© 2022 г. А.О. ИСХАКОВА, канд. техн. наук (shumskaya.ao@gmail.com),
Д.А. ВОЛЬФ, канд. техн. наук (runsolar@mail.ru),
Р.В. МЕЩЕРЯКОВ, д-р техн. наук (mrv@ieee.org)
(Институт проблем управления им. В.А. Трапезникова РАН, Москва)

СПОСОБ СНИЖЕНИЯ РАЗМЕРНОСТИ ПРОСТРАНСТВА ПРИЗНАКОВ ПРИ РАСПОЗНАВАНИИ РЕЧЕВЫХ ЭМОЦИЙ С ИСПОЛЬЗОВАНИЕМ СВЕРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ[1]

Рассматриваются архитектуры сверточных нейронных сетей, используемые для оценки эмоционального состояния человека по его речи. Решается задача повышения эффективности распознавания эмоций за счет снижения вычислительной сложности данного процесса. Для этого предлагается способ преобразования входных данных в форму, подходящую для алгоритмов машинного обучения.

Ключевые слова: распознавание речевых эмоций, речевой сигнал, звук, идентификация эмоционального состояния, выявление агрессии, классификация речевых сигналов, социо-киберфизическая система, сверточная нейронная сеть.

DOI: 10.31857/S0005231022060046, **EDN:** ACJOSQ

1. Введение

Основным средством человеческого общения является речь, которая содержит характеристические параметры, отражающие в том числе психоэмоциональное состояние говорящего. Распознавание эмоций человека играет важную роль во взаимодействии человека с компьютером, так как оно является дополнительным каналом информации. У людей распознавание эмоций является естественной частью речевого общения, в то время как способность распознавать автоматически с помощью программируемых устройств все еще остается предметом исследований. Возможность автоматического определения эмоций по голосу и речи человека необходима для развития успешных диалоговых систем [1], например, в процессах обучения, мониторинга пожилых людей, людей с ограниченными возможностями, в системах интерактивного развлечения и т.д. Задача идентификации эмоционального состояния человека востребована в различных сферах: телекоммуникации, индустрии развлечений, обучении, медицине и др.

[1] Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований в рамках проекта № 18-29-22104.

Решение задачи автоматического распознавания речевых эмоций (РРЭ, англ. SER) с помощью вычислительных систем является предметом исследований ученых на сегодняшний день [2]. Это позволит более эффективно решать задачи определения эмоциональной составляющей мультимедиа материалов, распространяющихся в виртуальной среде. Автоматический анализ содержимого, выявление гнева, злости, агрессии в видео- и аудиоматериалах позволит решить задачу классификации разнородного Интернет-контента по степени его деструктивного воздействия на пользователя [3, 4]. Разработка методов РРЭ в данном случае должна отвечать требованиям, продиктованным соответствующей платформой применения — исследовать материалы быстро и не затрачивая значимых вычислительных ресурсов. При таких условиях становится реальным создание социо-киберфизической системы управления и мониторинга информации в целях противодействия проявлению деструктивного воздействия на пользователей. Повышение скорости вычислений и снижение их сложности возможны за счет уменьшения размерности обучающих данных.

Цель данной статьи — представление нового способа снижения размерности пространства входных признаков, использующего современные сверточные нейронные сети, в задачах распознавания речевых эмоций.

2. Создание обучающих выборок и предобработки при разработке систем РРЭ

В качестве обучающих выборок (данных) в задачах РРЭ обычно используются признаки в виде коэффициентов LPC (Linear Predictive Coding — кодирование с линейным прогнозированием), LPCC (линейные прогнозирующие кепстральные коэффициенты) и MFCC (Mel Frequency Cepstral Coefficients — частотные кепстральные коэффициенты мел). Создается вектор признаков для каждого высказывания путем анализа глобальной статистики (среднее, медиана и т.д.) по всем кадрам [5]. Качество извлечения признаков напрямую влияет на точность распознавания речевых эмоций. Наиболее распространенный метод извлечения признаков включает MFCC [6]. Признаки MFCC получаются в результате применения кратковременного преобразования Фурье (STFT) к исходному сигналу с использованием типа постобработки, которая включает кепстральный анализ. Подробное описание процесса извлечения MFCC-признаков рассматривается в [7, 8]. MFCC были доминирующими функциями, используемыми для распознавания речи сверточными нейронными сетями (СНС). Успех использования СНС был обусловлен их способностью представлять спектр амплитуд речи в компактной форме в качестве информации для обучения и распознавания. Между тем MFCC содержат не только информацию об эмоциональных характеристиках, но и важную информацию о говорящем. Исследования, направленные на то, какие характерные признаки эмоций извлекать из речевого сигнала, имеют большое значение [9].

Недостатком является сложность качественной оценки признаков, что может влиять на снижение точности распознавания. Трудно гарантировать, что хорошие результаты могут быть достигнуты за счет использования различных баз данных, так как люди выражают эмоции по-разному, а признаки однозначного определения эмоций отсутствуют. Успех и производительность методов машинного обучения во многом зависят от выбора представленных данных [10, 11].

На основе выделяемого набора информативных признаков строится классификатор, который обучается на предварительно подготовленном наборе звуковых фрагментов. Наиболее популярными техниками классификации являются следующие: поиск ближайших соседей, метод опорных векторов, скрытые марковские модели, модель смеси нормальных распределений, модели на основе нечеткой логики, байесовские классификаторы максимума вероятности [12].

Классификация эмоциональных состояний производится в соответствии либо с задачами построения анализатора (оценки удовлетворенности, уровня стресса, усталости и т.п.), либо с выбранной моделью описания (набор базовых эмоций, непрерывная классификация и т.п.). Как правило, с ростом числа возможных вариантов классификации точность распознавания эмоциональных состояний снижается. Поэтому количество классов, используемых для обучения, выбирается небольшим.

3. Двумерные СНС для решения задач РРЭ

Основные виды СНС основаны на двух общих архитектурах: AlexNet и GoogleNet [13]. Ключевая идея СНС состоит в локальной связности и распределении весов нейронов, которые объединяются в слои. Каждый нейрон в слое получает входные данные от набора нейронов, расположенных в предыдущем слое. Активации, вычисленные каждым ядром, собираются в матрицы, которые называются картами признаков и представляют собой фактические выходные данные сверточных слоев. Последний слой СНС — это слой вывода фактического предсказания сети, он состоит из полностью связанных нейронов так, что каждый из них принимает в качестве входных данных все выходные данные предыдущих слоев. С учетом успеха проектирования архитектур СНС для классификации двумерных массивов классификация речевых эмоций следовала тенденции использования массивов спектральных величин, известных как речевые спектрограммы. Для решения проблемы РРЭ типичная СНС также предназначена для анализа речевых характеристик, которые представлены в виде многомерного массива [11].

В меньшей степени, чем AlexNet или GoogleNet, для приложений распознавания обычно используют более простые виды СНС, основанные на архитектурах типа LeNet-5 [14]. Выбор размера как традиционной (полносвязной) нейронной сети, так и СНС является сложной задачей. Например, для достижения приемлемой эффективности должны подстраиваться размеры весовых

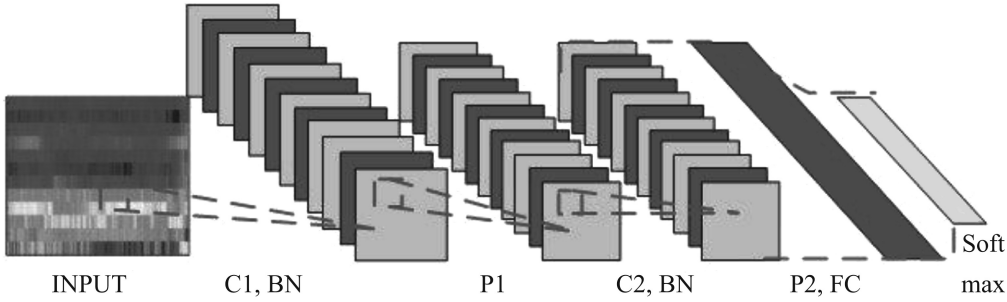

Рис. 1. Блок-диаграмма сверточной нейронной сети, предложенной Мишелем Валенти (Valenti-CNN).

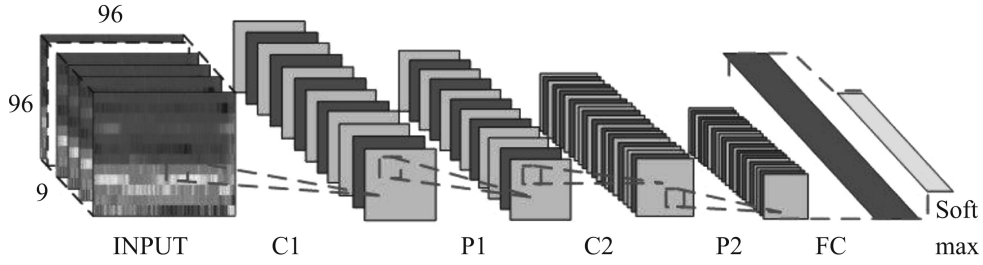

Рис. 2. Блок-диаграмма сверточной нейронной сети, предложенной Н. Хаджароласвади и Х. Демирелем (3D-CNN).

матриц в слоях. В общих случаях СНС выбираются эмпирически, в зависимости от характеристик обучающих данных, прошедших предварительную обработку.

В работе Мишеля Валенти [15] была предложена сеть CNN–Valenti-CNN (рис. 1). На вход сети подаются аудиопоследовательности в виде специально подготовленных логарифмических спектрограмм. Для этого применено кратковременное преобразование Фурье (STFT) с перекрытием окнами Хэмминга, далее абсолютные значения каждого полученного бина возведены в квадрат и применен мел-фильтр.

Еще одно интересное решение было предложено в работе Ноушини Хаджароласвади и Хасана Демиреля — 3D-CNN [16] (рис. 2). На вход сети там подается 88-мерный вектор, содержащий различные аудиохарактеристики в виде MFCC, частоты основного тона, интенсивности сигнала и т.д. Параллельно на вход подается частотный спектр каждого кадра.

Яфенг Ню и др. в [17] предложили оригинальную двумерную сверточную нейронную сеть (рис. 3), основанную на принципе визуализации сетчатки глаза и выпуклой линзы. На вход сети подаются спектрограммы разных размеров с эффектом, полученным при изменении фокусного расстояния. Таким образом достигалось увеличение числа тренировочных данных (аугментация)

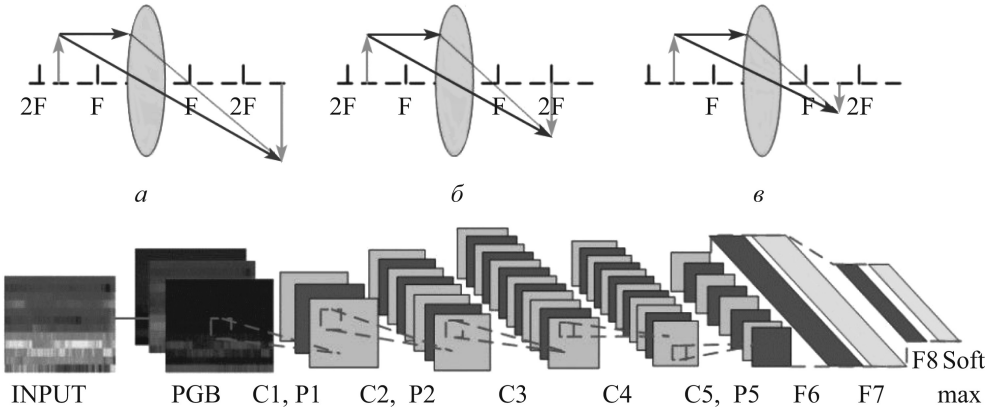

Рис. 3. Блок-диаграмма сверточной нейронной сети, основанной на принципе визуализации сетчатки глаза и выпуклой линзы.

путем изменения расстояния между спектрограммой и выпуклой линзой. Для этого были выбраны изображения в различных точках фокусирования, принадлежащие интервалам L1(F<L1<2F), L2(L2=2F) и L3(L3>2F).

В настоящее время СНС применяется к РРЭ многими исследователями, и в этом направлении уже достигнуты значительные результаты. Например:

1. Для сети, предложенной Яфенгом Ню и др., эксперименты проводились для речевых баз EmoDB [18] и SAVEE [19, 20]. Достигнута точность около 99% из семи видов эмоций.
2. Ч. Хуан и др. [21] обучили модель СНС, которая является стабильной и надежной в сложных сценах и превосходит некоторые хорошо зарекомендовавшие себя способы для решения задач РРЭ. Достигнуты результаты: точность 78% по базе SAVEE, 84% по базе Emo-DB.
3. С. Прасомфан [22] обнаружил эмоции, используя информацию внутри спектрограмм. Затем с помощью нейронной сети осуществил классификацию эмоции, используя базу EmoDB, и получил точность до 83,28% по пяти эмоциям.
4. Н. Семвал [23] предложила способ автоматического определения речевых эмоций с использованием многодоменных акустических моделей выбора и классификации. Этот подход был протестирован с базами EmoDB и BML (RED). Для мультиклассовой классификации достигается точность 80% для EmoDB и 73% для RED.

Однако для обучения глубокой нейронной сети требуется значительный объем данных, в то время как данные, предоставляемые существующими базами общих речевых эмоций, очень ограничены.

4. Одномерные СНС для решения задач РРЭ

Двумерные СНС были исследованы переходом к одномерным архитектурам, которые позволяют существенно снизить размерность обучающих при-

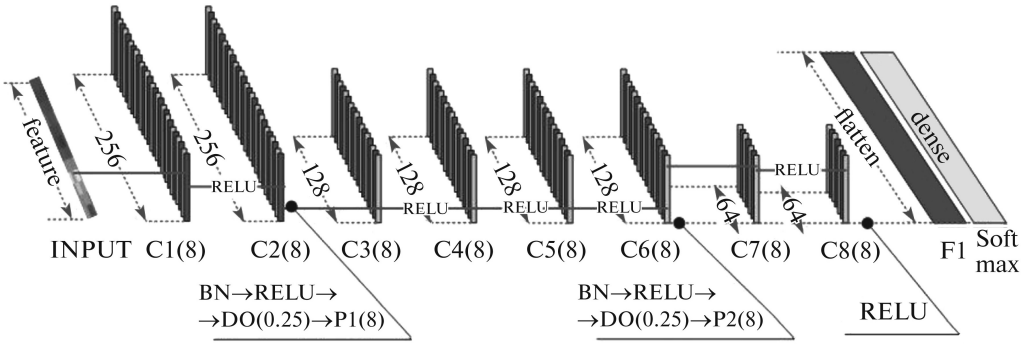

Рис. 4. Нейронная сеть (Reza 1-D CNN) для распознавания эмоций в речи, предложенная Реза Чу.

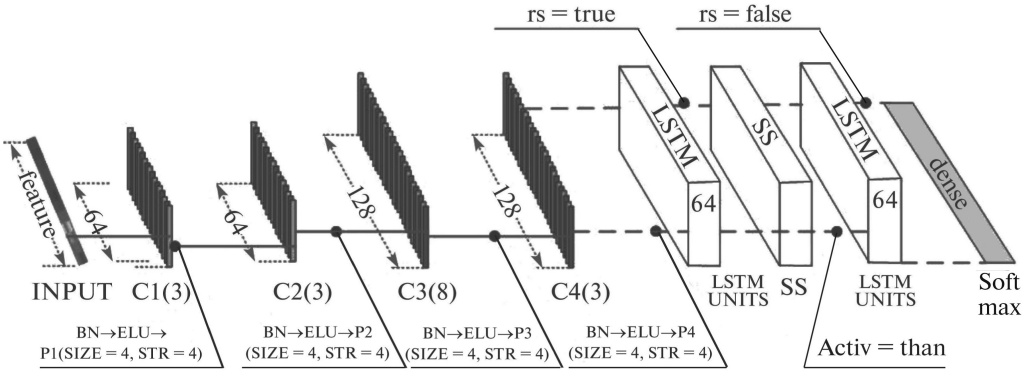

Рис. 5. Нейронная сеть (Vandana-Raian 1-D CNN-RNN) для распознавания эмоций в речи, предложенная В. Раджан.

знаков. Широкая популярность применения одномерных сверточных нейронных сетей для решения задач РРЭ возникла относительно недавно — начиная с 2019 г.

Так, Реза Чу в [24] предложил одномерную нейронную сеть — Reza 1-D CNN. Указанная нейронная сеть — это наиболее подходящая модель СНС для представления аудио-кортексиального органа слуховой системы человека в формальном описании (рис. 4, табл. 2 в справочной информации).

В это же время Ц. Чжао [25] предлагает СНС (реализация Vandana-Raian 1-D CNN-RNN [26]) с дополнительными рекуррентными слоями LSTM (рис. 5, табл. 3 в справочной информации). В отличие от модели Reza 1-D CNN в сети отсутствует регуляризация (dropout). Число сверточных ядер увеличивается в направлении выходного слоя с целью моделирования последовательностей. Полносвязный слой (Dense) получает выход из ячейки LSTM и рассчитывает логиты для каждого элемента выходной последовательности. Указанная нейронная сеть представляет собой гибридную архитектуру.

Можно заметить, что структура СНС для решения задач РРЭ имеет типичную архитектуру. Основное отличие заключается либо в расширении числа сверхточных ядер к полносвязному слою, либо к их уменьшению, а также отсутствием или наличием LSTM каскадов. В данной статье не рассматриваются параллельные архитектуры, так как такие сети нацелены на повышение точности классификации и используют иные акустические признаки дополнительно к MFFC. В настоящем исследовании допускается, что особенностей MFFC достаточно для того, чтобы решать задачу РРЭ.

5. Результаты экспериментальных расчетов

Для достижения поставленной цели была проведена собственная реализация рассмотренных выше архитектур нейронных сетей и проведено их обучение с наиболее популярными базами данных.

В первом эксперименте были обучены двумерные СНС для того, чтобы получить собственные оценки классификации. Во втором эксперименте был осуществлен переход к одномерным архитектурам. В третьем эксперименте проведены снижение размерности пространства обучающих признаков и сопоставление полученных результатов с предыдущим экспериментом.

Для тестирования двумерных СНС были выбраны следующие базы данных эмоций: Surrey Audio-Visual Expressed Emotion (SAVEE), Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [27], Toronto emotional speech set (TESS) [28], Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [29] и Emo-DB.

Для каждого акустического образца из базы были извлечены мелкепстральные коэффициенты со следующими параметрами: длительность аудио 1–4 с, частота дискретизации 44 100 Гц, 64 MFCC коэффициента.

Архитектура нейронной сети, предлагаемая Яфенг Ню и др., была заменена на архитектуру сети LeNet-5 [30]. В эксперимент была добавлена одно-

Таблица 1. Результаты тестирования сверточных нейронных сетей

Accuracy (оценки абсолютной точности для MFCC)	Сверточная нейронная сеть			
	LeNet-5	Valenti-cnn	3D-CNN	1D-cochlea-cnn
Входной слой	64 x 774	64 x 774	64 x 774	64
Обуч. параметров	5,942,666	60,614,922	1,642,954	160,202
CREMAD	0,39	0,44	0,43	0,41
SAVEE	0,48	0,5	0,5	0,6
RAVDESS	0,43	0,39	0,54	0,42
TESS	0,99	0,99	0,99	0,99
EMO-DB	0,34	0,13	0,31	0,64
UNITED	0,67	0,71	0,74	0,68

мерная СНС (1D-cochlea-cnn), рассматриваемая в [31]. После обучения нейронных сетей были получены результаты, которые представлены в табл. 1. Числовые значения в таблице показывают абсолютную точность классификации каждой из СНС для соответствующей базы. Решения для баз CREMAD, SAVEE, RAVDESS, TESS и Emo-DB являются частными случаями, а мультилингвальное решение United — общим (объединенная база).

Результаты экспериментов с применением одномерных сверточных нейронных сетей показывают, что одномерные СНС для задач РРЭ не уступают двумерным аналогам. В [31] представлены эксперименты с одномерной сверточной сетью для задачи распознавания эмоционального состояния агрессии, где достигается точность в 75%.

В эксперименте каждый признак – это массив, состоящий из 49 536–131 072 элементов. В общем случае на вход двумерных СНС подаются матрицы размерностями 32, ..., 64 на 774, ..., 2048, ..., N. Для снижения размерности пространства признаков была принята гипотеза о том, что признак, задающий эмоцию в речи, сохраняется в случае усреднения мел-кепстральных коэффициентов по частотной шкале [31].

Для следующего эксперимента были выбраны две базы CREAMD и IEMOCAP, которые были объединены в единую базу. Из нее были отобраны восемь эмоций в следующих пропорциях по гендерному типу: male_happy (радость) — 671, male_angry (злость) — 671, male_sad (печаль) — 671, female_angry — 600, female_happy — 600, female_sad — 600, male_neutral — 575, female_neutral — 512.

После приведения двумерных признаков MFCC (2D-MFCC) к среднему вектору получены одномерные MFCC признаки (1D-MFCC). Длина каждого обучающего признака представляла собой массив размерностью 2048 элементов.

Данные для обучения выбранных одномерных сетей получились следующими:

— размер тренировочных признаков для обучения — 7042;

— набор тестовых признаков — 2347 (кросс-валидация);

— объем тренировочных признаков для каждой эпохи — 50.

На рис. 6 показаны одномерные MFCC-признаки для последующего машинного обучения. Полученные признаки не масштабированы по временной шкале.

После трехсот эпох обучения точность данных проверки для сети Reza-1-D-CNN варьируется в пределах 26%, а для сети Vandana-Raian-CNN-RNN — в пределах 24%. На графиках ошибки (рис. 7) заметно, что модель не способна хорошо сходиться даже с восемью целевыми классами. Однако для речевой базы RAVDESS Р. Чу декларирует, что для сети Reza-1-D-CNN достигает более 70% точности. Осуществляется это за счет упрощения модели в виде разбиения MFCC-признаков только на мужские или женские эмоции. Для сети Vandana-Raian-CNN-RNN и базы Emo-DB достигается результат

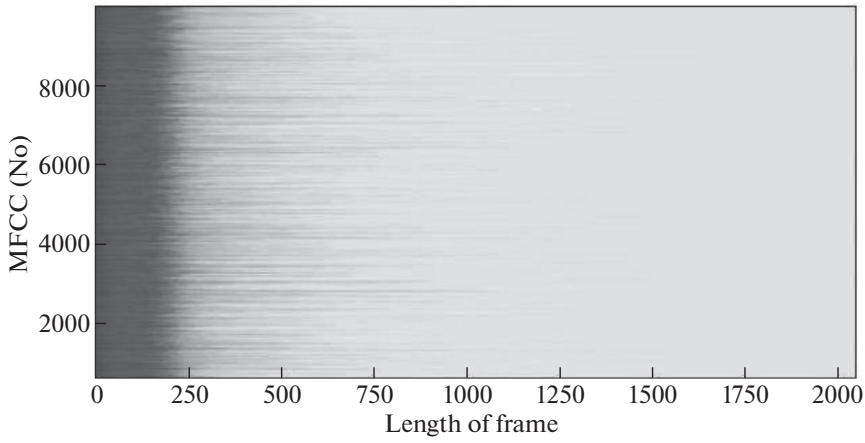

Рис. 6. Одномерные MFCC признаки на основе баз данных эмоций CREMAD и IMPOCAP.

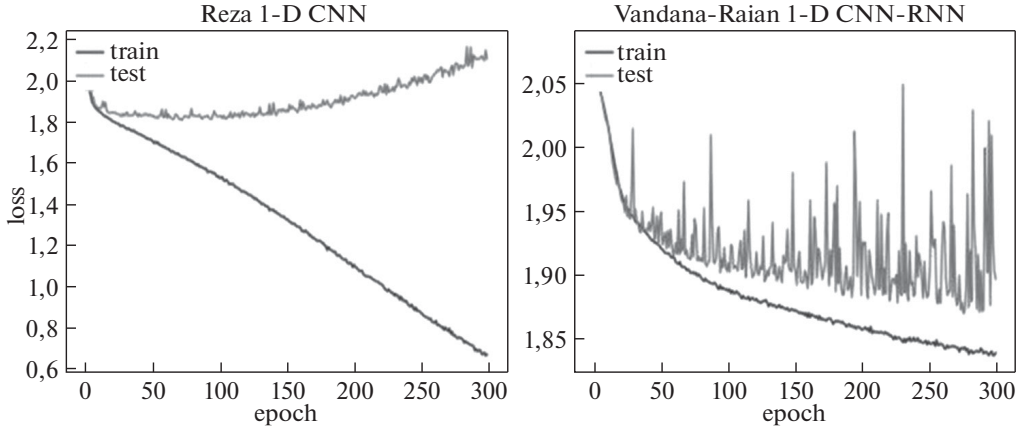

Рис. 7. Графики ошибки в процессе обучения моделей Reza-1-D-CNN и Vandana-Rajan-1-D-CNN-RNN с 1-D MFCC на основе баз данных эмоций CREMAD и IMPOCAP.

в 61%. Тем не менее для объединенных баз оценки классификации оставляют желать лучшего. Полученные результаты демонстрируют низкую эффективность классификации из-за усложнения структуры данных. Следует отметить, что в табл. 1 оценки для базы Emo-DB также невысоки.

В следующем эксперименте одномерные признаки MFCC были рассмотрены как временной ряд. Далее было применено преобразование Фурье к каждому из признаков. После преобразования были получены масштабированные признаки, представляющие собой массив из 64 элементов (рис. 8).

После повторного обучения точность данных проверки для сети Reza-1-D-CNN достигла 28%, а для сети Vandana-Raian-CNN-RNN — 27%. Графики

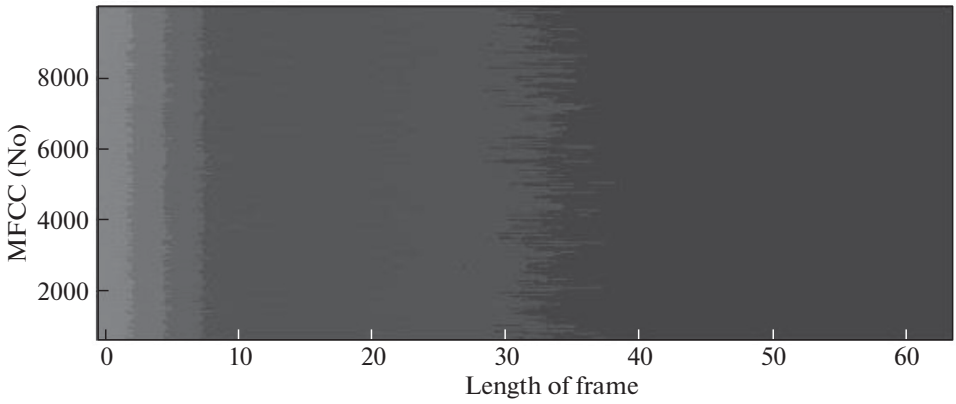

Рис. 8. 1-D-MFCC-FT признаки на основе баз данных эмоций CREMAD и IMPOCAP.

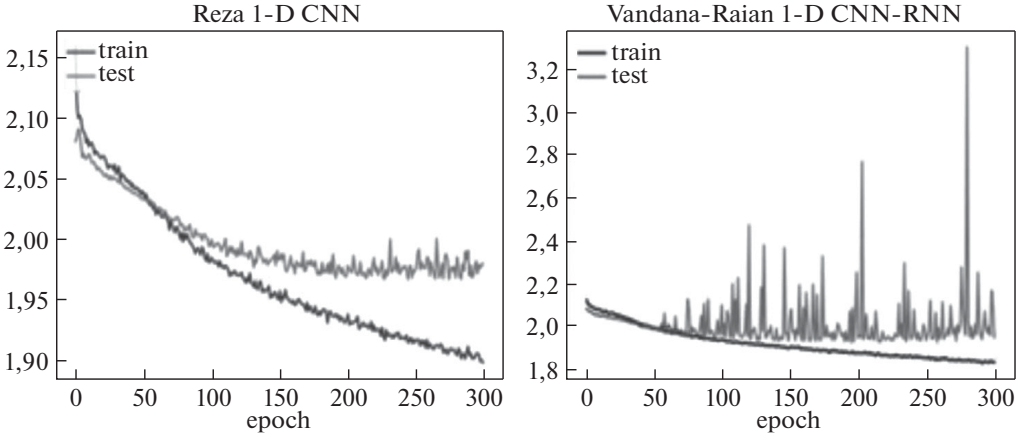

Рис. 9. Графики ошибки в процессе обучения моделей Reza 1-D CNN и Vandana-Rajan-1-D-CNN-RNN с 1-D MFCC-FT признаками, на основе баз данных эмоций CREMAD и IMPOCAP.

ошибок для сетей Reza-1-D-CNN и Vandana-Rajan-1-D-CNN-RNN с новыми признаками (1-D-MFCC-FT) показаны на рис. 9.

Из графиков видно, что оценки классификации согласуются с оценками предыдущего эксперимента. По сравнению со вторым экспериментом полученный способ позволяет снизить размерность обучающего признака в 32 раза.

Несмотря на то что расчет спектрограмм не полностью соответствует концепции сквозной сети, поскольку он допускает дополнительный этап предварительной обработки (преобразование 1D-MFCC в спектрограмму) перед моделью CHC, обработка минимальна, и наиболее важно, что сохраняется целостность сигнала. Предлагаемый подход к выделению признаков позволяет

значительно сократить длину обучающих признаков, обеспечивая простую трансформацию данных в новое пространство признаков. С практической точки зрения данный подход можно использовать для улучшения характеристик пространственного хранения или для вычислительной продуктивности алгоритмов обучения. Данный способ снижения размерности предлагается использовать в задачах РРЭ.

6. Заключение

Предложен подход приведения речевых данных, содержащих эмоциональную составляющую в речи, в форму, подходящую для алгоритмов машинного обучения. Очевидно, что качество и объем акустических признаков определяют, насколько хорошо алгоритмы машинного обучения способны обучаться. Следовательно, критически важно провести исследование и предварительную обработку признаков, прежде чем передавать их значения алгоритму обучения. Результаты эксперимента показывают, что небольшие сети, или сети, имеющие относительно малое число параметров, обладают недостаточной емкостью, а потому присутствует эффект недообученности, демонстрируется низкая эффективность, поскольку они не могут выявлять внутреннюю структуру сложных данных.

Предложенный авторами статьи подход для предобработки данных и выделения признаков способствует улучшению характеристик пространственного хранения и вычислительной продуктивности алгоритмов обучения. Полученные результаты важны для исследований, связанных с обработкой и анализом речевых сигналов, выделением определенных эмоциональных свойств говорящих [32]. Применение предложенного в статье метода в задачах анализа электронной информации позволит повысить эффективность работы за счет снижения вычислительной нагрузки, уменьшения пространства признаков и, соответственно, повышения скорости расчетов.

Справочная информация

1) Сокращения для конфигураций нейронных сетей:
Layer — слой;
LT — layer type (тип слоя);
SF — same filters (фильтры одного рода);
KS — kernel size (размер ядра свертки);
Strides — шаг свертки;
Activation — функция активации;
BN — batch Normalization (нормализация);
Dropout — регуляризация;
MP (P) — Max pooling (слой понижения размерности);
LSTM — Long short-term memory (слой с рекуррентной нейронной сетью);
AA — attention activation (слой активации рекуррентного слоя);

Flatten — полносвязный слой;

Dense — выходной полносвязный слой.

2) Конфигурации нейронных сетей (табл. 2, 3)

Таблица 2. Конфигурация одномерной нейронной сети — Reza-1-D-CNN

Layer	LT	SF	KS	Strides	Padding	BN	Activation	Dropout
1	CNN (SF)	256	8	1	same		ReLu	
2	CNN (SF)	256	8	1	same	+	ReLu	0,25
3	MP (P)		8	1				
4	CNN (SF)	128	8	1	same		ReLu	
5	CNN (SF)	128	8	1	same		ReLu	
6	CNN (SF)	128	8	1	same		ReLu	
7	CNN (SF)	128	8	1	same	+	ReLu	0,25
8	MP (P)		8	1				
9	CNN (SF)	64	8	1	same		ReLu	
10	CNN (SF)	64	8	1	same		ReLu	
11	flatten							
12	Dense						Softmax	

Таблица 3. Конфигурация одномерной нейронной сети — Vandana-Raian 1-D CNN-RNN

Layer	LT	SF	KS	Strides	Padding	BN	Activation	Dropout
1	CNN (SF)	64	3	1	same	+	elu	
2	MP (P)		4	4				
3	CNN (SF)	64	3	1	same	+	elu	
4	MP (P)		4	4				
5	CNN (SF)	128	3	1	same	+	elu	
6	MP (P)		4	4				
7	CNN (SF)	128	3	1	same	+	elu	
8	MP (P)		4	4				
9	LSTM	64						
10	AA						tanh	
11	LSTM	64						
12	Dense						Softmax	

СПИСОК ЛИТЕРАТУРЫ

1. *Мещеряков Р.В., Бондаренко В.П.* Диалог как основа построения речевых систем // Кибернетика и системный анализ. 2008. № 2. С. 30–41.

2. *Papakotas M., Siantikos G., Giannakopoulos T. et al.* IoT Applications with 5G Connectivity in Medical Tourism Sector Management: Third-Party Service Scenarios // GeNeDis 2016. Advances in Experimental Medicine and Biology. 2016. V. 989. P. 155–164. 2016. https://doi.org/10.1007/978-3-319-57348-9_12

3. *Okhapkin V., Okhapkina E., Iskhakova A. et al.* Application of neural network modeling in the task of destructive content detecting // CEUR workshop proceedings. Proceedings of the 3rd International Conference on R. Piotrowski's Readings in Language Engineering and Applied Linguistics, PRLEAL 2019. St. Petersburg, Russia, 2020. P. 85–94.

4. *Iskhakova A., Iskhakov A., Meshcheryakov R.* Research of the estimated emotional components for the content analysis // Journal of Physics: Conference Series. 2019. V. 1203. P. 1–10. https://doi.org/10.1088/1742-6596/1203/1/012065

5. *Scheirer E., Slaney M.* Construction and evaluation of a robust multifeature speech/music discriminator // IEEE International Conference on Acoustics, Speech, and Signal Processing. Munich, Germany, 2002. P. 1331–1334. https://doi.org/10.1109/ICASSP.1997.596192

6. *Hossan M.A., Memon S., Gregory M.A.* A novel approach for MFCC feature extraction // 2010 4th International Conference on Signal Processing and Communication Systems. Gold Coast, QLD, Australia, 2010. P. 1–5. https://doi.org/10.1109/ICSPCS.2010.5709752

7. *Logan B.* Mel Frequency Cepstral Coefficients for Music Modeling. https: //ismir2000.ismir.net/papers/logan_abs.pdf

8. *Rabiner L.R., Juang B.H.* Fundamental of Speech Recognition. USA: Prentice Hall, 1993.

9. *Nwe T.L., Foo S.W., Silva L.C.* Speech emotion recognition using hidden Markov models // Speech Communication. 2003. V. 41. No. 4. P. 603–623. https://doi.org/10.1016/S0167-6393(03)00099-2

10. *Zou D., Niu Y., He Z., Tan H.* A breakthrough in speech emotion recognition using deep retinal convolution neural networks. https: //arxiv.org/abs/1707.09917

11. *Lim W., Jang D., Lee T.* Speech Emotion Recognition using Convolutional and Recurrent Neural Networks // 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). Jeju, Korea (South), 2016. P. 1–4. https://doi.org/10.1109/APSIPA.2016.7820699

12. *Prasomphan S.* Improvement of speech emotion recognition with neural network classifier by using speech spectrogram // 2015 International Conference on Systems, Signals and Image Processing (IWSSIP). London, UK, 2015. P. 73–76. https://doi.org/10.1109/IWSSIP.2015.7314180

13. *Pakoci E., Popovic B., Pekar D.* Improvements in Serbian Speech Recognition using Sequence-Trained Deep Neural Networks // SPIIRAS Proceedings. 2018. Vol. 3(58). P. 53-76. https://doi.org/10.15622/sp.58.3

14. *Bengio Y., Hinton G.* Deep learning // Nature. 2015. V. 521. P. 436–444. https://doi.org/10.1038/nature14539.

15. *Valenti M., Squartini S., Diment A. et al.* A convolutional neural network approach for acoustic scene classification // 2017 International Joint Conference on Neural Networks (IJCNN). Anchorage, AK, 2017. P. 1547–1554. https://doi.org/10.1109/IJCNN.2017.7966035

16. *Hajarolasvadi N., Demirel H.* 3D CNN-Based Speech Emotion Recognition Using K-Means Clustering and Spectrograms // Entropy. 2019. V. 21(5) 479. P. 1–17. https://doi.org/10.3390/e21050479

17. *Niu Y., Zou D., Niu Y., He Z., Tan H.* A breakthrough in speech emotion recognition using deep retinal convolution neural networks. Preprint. https://arxiv.org/abs/1707.09917

18. *Burkhardt F., Paeschke A., Rolfes M., Sendlmeier W.F., Weiss B.* A Database of German Emotional Speech // INTERSPEECH 2005 — Eurospeech, 9th European Conference on Speech Communication and Technology. Lisabon, Portugal, 2005. P. 1–4. https://doi.org/10.21437/Interspeech.2005-446

19. *Haq S., Jackson P.J.B., Edge J.D.* Audio-Visual Feature Selection and Reduction for Emotion // Proceedings of the International Conference on Auditory-Visual Speech Processing 2008, Tangalooma Wild Dolphin Resort, Moreton Island, Queensland, Australia, 2008. P. 185–190.

20. *Haq S., Jackson P.J.B.* Speaker-Dependent Audio-Visual Emotion Recognition // Proceedings of the International Conference on Auditory-Visual Speech Processing, Norwich, UK, 2009. P. 53–58.

21. *Huang Z., Dong M., Mao Q., Zhan Y.* Speech Emotion Recognition Using CNN // MM '14: Proceedings of the 22nd ACM international conference on Multimedia. Orlando, Florida, USA, 2014. P. 801–804. https://doi.org/10.1145/2647868.2654984

22. *Prasomphan S.* Improvement of speech emotion recognition with neural network classifier by using speech spectrogram // 2015 IEEE International Conference on Systems, Signals and Image Processing. London, UK, 2015. P. 73–76. https://doi.org/10.1109/IWSSIP.2015.7314180

23. *Semwal N., Kumar A., Narayanan S.* Automatic speech emotion detection system using multi-domain acoustic feature selection and classification models // 2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA). New Delhi, India, 2017. P. 1–6.

24. *Chu R.* Speech Emotion Recognition with Convolutional Neural Network. 2019. https://towardsdatascience.com/speech-emotion-recognition-with-convolution-neural-network-1e6bb7130ce3

25. *Jianfeng Z., Mao X., Chen L.* Speech emotion recognition using deep 1D & 2D CNN LSTM networks // Biomedical Signal Processing and Control. 2019. V. 47. P. 312–323. https://doi.org/10.1016/j.bspc.2018.08.035

26. *Rajan V.* 1D Speech Emotion Recognition. 2021. https: //github.com/vandana-rajan/1D-Speech-Emotion-Recognition

27. *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). P. 1–35. https://doi.org/10.1371/journal.pone.0196391

28. *Dupuis K., Pichora-Fuller M.K.* Toronto emotional speech set (TESS). https://doi.org/10.5683/SP2/E8H2MF https://dataverse.scholarsportal.info/dataset.xhtml?persistentId= doi:10.5683/SP2/E8H2MF

29. *Cao H., Cooper D.G., Keutmann M.K. et al.* CREMA-D: Crowd-sourced emotional multimodal actors dataset // IEEE transactions on affective computing. 2014. V. 5(4). P. 377–390. https://doi.org/10.1109/TAFFC.2014.2336244

30. *Franti E., Ispas I., Dragomir V. et al.* Voice Based Emotion Recognition with Convolutional Neural Networks for Companion Robots // Romanian Journal of Information Science and Technology. 2018. V. 20(3). P. 222–240.

31. *Iskhakova A., Wolf D., Meshcheryakov R.* Automated Destructive Behavior State Detection on the 1D CNN-Based Voice Analysis // Speech and Computer. SPECOM 2020. Lecture Notes in Computer Science. 2020. V. 12335. P. 184–193. https://doi.org/10.1007/978-3-030-60276-5_19

32. *Исхакова А.О., Вольф Д.А., Исхаков А.Ю.* Неинвазивный нейрокомпьютерный интерфейс для управления роботом // Высокопроизводительные вычислительные системы и технологии. 2021. Том 5. № 1. С. 166–171.

Статья представлена к публикации членом редколлегии О.П. Кузнецовым.

Поступила в редакцию 17.11.2021

После доработки 19.01.2022

Принята к публикации 26.01.2022