2. *Янь, Ма.* Анализ характеристик затрат-выпуска белорусских научно-технических инноваций / Ма Янь. — Contemporary Economy, 2020. — С. 26—29.

3. *Хуэй, Ван.* Языковая ситуация в странах, расположенных вдоль «пояса и пути» / Ван Хуэй, Ван Ялань. — Исследование языковой стратегии, 2016. — С. 13—19.

УДК 004.934.2

Е. Г. Шапович[1], А. В. Шах[2]

Учреждение образования «Барановичский государственный университет», Барановичи, Республика Беларусь, [1]evgeniy.shapovich@gmail.com , [2]shah.al.vas@gmail.com

РАСПОЗНАВАНИЕ ЭМОЦИЙ ПО РЕЧИ

В этой работе проводится обширное сравнение различных подходов к системам распознавания речи по эмоциям. Анализы проводились на аудиозаписи из аудиовизуальной базы данных эмоциональной речи и песен Райерсона. После предварительной обработки необработанных аудиофайлов считались такие функции как Log-Mel, спектрограмма, кепстральные коэффициенты Mel-частоты (MFCC), высота тона и энергия. Значение этих характеристик для классификации эмоций сравнивается с применением таких методов, как Long Short Term Memory (LSTM), сверточные нейронные сети (CNN), скрытые марковские модели (HMM).

Нейронные сети (DNN). По 14-классной классификации (2 пола × 7 эмоций) задача, точность 68 % была достигнута с 4-слойной 2-мерной CNN с использованием спектрограммы Log-Mel. Мы также наблюдаем, что при распознавании эмоций выбор звуковых характеристик влияет на результаты гораздо больше, чем сложность модели.

Ключевые слова: эмоции; нейронные сети; распознавание; искусственный интеллект; автоматизация.

E. G. Shapovich[1], A. V. Shakh[2]

Baranavichy State University, Baranavichy, the Republic of Belarus, [1]evgeniy.shapovich@gmail.com , [2]shah.al.vas@gmail.com

SPEECH RECOGNITION OF EMOTIONS

In this paper, we conduct an extensive comparison of different approaches to speech recognition systems based on emotions. The analyses were performed on audio recordings from the Ryerson emotional speech and songs audio-visual database. After

preprocessing the raw audio files, such features as Log-Mail, spectrogram, Mel-Frequency cepstral coefficients (MFCC), pitch, and energy were counted. The significance of these characteristics for emotion classification is compared using methods such as Long Short Term Memory (LSTM), convolutional neural networks (CNN), and hidden Markov models (HMM).

Neural networks (DNNs). According to the 14-class classification (2 genders × 7 emotions) task, 68 % accuracy was achieved with a 4-layer 2-dimensional CNN using a Log-Mel spectrogram. We also observe that in emotion recognition, the choice of sound characteristics affects the results much more than the complexity of the model.

Key words: emotions; neural networks; recognition; artificial intelligence; automation.

Введение. Согласно докладу Организации Объединенных Наций [1], в ближайшие пять лет все большее число людей будут взаимодействовать с голосовыми помощниками, чем со своими партнерами. С распространением виртуальных помощников (VPA), таких как Siri, Alexa и Google Assistant, в наших повседневных взаимодействиях они выполняют роль быстрого и точного ответа на наши вопросы и выполнения наших запросов. Хотя эти помощники понимают наши команды, они недостаточно искусны в распознавании нашего настроения и реагирует соответственно. Поэтому важно разработать эффективную систему распознавания эмоций, которая может расширить возможности этих помощников и революционизировать всю отрасль.

Речь — это форма общения, способная эффективно передавать информацию. Она содержит два типа информации, а именно лингвистическую и паралингвистическую. Первая относится к вербальному содержанию, лежащему в основе языкового кода, в то время как вторая относится к неявной информации, такой как язык тела, жесты, выражения лица, тон, высота тона, эмоции и т. д. Паралингвистические характеристики могут помочь понять психическое состояние человека (эмоции), пол, отношение, диалект и многое другое [2]. Записанная речь имеет ключевые особенности, которые могут быть использованы для извлечения информации, такой как эмоции, структурированным способом. Получить такую информацию было бы бесценно для облегчения более естественных разговоров между виртуальным помощником и пользователем, поскольку эмоции окрашивают повседневные человеческие взаимодействия.

Существует два широко используемых представления эмоций: непрерывное и дискретное. В непрерывном представлении эмоция

высказывания может быть выражена в виде непрерывных значений по нескольким психологическим измерениям. Эмоцию можно охарактеризовать в двух измерениях: активация и валентность. Активация — это количество энергии, необходимое для выражения определенной эмоции. Исследования показали, что радость, гнев и страх могут быть связаны с высокой энергией и высотой речи, тогда как печаль может быть связана с низкой энергией и медленной речью. Валентность дает больше нюансов и помогает различать такие эмоции, как гнев и счастье, поскольку повышенная активация может указывать и на то, и на другое. В дискретном представлении эмоции могут быть дискретно выражены в виде определенных категорий, таких как гнев, печаль, счастье и т. д.

Это исследование фокусируется на выявлении наилучшей звуковой функции и архитектуры модели для распознавания эмоций в речи. Эксперименты проводились на базе данных «Аудиовизуальной базы данных эмоциональной речи и песен Райерсона (RAVDESS)» [3]. Надежность модели оценивалась путем прогнозирования эмоций речевых высказываний на совершенно другом наборе данных-наборе данных “Toronto Emotional Speech Set (TESS)”. Четырехслойная архитектура 2D-CNN с функциями Log-Mel Spectrogram audio дала максимальную точность 70 % на тестовом наборе и 62 % на наборе данных TESS.

Основная часть. Существует три основных компонента проектирования нейронной сети: выбор набора данных эмоциональной речи, выбор признаков из аудиоданных и классификаторы для обнаружения эмоций [4]. Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset — это проверенная мультимодальная база данных эмоциональной речи и песен. Эта гендерно сбалансированная база данных состоит из голосов 24 профессиональных актеров, каждый из которых выполняет 104 уникальных вокализации с эмоциями, которые включают в себя: счастье, печаль, гнев, страх, удивление, отвращение, спокойствие и нейтральность [3]. Каждый актер разыгрывал по два высказывания для каждой эмоции: «дети разговаривают у двери» и «собаки сидят у двери». Эти утверждения также были записаны в двух различных эмоциональных интенсивностях, нормальной и сильной, для каждой эмоции, за исключением нейтральной. Существует в общей сложности 1 440 речевых высказываний и 1 012 песенных высказывания.

Из рисунка 1 мы видим, что примерно 73 % выбранных эмоций были хорошо разыграны актерами, что обеспечивает надежность классификации эмоций и аудиоконтента. Кроме того, мы также наблюдаем, что людям-оценщикам было трудно различать нейтральные и спокойные эмоции. Распределение данных по полу и классам эмоций показано на рисунке 2.

		Actor intended emotion							
		Neutral	Calm	Happy	Sad	Angry	Fearful	Disgust	Surprise
Rater chosen emotion	Neutral/Calm	86.6	69.9	14.25	17.12	4.03	4.5	4.36	7.03
	Happy	0.63	17.27	68.44	1.48	0.23	0.59	0.59	6.56
	Sad	4.65	6.06	2.29	60.85	1.02	6.58	8.65	0.76
	Angry	3.82	1.02	1.79	2.9	81.32	4.79	6.48	2.78
	Fearful	0.63	0.66	1.67	9.64	1.39	70.71	2.31	2.22
	Disgust	1.15	1.46	0.78	3.09	8.37	1.81	69.77	3.28
	Surprise	0.28	0.33	7.88	0.69	1.2	7.76	4.13	72.29
	None	2.26	3.3	2.9	4.24	2.45	3.26	3.72	5.07

Рисунок 1 — Проверка набора данных RAVDESS (точность распознавания), % [3]

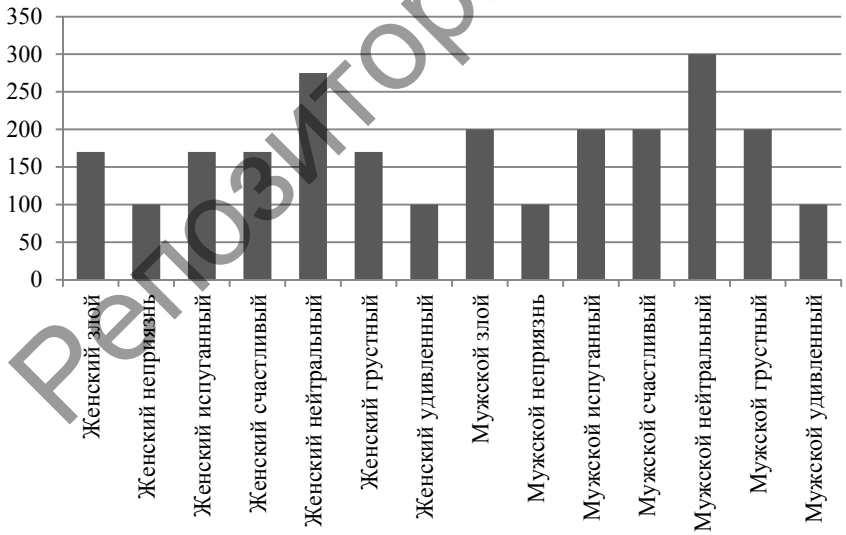

Рисунок 2 — Распределение данных по полу и эмоциям, количество аудиозаписей

В наборе данных есть недостатки. Во-первых, способ выражения эмоций сильно зависит от языка, акцента, диалекта и культурного фона. Модель, обученная распознавать эмоции на английском наборе данных, может быть не в состоянии идентифицировать эмоции в речевых высказываниях на китайском/индийском языках. Набор данных RAVDESS, естественно, страдает предвзятостью отбора, поскольку набор данных был создан с использованием 24 англоговорящих актеров из Канады, которые демонстрируют сильные североамериканские характеристики. Во-вторых, набор данных был создан с использованием обученных актеров, а не с использованием естественных примеров эмоций. Из-за этого ограничения модель должна быть тщательно проверена перед развертыванием. Последним ограничением было включение только двух утверждений, ограничивающих лексическую вариативность базы данных. Таким образом, бремя понимания эмоциональных паттернов, а не самих слов, в значительной степени ложится на разработку и моделирование функций.

Каждый аудиофайл содержит числовой идентификатор из 7 частей, каждый из которых обозначает модальность, вокальный канал, эмоцию, эмоциональную интенсивность, высказывание, повторение и актера соответственно. Соглашение об именовании следовало шаблону, в котором нечетные и четные актеры обозначали мужской и женский пол соответственно. Целевой переменной является эмоция, к которой была отнесена аудиозапись.

Каждая запись длилась примерно 3 секунды. Аудиозаписи были обрезаны, чтобы убрать тишину как в начале, так и в конце. Благодаря тому, что аудиозаписи были профессионально записаны, в данных существовали мельчайшие шумовые паттерны. Чтобы отфильтровать шум от поврежденного сигнала и обеспечить четкую версию лежащего в основе сигнала использовался фильтр. После фильтрации наблюдается значительное улучшение качества набора данных без ущерба для какого-либо важного аудиоконтента.

Звуковые функции могут быть широко классифицированы на две категории, а именно функции временной области и функции частотной области. Особенности временной области включают кратковременную энергию сигнала, скорость пересечения нуля, максимальную амплитуду, минимальную энергию, энтропию энер-

гии. Эти функции очень легко извлекаются и обеспечивают более простой способ анализа аудиосигналов. В условиях ограниченных данных особенности частотной области выявляют более глубокие паттерны в звуковом сигнале, которые потенциально могут помочь нам идентифицировать лежащую в основе сигнала эмоцию. Особенности частотной области включают спектрограммы, кепстральные коэффициенты Mel-частоты (далее — MFCC), спектральный центроид, спектральный роллофф, спектральную энтропию и коэффициенты цветности.

В ходе исследовательского анализа данных был проведен обширный анализ каждого признака. Однако для достижения поставленных целей мы ограничились двумя основными характеристиками, а именно кепстральными коэффициентами MFCC и Мел-спектрограммами.

MFCC — это представление кратковременного спектра мощности звука путем преобразования звукового сигнала через серию шагов, имитирующих человеческий слух. Шкала Mel важна тем, что она лучше аппроксимирует человеческое восприятие звука в отличие от линейных шкал [4]. В теории фильтра-источника источником являются голосовые связки, а фильтр представляет голосовой тракт. Длина и форма голосового тракта определяют, как звук выводится из человека, а кепструм (спектр логарифма спектра) может описывать фильтр, то есть представлять звук структурированным образом [4]. MFCC — это коэффициенты, которые захватывают огибающую кратковременного спектра мощности.

Спектрограмма — это временная и частотная репрезентация звукового сигнала. Различные эмоции демонстрируют различные паттерны в энергетическом спектре. Mel-спектрограмма — это представление звукового сигнала в Мел-масштабе. Логарифмическая форма mel-спектрограммы помогает лучше понять эмоции, потому что люди воспринимают звук в логарифмическом масштабе.

Таким образом, спектрограмма log-mel соответствует представлению времени и частоты log-mel, которое было получено при вычислении MFCC. Функции MFCC и логарифмические спектрограммы могут быть представлены в виде изображений, и эти изображения могут быть переданы в методы глубокого обучения, такие

как сверточные нейронный сети [5] (далее — CNN) и рекуррентные нейронные сети, чтобы классифицировать эмоции звука.

MFCC были извлечены с размером окна 10 мс и длиной прыжка 5 мс. Количество MFCC на кадр также было настроено в нашем анализе, и было обнаружено, что прирост производительности не превышает 25 функций на кадр. Кроме того, такие функции, как высота тона, величина и среднеквадратичная энергия, их дельты и дельта-дельты также были добавлены в MFCC. 128 Log-Mel спектрограммы были извлечены из входных аудиосигналов с размером окна и длиной прыжка 0,014 сек и 0,0035 сек соответственно.

Эксперименты по распознаванию эмоций широко подразделяются на эксперименты, зависящие от говорящего и независимые от говорящего. Эксперименты, зависящие от говорящего, содержат звуковые примеры различных эмоций одного и того же актера в наборах данных обучения, валидации и тестирования. Это означает, что набор данных разбивается случайным образом, и обучающие данные могут чрезмерно соответствовать одному конкретному актеру, что приводит к смещению модели. Следовательно, случайное разделение является одной из форм утечки данных для этой задачи и нецелесообразно.

С другой стороны, независимые от говорящего эксперименты — это эксперименты, в которых обучающие, валидационные и тестовые данные состоят из звуковых экземпляров от разных участников. Независимое от говорящего обучение гарантирует, что модели устойчивы и смогут идентифицировать эмоции независимо от актера. Таким образом, аудио-экземпляры актеров 1—20, актеров 21 и 22 и актеров 23 и 24 использовались для обучения, валидации и тестирования соответственно. Поскольку и валидационные, и тестовые наборы данных содержат экземпляры как мужских (нечетные актеры), так и женских (четные актеры) актеров, мы позаботились о том, чтобы модели не были настроены на определенный пол. Кроме того, разделив его таким образом, мы также обеспечили отсутствие утечки характеристик актера в наборе данных.

Учитывая, что данные распределены почти одинаково, точность является допустимой метрикой для сравнения производительности моделей. Таким образом, метрика выбора модели

была выбрана невзвешенная точность. Для обучающих моделей с входами фиксированного размера все входные аудиосигналы бьии разделены на 3 секунды путем их соответствующей обрезки или заполнения.

Все модели были обучены для 100 эпох с различными размерами партий (в зависимости от сложности архитектуры). Первоначально модели обучались с помощью оптимизатора стохастического градиентного спуска (далее — SGD). Из-за более медленной конвергенции с SGD был использован оптимизатор ADAM для более поздних экспериментов с параметрами по умолчанию. Модели были сохранены путем мониторинга точности набора валидации.

Мы обучили базовую архитектуру глубокой нейронной сети (далее — DNN) с функциями MFCC и обнаружили, что сеть не может понять записи из класса «Удивление». Это может быть связано с нехваткой данных в классе. Следовательно, класс «Удивление» был удален из набора данных, и дальнейший анализ проводился только с использованием 6 классов — злой, грустный, нейтральный, неприязнь, счастливый и испуганный. Увеличение числа MFCC и добавление дельт дало нам лучшую производительность DNN.

Также параллельно были реализованы 1D CNN и 2D CNN на 6 классах классификации эмоций и 12 классах классификации «гендер + эмоция», и было замечено, что включение гендера дает лучшую производительность. Поскольку CNN являются естественными экстракторами функций, архитектуры 1D CNN и 1D CNN-LSTM также обучались на необработанном аудиовходе.

2D CNN были реализованы на таких инженерных функциях, как MFCC и Log-mel спектрограмма. Обучение 2D CNN началось с 2 сверточных слоев с фильтрами размером 3 × 3 и фильтрами максимального объединения с размером 2 × 2 с шагом 2. Они были настроены путем добавления большего количества сверточных слоев и увеличения размеров фильтров в начальных слоях. Было обнаружено, что увеличение глубины за пределами 4 слоев не улучшает производительность. В ходе последующих экспериментов к нашему анализу был добавлен класс «Удивление».

Конечным слоям объединения был присвоен более высокий размер фильтра и шаг 4. Это помогло уменьшить количество па-

раметров в полностью связном слое, когда сверточная карта объектов сглажена.

Мы также провели эксперименты с другой архитектурой CNN, где вместо выравнивания конечной сверточной карты объектов было выполнено глобальное среднее объединение для получения карт объектов фиксированной длины. Это решило проблему большого количества параметров в полностью связанных слоях.

Дополнительным преимуществом этой архитектуры является то, что она может обрабатывать входные данные переменного размера, что типично для речевых данных.

Глобальный средний слой объединения позволил нам увеличить размер фильтров в начальных сверточных слоях, что привело к повышению производительности. Лучшая модель по прогнозированию 14 классов была получена с фильтрами 12 × 12 и 7 × 7 в первом и втором слоях соответственно. График зависимости точности обучения от количества эпох обучения представлен на рисунке 3.

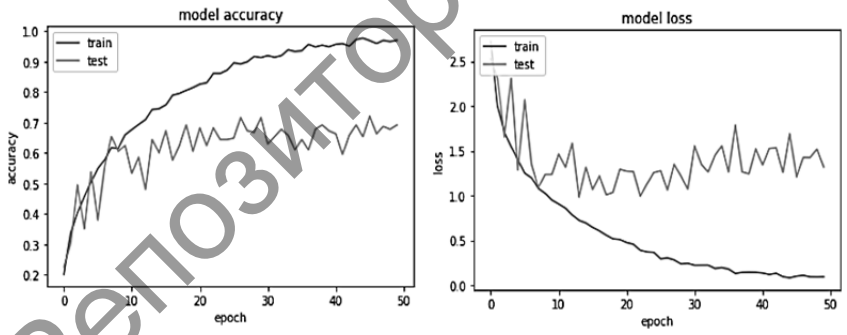

Рисунок 3 — График зависимости точности
и потерь обучения от количества эпох обучения

Заключение. Мы провели углубленный анализ различных методов конструирования признаков и моделирования для распознавания эмоций. Мы получаем гораздо лучшие результаты с помо-

щью инженерных функций, таких как MFCCs и Log-mel спектрограмма, чем необработанный аудиовход, что, скорее всего, связано с нехваткой данных.

Хотя MFCC являются широко используемыми функциями для распознавания эмоций на основе речи, мы считаем, что функции логарифмической спектрограммы определенно лучше справляются с этой задачей. Мы наблюдаем, что гендерная классификация эмоций приводит к более высокой производительности. Это происходит из-за разницы в высоте и энергии среднего мужского и среднего женского голоса, что делает паттерны мужских эмоций отличными от женских.

Кроме того, добавление таких функций, как высота тона и энергия, в MFCC улучшило характеристики модели, что означает, что в функциях MFCC отсутствует информация о высоте тона и энергии, которая имеет решающее значение для прогнозирования эмоций. Мы также наблюдаем, что 2D CNN дает лучшую производительность (86 %). Как уже упоминалось ранее, более низкая точность может быть объяснена субъективным характером восприятия эмоций человеком, что существенно усложняет нашу проблему.

Список цитируемых источников

1. United Nations Educational, Scientific, and Cultural Organization [Electronic resource] : I'd blush if I could: closing gender divides in digital skills through education / EQUALS Skills Coalition . — ЮНЕСКО, 2019. — Mode of access: http://unesdoc.unesco.org/images/0021/002170/217073e.pdf . — Date of access: 15.03.2021.

2. *Ахманова, О. С.* Паразык // Словарь лингвистических терминов / О. С. Ахманова — Изд. 4-е, стер. — М. : КомКнига, 2007. — 576 с.

3. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [Electronic resource] : A dynamic, multimodal set of facial and vocal expressions in North American English / S. R. Livingstone, F. A. Russo. — HUNGARY : University of Pecs Medical School, 2017. — Mode of access: https://zenodo.org/record/1188976#.YFHHsWgzaUk/ . — Date of access: 15.03.2021.

4. *Головко, В. А.* Нейросетевые технологии обработки данных : учеб. пособие / В. А. Головко, В. В. Краснопрошин. — Минск : БГУ, 2017. — 263 с.

5. *Шапович, Е. Г.* Классификация диатомовых водорослей для определения качества воды / Е. Г. Шапович // Беларусь и Китай : многовекторность сотрудничества : материалы III науч.-практ. кругл. стола в рамках реализации цикла науч.-практ. мероприятий «Цифровой конгресс-2020», Барановичи, 13 марта 2020 г. / М-во образования Респ. Беларусь, Баранович. гос. ун-т ; редкол.: В. В. Климук (гл. ред.), А. В. Прадун (отв. ред.). — Барановичи : БарГУ, 2020. — С. 104—111.

UDC 37.01

Qinghua Yang[1], Honglu Wang[2], Yuchuan Zhang[3], Baili Feng[4]
[1, 2, 3]Northwest A&F University, College of Agronomy, State Key Laboratory of Crop Stress Biology in Arid Areas, Yangling, the People's Republic of China,
[1]qinghuayang@nwafu.edu.cn , [2]1421960977@qq.com , [3]458257924@qq.com
[4]Northwest A&F University, College of Agronomy, Yangling, the People's Republic of China, fengbaili@nwsuaf.edu.cn

RESEARCH ON THE TRAINING MODE OF INTERNATIONAL AGRICULTURAL TALENTS BASED ON THE CONSTRUCTION OF OVERSEAS SCIENCE AND TECHNOLOGY DEMONSTRATION PARKS

Since the "Belt and Road" initiative was put forward, China and Belarus have increasingly frequent agricultural trade cooperation. Talents are the fulcrum and key to the construction of the "Belt and Road". However, the current training of international agricultural talents in Chinese universities still has problems such as a single training structure and low level of internationalization. In order to promote the agricultural cooperation between China and Belarus and adapt to the extensive advancement and long-term development of the "Belt and Road" construction, the training of international talents in agricultural universities in China is very important. China and Belarus should rely on the Belarusian Agricultural Science and Technology Demonstration Park, build an international exchange platform, create an innovative training model for agricultural talents with an international perspective, and provide strong talent support for bilateral agricultural cooperation and development.

Key words: One Belt One Road; Belarus; agricultural talent training.