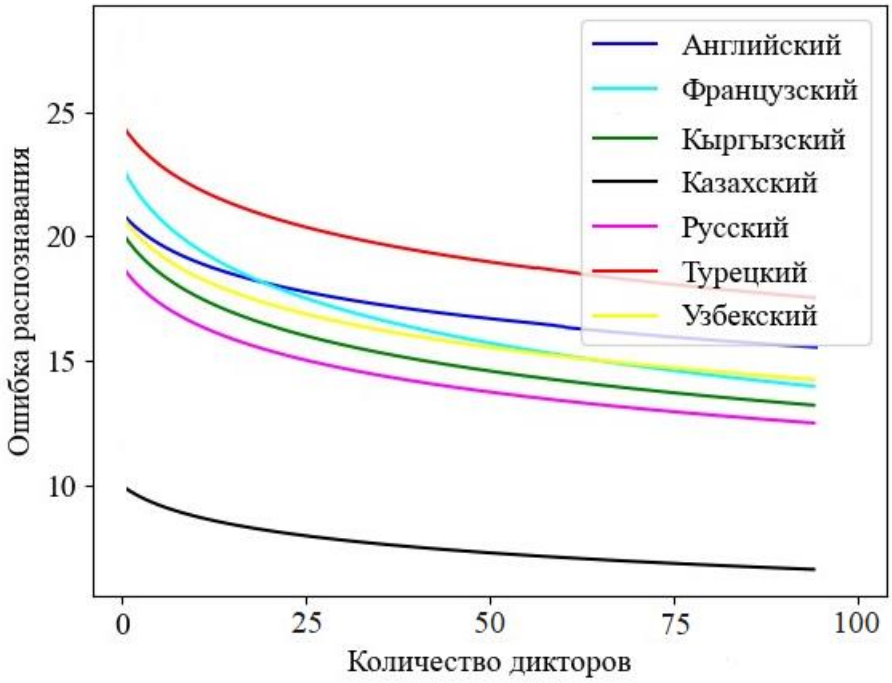

Рисунок 2 - Графики зависимости ошибки распознавания человеческого голоса по высказываниям на различных языках от количества дикторов для нейронной сети CNN

Наконец, на рисунке 3 представлены графики зависимости ошибки распознавания человеческого голоса по высказываниям на различных языках от количества дикторов для нейронной сети MLP.

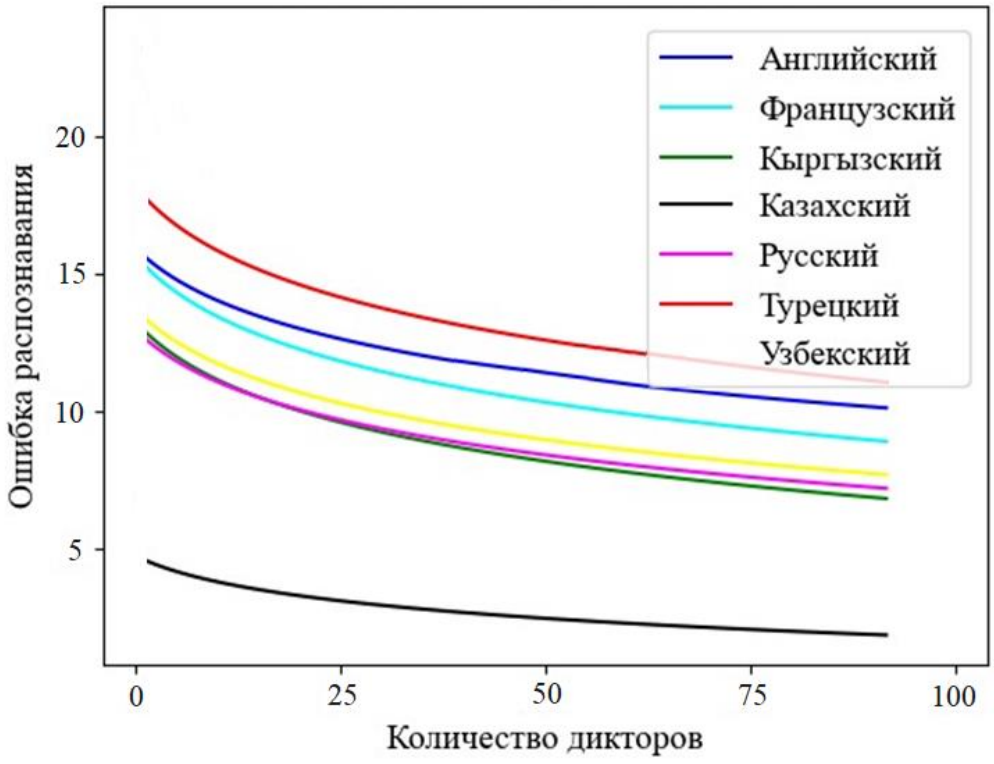

Рисунок 3 - графики зависимости ошибки распознавания человеческого голоса по высказываниям на различных языках от количества дикторов для нейронной сети RNN

Еще раз напомним, что все эти нейронные сети обучались на наборе речевых данных только на казахском языке. А тестирование сетей производились с помощью высказываний на различных языках, в том числе и на казахском языке.

Как видим, из рисунков 1, 2 и 3, человеческий голос по высказываниям на казахском языке распознается с заметным отрывом лучше, чем на остальных языках всеми тремя нейронными сетями. Еще одним интересным фактом является то, что разные сети на второе место ставят разные языки. Например, для RNN на втором месте после казахского языка расположился киргизский язык, а для CNN and MLP - русский язык. И во всех случаях аутсайдером оказался турецкий язык, хотя он входит в одно и тоже семейство языков вместе с казахским языком.

Все предыдущие графики, показанные на рисунках 1–3, указывают на наличие однозначной зависимости точности распознавания человеческого голоса нейронными сетями от количества дикторов, использованных при обучении этих сетей. Очевидно, что с ростом количества дикторов, падает ошибка распознавания человеческого голоса, т.е. растет эффективность нейронной сети. В результате зависимость ошибки распознавания от количества дикторов для сети MLP (на примере казахского языка) можно представить следующим образом:

$$y = 29{,}10 - 3{,}10 \cdot \log(x), \tag{1}$$

а для сети CNN также для казахского языка данную зависимость можно выразить с помощью следующего выражения:

$$y = 12{,}62 - 1{,}13 \cdot \log(x), \tag{2}$$

наконец, для сети RNN имеется следующая закономерность:

$$y = 12{,}77 - 1{,}12 \cdot \log(x), \tag{3}$$

где во всех формулах (1), (2) и (3) y – означает величину ошибки распознавания человеческого голоса нейронной сетью, а x – это количество дикторов, чьи высказывания были использовании при обучении нейросетей.

Расчеты показывают, что если эти закономерности верны, то для того, чтобы получить ошибку распознавания человеческого голоса не более 3% (точность не менее 97%) необходимо иметь не меньше 4.5 тыс. образцов голосов для обучения нейронных сетей.

Обсуждение.

В начале предполагалось, что эффективность работы нейронных сетей слабо зависит от языка говорящего, так как все языки имеют большое количество схожих фонем. Таким образом, ожидалось, что, обучив какую-либо нейронную сеть на одном языке, можно с достаточно хорошей точностью обнаруживать человеческие голоса на других языках с помощью этой же нейронной сети. Но эти ожидания в полной мере не оправдались. Все же имеется зависимость от языка. Это можно увидеть на рисунках 1, 2 и 3, где обученные на казахском языке нейронные сети MLP, CNN and RNN гораздо лучше обнаруживают голоса на этом же языке, чем голоса на других языках. Но в то же время мы натолкнулись на странный результат. Несмотря на то, что казахский, киргизский, узбекский и турецкий языки считаются родственными, т.е. относятся к одной и той же группе тюркских языков, обученные на казахском языке нейронные сети лучше остальных распознает голоса на русском языке. Это, по крайне мере, можно увидеть на рисунках 1 и

2. Возможно, это объясняется тем, что турецкий язык, имеет гораздо большее фонетическое отличие от казахского языка, чем тот же русский язык. Таким образом, примененные методы и результаты данного исследования могли бы быть полезными для изучения фонетических сходств между собой разных языков.

Заключение.

Таким образом, результаты исследования показали, что эффективность работы нейронных сетей зависит от языка, на котором они обучались. Это важное наблюдение указывает на необходимость учета языковых особенностей при обучении и применении нейронных сетей для распознавания речи. Дальнейшие исследования в этой области могут быть направлены на разработку методов, которые позволят улучшить универсальность и обобщающую способность нейронных сетей в задачах распознавания речи на различных языках.

В заключение хотелось бы отметить, что результате нашего исследования выявлены интересные фонетические особенности, которые могут быть обнаружены при обучении нейронных сетей на разных языках. Несмотря на родственные связи между рассматриваемыми тюркскими языками, наблюдается более успешное распознавание русского языка, что может быть обусловлено фонетическими различиями между языками. Эти результаты могут быть полезны для дальнейшего изучения фонетических сходств и различий между различными языками, а также для разработки более эффективных методов обучения нейронных сетей для распознавания речи на разных языках.

Благодарность. Работа выполнена при финансовой поддержке КН МОН РК по программе грантового финансирования научных исследований, грант AP19678995 «Разработка метода распознавания дикторов с применением глубоких нейронных сетей при ультракороткой продолжительности чистой речи»

ЛИТЕРАТУРА

[1] Sehgal A., Kehtarnavaz N. A Convolutional Neural Network Smartphone App for Real-Time Voice Activity Detection. (2018) IEEE Access, 6, pp. 9017 – 9026. https://doi.org/10.1109/ACCESS.2018.2800728

[2] Mukherjee H., Obaidullah S.M., Santosh K.C., Phadikar S., Roy K. Line spectral frequency-based features and extreme learning machine for voice activity detection from audio signal (2018) International Journal of Speech Technology, 21 (4), pp. 753 - 760, https://doi.org/10.1007/s10772-018-9525-6

[3] Ali, Z., & Talha, M. (2018). Innovative Method for Unsupervised Voice Activity Detection and Classification of Audio Segments. IEEE Access, 6, 15494–15504. https://doi.org/10.1109/access.2018.2805845

[4] Jung, Youngmoon & Kim, Younggwan & Choi, Yeunju & Kim, Hoirin. (2018). Joint Learning Using Denoising Variational Autoencoders for Voice Activity Detection. 1210-1214. https://doi.org/10.21437/Interspeech.2018-1151

[5] Yoshimura, T., Hayashi, T., Takeda, K., & Watanabe, S. (2020). End-to-End Automatic Speech Recognition Integrated with CTC-Based Voice Activity Detection. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). https://doi.org/10.1109/icassp40776.2020.9054358

[6] Bredin H., Laurent A. End-to-end speaker segmentation for overlap-aware resegmentation (2021) Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 4, pp. 2463 – 2467. https://doi.org/10.21437/Interspeech.2021-560

[7] Lavechin M., Gill M.-P., Bousbib R., Bredin H., Garcia-Perera L.P. End-to-end domain-adversarial voice activity detection (2020) Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2020-October, pp. 3685 - 3689, https://doi.org/10.21437/Interspeech.2020-2285

[8] Cornell S., Omologo M., Squartini S., Vincent E. Detecting and counting overlapping speakers in distant speech scenarios (2020) Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2020-October, pp. 3107 – 3111 https://doi.org/10.21437/Interspeech.2020-2671

[9] Tan, X., & Zhang, X.-L. (2021). Speech Enhancement Aided End-To-End Multi-Task Learning for Voice Activity Detection. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). https://doi.org/10.1109/icassp39728.2021.9414445

[10] Varzandeh R., Adiloğlu K., Doclo S., Hohmann V. Exploiting periodicity features for joint detection and doa estimation of speech sources using convolutional neural networks (2020) ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2020-May, art. no. 9054754, pp. 566 – 570. https://doi.org/10.1109/ICASSP40776.2020.9054754

[11] Medetov, B., Kulakayeva, A., Zhetpisbayeva, A., Albanbay, N., Kabduali, T. Identifying the regularities of the signal detection method using the Kalman filter. // Eastern-European Journal of Enterprise Technologies, 2023, 5(9(125)), pp 26–34. https://doi.org/10.15587/1729-4061.2023.289472

[12] Mussakhojayeva, S., Khassanov, Y. , Varol, H.A.: KSC2: An Industrial-Scale Open-Source Kazakh Speech Corpus. In: Proceedings of the 23rd INTERSPEECH Conference: pp. 1367-1371. 2022. https://doi.org/10.21437/Interspeech.2022-421

[13] Mussakhojayeva S., Khassanov Y., Atakan Varol H. (2021) A Study of Multilingual End-to-End Speech Recognition for Kazakh, Russian, and English. In: Karpov A., Potapova R. (eds) Speech and Computer. SPECOM 2021. Lecture Notes in Computer Science, vol 12997. Springer, Cham. https://doi.org/10.1007/978-3-030-87802-3_41

[14] Mussakhojayeva, S.; Dauletbek, K.; Yeshpanov, R.; Varol, H.A. Multilingual Speech Recognition for Turkic Languages. Information 2023, 14, 74 https://doi.org/10.3390/info14020074

[15] Musaev M., Mussakhojayeva S., Khujayorov I., Khassanov Y., Ochilov M., Atakan Varol H. (2021) USC: An Open-Source Uzbek Speech Corpus and Initial Speech Recognition Experiments. In: Karpov A., Potapova R. (eds) Speech and Computer. SPECOM 2021. Lecture Notes in Computer Science, vol 12997. Springer, Cham. https://doi.org/10.1007/978-3-030-87802-3_40

Бекболат Медетов, PhD, қауымдастырылған профессор, С.Сейфуллин атындағы Қазақ агротехникалық зерттеу университеті, Астана, Қазақстан, bm02@mail.ru

Айгуль Нурланкызы, магистр, аға оқытушы, Satbayev University, Energo University, Алматы, Қазақстан, nurlankyzyaigulya@gmail.com

Айгуль Кулакаева, PhD, қауымдастырылған профессор м.а., Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан, a.kulakayeva@iitu.edu.kz

Айнур Жетписбаева, PhD, қауымдастырылған профессор м.а., С.Сейфуллин атындағы Қазақ агротехникалық зерттеу университеті, Астана, Қазақстан, aigulji@mail.ru

Тимур Намазбаев, магистр, аға оқытушы, әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан, timur.namazbayev@gmail.com

ЖАСАНДЫ НЕЙРОНДЫҚ ЖЕЛІЛЕР АРҚЫЛЫ ТІЛДІҢ АДАМ ДАУЫСЫН ТАНУ ДӘЛДІГІНЕ ӘСЕРІН БАҒАЛАУ

Аңдатпа. Бұл жұмыс жасанды нейрондық желілерді қолдана отырып, тілдің адам дауысын тану дәлдігіне әсерін бағалауға арналған. Сонымен, дәстүрлі VAD жүйелері алгоритмдік әдіс болып табылатын сигнал энергиясы мен энтропиясын талдаумен жұмыс істейді. Алайда нақты өмірде алгоритмдер арқылы адам дауысының параметрлерін дәл сипаттау мүмкін емес. Осыған байланысты сөйлеуді танудың заманауи технологияларында жасанды нейрондық желілер қолданылады. Жасанды нейрондық желілерге негізделген әдістер адамның дауысын тану саласында әсерлі нәтижелерге қол жеткізеді.

Осы жұмыстағы зерттеу нәтижелері сөйлеуді тану үшін нейрондық желілерді оқыту және қолдану кезінде тілдік ерекшеліктердің маңыздылығын көрсетеді. Осы саладағы қосымша зерттеулер әртүрлі тілдерде сөйлеуді танудағы нейрондық желілердің әмбебаптығы мен жалпылау қабілетін жақсартатын әдістерді әзірлеуге назар аударуы мүмкін. Бұл нәтижелер сөйлеуді тану технологияларын дамыту үшін маңызды және әртүрлі салаларда, соның ішінде көп тілді сөйлеуді тану жүйелерін дамытуда қолданылуы мүмкін.

Сондай-ақ, осы зерттеу аясында қызықты фонетикалық ерекшеліктер жарияланды. Қазақ тілінің басқа түркі тілдерімен туыстық байланыстарына қарамастан, орыс тілін неғұрлым табысты тану байқалды. Бұл нәтижелер тілдер арасындағы фонетикалық ұқсастықтар мен айырмашылықтарды зерттеуге және әртүрлі тілдерде сөйлеуді тану үшін нейрондық желілерді оқытудың тиімді әдістерін әзірлеуге пайдалы болуы мүмкін.

Түйінді сөздер. Дауыстық белсенділік детекторы, жасанды нейрондық желілер, көп қабатты перцептрон (MLP), қайталанатын нейрондық желі (RNN), конволюциялық нейрондық желі (CNN).

Bekbolat Medetov, PhD, associate professor, S. Seifullin Kazakh Agro Technical Research University, Astana, Kazakhstan, bm02@mail.ru

Aigul Nurlankyzy, master, senior lecturer, Satbayev University, Energo University, Almaty, Kazakhstan, nurlankyzyaigulya@gmail.com

Aigul Kulakayeva, PhD, associate professor, International University of Information Technology, Almaty, Kazakhstan, a.kulakayeva@iitu.edu.kz

Ainur Zhetpisbayeva, PhD, associate professor, S. Seifullin Kazakh Agro Technical Research University, Astana, Kazakhstan, aigulji@mail.ru

Timur Namazbayev, master, senior lecturer, Al-Farabi Kazakh National University, Almaty, Kazakhstan, timur.namazbayev@gmail.com

ASSESSMENT OF THE INFLUENCE OF LANGUAGE ON THE ACCURACY OF HUMAN VOICE RECOGNITION USING ARTIFICIAL NEURAL NETWORKS

Abstract. This study is devoted to assessing the influence of language on the accuracy of human voice recognition using artificial neural networks. Thus, traditional VAD systems work by analyzing the energy and entropy of a signal, which is an algorithmic method. However, in real life, it is almost impossible to accurately describe the parameters of human voice using algorithms. In this regard, artificial neural networks are used in modern speech recognition technologies. Artificial neural network-based methods have achieved impressive results in the field of human voice recognition.

The results of this study indicate the importance of language features in learning and the use of neural networks for speech recognition. Further research in this area may focus on the development of methods that will improve the versatility and generalization ability of neural networks in speech recognition in various languages. These results are important for the development of speech recognition technologies, and can be used in various fields, including the development of multilingual speech recognition systems.

Moreover, within the framework of this study, interesting phonetic features were made public. Despite the kinship of the Kazakh language with other Turkic languages, there was more successful recognition of the Russian language. These results can be useful for studying phonetic similarities and differences between languages, as well as for developing effective training methods for neural networks for speech recognition in different languages.

Keywords. Voice Activity Detector, artificial neural networks, multi-layered perceptron (MLP), recurrent neural network (RNN), convolutional neural network (CNN).

****