ИЗВЕСТИЯ РАН. ТЕОРИЯ И СИСТЕМЫ УПРАВЛЕНИЯ, 2022, № 3, с. 116–121

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

УДК 004.8

ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА ДЛЯ ОПРЕДЕЛЕНИЯ ЭМОЦИЙ НА АУДИОЗАПИСИ С ПОМОЩЬЮ МЕЛ-СПЕКТРОГРАММ

© 2022 г. Л. А. Деревягин[а], В. В. Макаров[а,*], В. И. Цурков[б], А. Н. Яковлев[а]

[а]МФТИ (НИУ), Москва, Россия

[б]ФИЦ ИУ РАН, Москва, Россия

**e-mail: viktor.makarov@phystech.edu*

Поступила в редакцию 11.12.2021 г.
После доработки 20.12.2021 г.
Принята к публикации 31.01.2022 г.

Предлагается архитектура нейронной сети для решения задачи определения эмоции человека на аудиозаписи. Под эмоциями понимаются страх, радость, грусть, гнев, спокойствие и нейтральность. Для обучения используются библиотечные данные. С помощью конвертации аудиофайла в изображение спектрограмм с мел-шкалой (эмпирически выбранная логарифмическая зависимость воспринимаемой органами слуха человека громкости звуковых колебаний от их частоты) сохраняются психофизические свойства аудиозаписи и применяются методы для классификации графических файлов, в том числе слои свертки (пофрагментное умножение матриц значений пикселей на заданные матрицы с возможным уменьшением размерности картинки)

DOI: 10.31857/S0002338822030040

Введение. Распознавание эмоций человека по аудиозаписи произносимого текста является важной научно-исследовательской проблемой, которая затрагивает множество дисциплин и областей [1–3]. Эта тематика актуальна в таких сферах, как медицина, психология [4] и безопасность. В работе рассматриваются подходы к распознаванию эмоций человека на монофоническом аудиоматериале. Анализ преобразованных из аудиофайлов так называемых мел-спектрограмм осуществляется при помощи сверточных нейронных сетей. Поскольку эти специфические спектрограммы представлены в виде картинки, то используется опыт в области классификации изображений [5].

Идея данного исследования была сформирована при анализе алгоритма CREPE, представленного в [6] и являющегося продолжением работ над алгоритмами YIN [7] и pYIN [8]. Упомянутые публикации являются инновационным для задач определения частоты основного тона (также называемой ЧОТ, F0 или Fundamental Frequency) в монофоническом аудиоматериале. В центральном месте алгоритма CREPE находится сверточная нейронная сеть, производящая обучение на непосредственно аудиосигнале во временной области. Базовый подход удалось реализовать в рамках этой статьи.

1. Краткий обзор существующих подходов и постановка задачи. Рассматривая задачу распознавания эмоций человека по аудиозаписи произносимого текста, остановимся на существующих на настоящий момент решениях в этой области [9]. Несмотря на определенную субъективность при оценке такой характеристики, как проявление эмоционального состояния на голосовой фонограмме, некоторые наборы данных отвечают всем необходимым признакам.

Чаще всего для оценки эмоций выделяют просодические (характеризующие речевую мелодию, темпоральные и тембральные особенности голоса) и спектральные характеристики аудиофайла с последующей классификацией полученных данных.

Близкими к данному исследованию выступают архитектуры, разработанные в университетах Пассау в Германии [10], Калифорнии [11] и Техасса [12]. Сравнительный анализ вышеуказанных алгоритмов, основанных на CREPE, привел к выявлению следующих недостатков.

1. Использование частичного обучением учителя в [10] приводит к нестабильным промежуточным результатам и потере точности.

2. Применение генеративно-состязательных сетей в [11] подразумевает повышение качественных требований к набору данных, а также усложненному процессу обучения и генерации результатов.

3. Сложный алгоритм, описанный в [12], имеет в основе адаптацию алгоритмов обучения с помощью метода опорных векторов, примененного к синтетическим данным (автоматически сгенерированных алгоритмом), для дальнейшего использования с реальными данными (доменная адаптация). Помимо сложности имплементации такая система имеет увеличенную вычислительную стоимость и базируется на условных правилах (так называемая “rule-based система”). При наличии образцов данных на момент прогнозирования, выходящих за установленный набор, точность такой системы окажется ниже расчетной.

У всех представленных алгоритмов имеется следующий недостаток: отсутствие непосредственного анализа информации сигнала, так как обучение производится при помощи данных, либо полученных от внутренних преобразований сети, либо от препроцессирующих алгоритмов.

Учитывая описанные выше недостатки, при проектировании системы, базирующейся на CREPE, необходимо рассмотреть принцип работы трекера частоты основного тона. Данная система имеет следующие входные и выходные данные.

В качестве входных данных взяты 1024 выдержки из аудиосигнала во временной области с частотой дискретизации 22 кГц. Они обрабатываются при помощи шести сверточных слоев.

Выходными данными является тензор размерностью 2048, который затем передается на полносвязный выходной слой классификации с активирующей функцией сигмоидой размерностью в 360 нейронов. Каждый из 360 элементов выходного вектора соответствует конкретному значению высоты звука, выражаемой в центах.

Цент — единица частотного интервала, равная сотой части полутона или 1/1200 части октавы (поскольку в октаве 12 полутонов), что дает шкалу высот звука, в которой 100 центов равняется одному полутону. Таким образом данная шкала покрывает диапазон звуков с интервалами в 20 центов в диапазоне частот от 32.70 до 1975.5 Гц.

Ключевой характеристикой голосового сигнала является частота основного тона. С музыкальной точки зрения — это образующая для всех остальных звуков натурального звукоряда, а для человеческой речи — частота колебаний голосовых связок. Она присуща непосредственно их обладателю, а ее повышение воспринимается слушателем как повышение высоты звука. Таким образом, возможно следующее предположение: решение задачи определения эмоций по монофоническому аудиоматериалу можно осуществить, использовав набор инструментов, схожий с задачей определения частоты основного тона алгоритма CREPE.

Постановка задачи на распознавание выглядит следующим образом. Определяются эмоции: страх, печаль, радость, грусть, нейтральность, спокойствие. Каждой из них вводятся неотрицательные коэффициенты (вероятности), сумма которых равна единице. Цель распознавания – нахождение максимального коэффициента. Он и определяет искомую эмоцию, т.е. решает поставленную задачу.

2. Выбор архитектуры нейронной сети. В соответствии с основной идеей работы алгоритма CREPE (непосредственная работа над характеризующей сигнал графической информацией) были рассмотрены несколько вариантов сверточных нейронных сетей с некоторыми различиями во внутренней архитектуре. Данные различия включают в себя: кардинальные отличия архитектур, отличное от CREPE количество слоев свертки, разное количество групп слоев, применение дополнительных техник предотвращения эффекта переобучения (over-fitting, droput, regularization).

Нейронная сеть, как известно, носит такое название в силу того, что состоит из некоторого количества вычислительных единиц – нейронов. Эти единицы способны получать, обрабатывать и отправлять любую информацию дальше. Делятся нейроны на три основных вида (входной, выходной, скрытый) и два вспомогательных (нейроны смещения, контекстный). Чтобы улучшить обработку информации при наличии большого количества нейронов, их совмещают в слой. Они также разделяются на входной, выходной и скрытый слой. Общий принцип работы основан на том, что каждый нейрон имеет два параметра: входные и выходные данные. Дальнейшие действия сводятся к простому циклу: входной нейрон или слой получает введенную информацию, после чего обрабатывает ее и отдает на скрытый нейрон или слой. Во всех последующих скрытых нейронах или слоях информация обрабатывается, и каждая последующая передача сопровождается собранной информацией каждого нейрона или слоя. В конце функция активации нормализует все полученное и отдает на выходной нейрон или слой, который выводит результат.

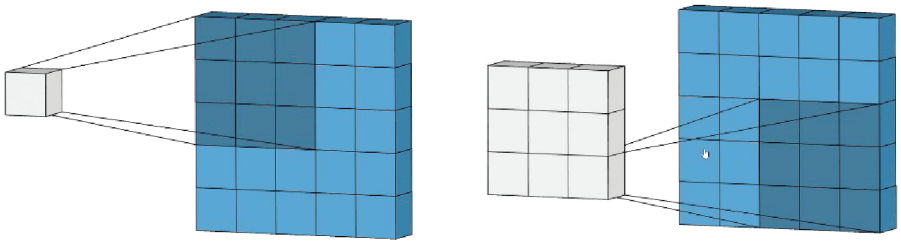

Рис. 1. Демонстрация принципа работы свертки

Чтобы решить более значительную задачу, например задачу классификации, нейроны собирают в общую систему – искусственную нейронную сеть. Как известно, при наличии в нейронной сети более одного скрытого слоя такую сеть принято называть глубокой. В большом многообразии различных архитектур были выделены и рассмотрены лишь подходящие для цели исследования. В качестве оптимальных для данной задачи изначально рассматривались: полносвязная нейронная сеть, когнитрон, перцептрон и сверточная нейронная сеть.

В полносвязной нейронной сети присутствует множество простых процессоров, которые сами по себе могут только совершать тривиальные операции. Каждому такому процессору (т.е. нейрону) назначается одна из задач: входные принимают набор данных, обрабатывающие совершают простые математические операции над набором, выходные используются для дальнейшей передачи. В итоговом счете каждому пикселю изображения ставится в отношение один нейрон. Это имеет место в большинстве вариантов таких архитектур. Такой подход в машинном обучении прост в использовании. Однако расчеты занимают большое количество времени и задействованных нейронов, а качественная оценка результатов может разниться из-за плохого качества изображения или наличия шума, не видного человеческому глазу. Упомянутые причины снимают приоритет с данного выбора.

Как известно, когнитрон и перцептрон являются двумя сходными архитектурами. Оба варианта в основе имеют принцип обработки изображения человеческим мозгом зрительной корой, но есть различия во внутренней архитектуре. В перцептроне клетки одного слоя не связаны между собой, но соседние слои полностью связаны. При обработке объекта нейроны реагируют на него и дают сигнал (по аналогии с реакцией зрительной коры мозга на попадание света на сетчатку глаза). В когнитроне имеется иерархическая многослойная организация, в которой нейроны между слоями связаны только локально. Несомненно, достоинством, общим для двух архитектур является то, что когнитрон и перцептрон дают более точные результаты, по сравнению с полносвязными нейронными сетями. Но стоит отметить, что даже малейшие изменения изображения могут восприниматься ими как совершенно новый объект изучения (что требует постоянного дополнения набора данных для более полного охвата предметной области задачи).

Как известно, в сверточной нейронной сети имеются слои, выполняющие операцию свертки. Каждый фрагмент изображения умножается на матрицу (ядро) свертки поэлементно, а результат суммируется и записывается в аналогичную позицию выходного изображения (рис. 1). Информация проходит распределение определенных свойств изображения, в которых выделяются более абстрактные детали. Структуру и распределение этих абстрактных признаков и ядро свертки нейронная сеть определяет самостоятельно в процессе обучения, обретая способность фильтрации деталей и выделения существенных признаков.

По причине того, что сверточная нейронная сеть нацелена на высокую точность распознавания образов и лучшую из предложенных работу по классификации изображения, данный вариант рассматривался как наиболее приоритетный, что выделяет эту архитектуру как самую эффективную для дальнейшей работы.

3. Выбор модели голосового аудиоматериала. Для работы с аудиофайлами и их последующей обработки предварительно необходимо рассмотреть модель создания звуковых волн в речевом тракте. Несмотря на то, что при исследовании нет возможности создать трехмерную модель траектории движения звуковой волны, достаточно описать общие характеристики данных акустических процессов с учетом аэродинамических свойств. Теория речеобразования достаточно полно описывает приведенную модель.

Как известно, речевым сигналом называется функция возбуждения с откликами линейных фильтров. В этом случае в качестве функции возбуждения выступает шум. В пределах 90–300 Гц

колеблется основной тон человеческой речи, который является уникальным для каждого отдельно взятого индивида. В пределах 90–180 Гц располагается частота мужских голосов и в пределах 185–300 Гц – частота женских и детских голосов. Набор гармоник, кратных основному тону, представляет щелчок голосовой щели. Падение уровня энергии гармоник напрямую зависит от увеличения частоты, 18 кГц – это максимальная граничная частота речевого сигнала, но для тракта достаточно частоты до 3500 Гц. При таком частотном ряде часть фонем не воспринимается человеческим ухом.

Резонансные полости речевого тракта напрямую используются щелчком голосовой щели. В этот момент часть гармоник, кратных основному тону, резонируют и созданные в спектре локальные максимумы образуют области концентрации энергии, которые называются формантами. Четыре форманты служат для создания гласных фонем, а любые другие изменения образуют согласные звуки. Все вышеперечисленное называют фонемами. Однако форманта также может служить для составления метрик на аудиоматериале речи человека, так как принадлежит к статическим характеристикам речи.

Если рассматривать образование речи как создание легкими, бронхами и трахеей акустической волны, которая образует речь посредством изменения траектории в голосовом тракте, то голосовой тракт (совокупность вышеназванных органов) можно представить как резонатор с несколькими пиками амплитудной частотной характеристики, частоты которых определяют вид произносимой фонемы и соответственно состояние человека.

Реализованный на начальном этапе исследований простой алгоритм производит перевод итогового аудиосигнала в соответствующий набор параметров в рамках описанного теоретического материала и в последствии – в графический вид. Благодаря своей информативности в сравнении с остальными вариантами была выбрана спектрограмма – двумерная диаграмма с прямой зависимостью, где по вертикальной оси показана частота, по горизонтальной оси – время, а амплитуда на определенной частоте в каждый конкретный момент времени представлена цветом.

Однако, несмотря на большую меру информативности спектрограмм, на этапе первичного обучения сверточной нейронной сети не было получено должной ориентировочной точности классификации эмоций, что привело к выдвижению гипотезы о применении психофизической шкалы.

Известно, что человеческое ухо более чувствительно к изменениям звука на низких частотах, чем на высоких. Это значит, что если частота звука изменится со 100 на 120 Гц, то человек с очень высокой вероятностью распознает это изменение. Однако изменение частоты с 10000 на 10020 Гц сложнее для восприятия человеческим ухом.

Такая особенность слуха учтена в одной из единиц измерения высоты звука — мел. Она основана на психофизиологическом восприятии звука человеком и логарифмически зависит от частоты, что непосредственно приводит к использованию мел-спектрограмм:

$$m = 1127\ln\left(1 + \frac{f}{700}\right), \tag{3.1}$$

где m – высота звука в мелах, f – частота звука в Гц.

Мел-спектрограмма – это вариант спектрограммы, где частота выражена не в Гц (рис. 2), а в мелах (рис. 3). Переход к мелам происходит с помощью применения шкалирования исходной спектрограммы.

4. Анализ мел-спектрограмм. Для обучения был выбран набор данных Ravdess [13], представляющий из себя 4-секундные аудиозаписи, на которых актеры произносят два предложения на английском языке по 2 раза каждое (обычное произношение и нараспев). Предложения произносятся 2 раза для записи сильного и слабого проявления. Каждая запись длится в среднем 4 с, в первой и последней секундах присутствует лидирующий и заключительный отрезок без звука. Аудиоматериал записан в стереоформате, частота семплирования равна 48 кГц. Каждый аудиофайл имеет метку с эмоцией (нейтральность, радость, спокойствие, грусть, злость, страх, отвращение, удивление), которую испытывал актер при записи. При помощи вспомогательных библиотеки функций librosa и matplotlib производится первичная обработка файлов: приведение материалов к моноформату, децимация аудиофайлов до частоты 22 кГц, подготовка изображения с мел-спектрограммой аудизаписи с разрешением 640 на 480 пикселей (рис. 3). Именно на этом этапе все аудизаписи были преобразованы в изображения, в которых по горизонтальной оси приведено время, по вертикальной оси — частота. Третье измерение с указанием амплитуды на определенной частоте в конкретный момент времени представлено интенсивностью желтого цвета каждой точки изображения.

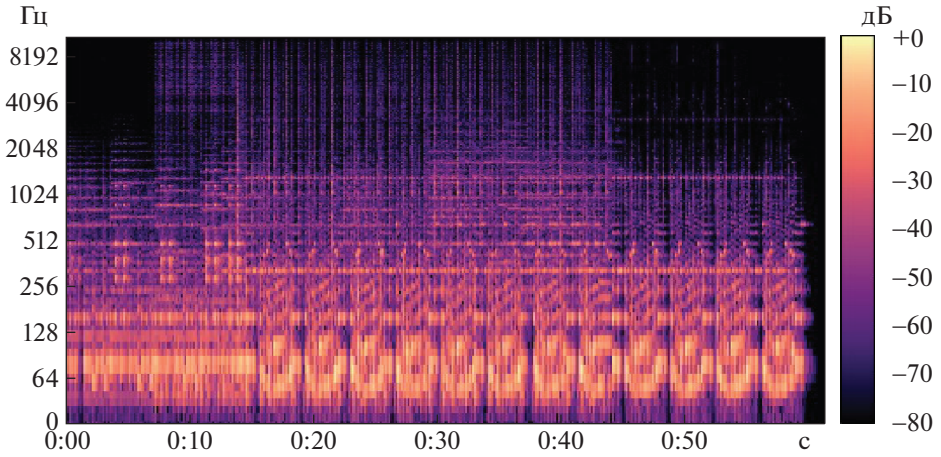

Рис. 2. Пример изображения со спектрограммой аудиофайла

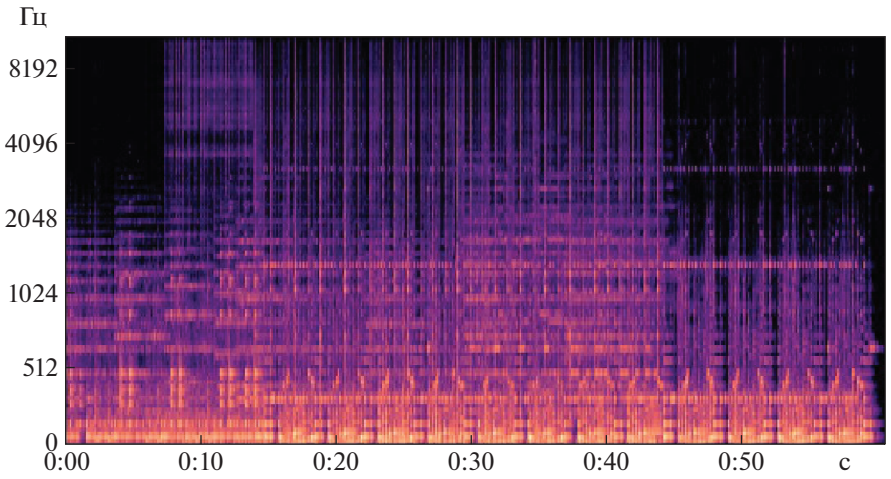

Рис. 3. Пример изображения с мел-спектрограммой аудиофайла

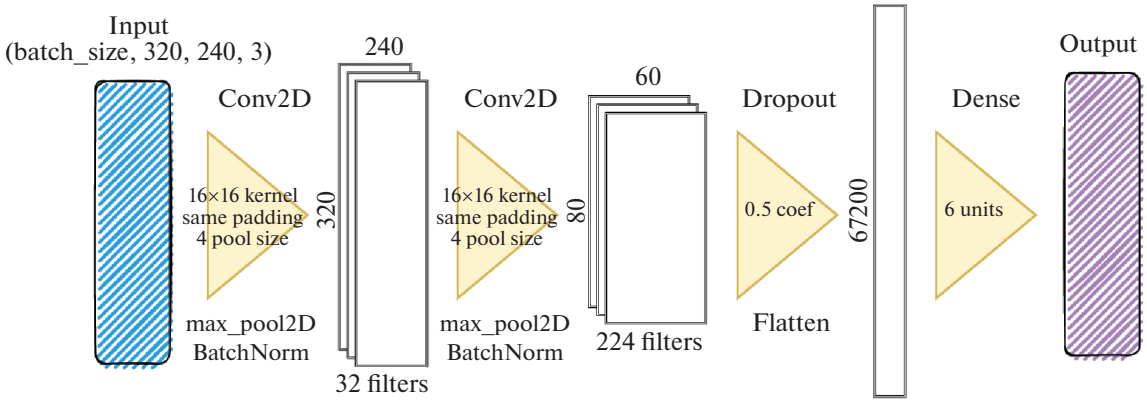

Рис. 4. Архитектура нейронной сети

Посредством библиотеки Keras производится загрузка, нормализация, разделение на обучающую и тестовую выборки, обучение и выбор наилучших моделей, составление матрицы ошибок и классификационного отчета по распознаванию эмоций. В результате тестирований архитектур для обучения нейронной сети с применением методов оптимизации гиперпараметров были получены следующие значения точности валидации (количество правильно распознанных эмоций в валидационной выборке в процентах):

1conv_2blocks – 80.2,
2conv_blocks – 78.29,
1conv_3blocks – 72.56,
2conv_3blocks – 78.81.

По результатам обучения отмечено следующее: сверточные сети, имеющие в своей архитектуре один сверточный слой, показали менее подверженный отклонениям результат по сравнению с сетями, имеющими два сверточных слоя;

Наивысший показатель валидационной точности имеет сеть с архитектурой, представленной на рис. 4:

входной слой (Input),

два блока, каждый из которых состоит из сверточного слоя, слоя нормализации пакетов (Batch Normalization) и функции активации ReLU,

слой выброса (Dropout),

выравнивающий слой (Flatten),

полносвязный слой (Dense) с функцией активации softmax.

Заключение. К сожалению, на сегодняшний день отсутствуют русскоязычные наборы данных для оценки эффективности представленной системы. Исследование может быть дополнено по факту подготовки таких материалов. В дальнейших исследованиях планируется оценить возможность повышения точности классификации при обогащении аудиосигналов информацией об уровне стресса исследуемых лиц.

В настоящее время проводятся работы с русскоязычными текстами, и по предварительным оценкам точность распознавания эмоций будет более высокой. Рассматривается возможность повышения точности распознавания с добавлением к текущим данным информации об уровне стресса говорящего, которые можно получить при помощи полиграфа.

СПИСОК ЛИТЕРАТУРЫ

1. *Александров А.А., Кирпичников А.П., Ляшева С.А., Шлеймович М.П.* Анализ эмоционального состояния человека на изображении // Вестн. технологического ун-та. 2019. Т. 22. № 8. С. 120–123.
2. *Заболеева-Зотова А.В.* Развитие системы автоматизированного определения эмоций и возможные сферы применения // Открытое образование. 2011. № 2. С. 59–62.
3. *Люсин Д.В.* Современные представления об эмоциональном интеллекте // Социальный интеллект: теория, измерение, исследования / Под ред. Д.В. Люсина, Д.В. Ушакова. М.: Изд-во Ин-та психологии РАН, 2004. С. 29–36.
4. *Гранская Ю.В.* Распознавание эмоций по выражению лица: Автореф. дис. ... канд. психологических наук по специальности 09.00.01. СПб., 1998.
5. *Bhatnagar S.,Ghosal D., Kolekar M.H.* Classification of Aashion Article Images Using Convolutional Neural Networks // Fourth Intern. Conf. on Image Information Processing (ICIIP). Waknaghat 2017. P. 1–6. https://doi.org/10.1109/ICIIP.2017.8313740.
6. *Kim J.W., Salamon J., Li P., Bello J.P.* CREPE: A Convolutional Representation for Pitch Estimation // Music and Audio Research Laboratory. N. Y.: Center for Urban Science and Progress, New York University, 2018. URL: https://arxiv.org/pdf/1802.06182.pdf
7. *Cheveigne A., Kawahara H.* YIN, A Fundamental Frequency Estimator for Speech and Music // Ircam-CNRS. Wakayama University, 2002. URL: http://recherche.ircam.fr/equipes/pcm/cheveign/ps/2002_JASA_YIN_proof.pdf
8. *Mauch M., Dixon S.* PYIN: A Fundamental Frequency Estimator Using Probabilistic Threshold Distributions. London: Queen Mary University of London, Centre for Digital Music, 2014. URL: http://matthiasmauch.de/_pdf/mauch_pyin_2014.pdf
9. *Sonmez Y.Ü.,Varol A.* New Trends in Speech Emotion Recognition // 7th Intern. Sympos. on Digital Forensics and Security (ISDFS). Barcelos 2019. P. 1–7. https://doi.org/10.1109/ISDFS.2019.8757528
10. *Deng J., Xu X., Zhang Z., Frühholz S., Schuller B.* Semisupervised Autoencoders for Speech Emotion Recognition // IEEE/ACM Transactions on Audio, Speech, and Language Processing. V. 26. № 1. P. 31–43. 2018. https://doi.org/10.1109/TASLP.2017.2759338
11. *Chang J., Scherer S.* Learning Representations of Emotional Speech with Deep Convolutional Generative Adversarial Networks // IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP). Toronto. 2017. P. 2746–2750. https://doi.org/10.1109/ICASSP.2017.7952656
12. *Abdelwahab M., Busso C.* Incremental Adaptation Using Active Learning for Acoustic Emotion Recognition // IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP). Toronto. 2017. P. 5160–5164. https://doi.org/10.1109/ICASSP.2017.7953140
13. *Livingstone S.R., Russo F.A.* The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A Dynamic, Multimodal Set of Facial and Vocal Expressions in North American English // PLoS ONE. 2018. V. 13. № 5. C. 1–35.