МРНТИ 20.19.29

10.51889/2959-5894.2024.86.2.021

Н.О. Мекебаев[1*], Д.К. Даркенбаев[2], А. Алтыбай[2]

[1]Казахский национальный женский педагогический университет, г. Алматы, Казахстан
[2]Казахский национальный университет им. аль-Фараби, г. Алматы, Казахстан
[*]*e-mail: nurbapa@gmail.com*

НЕЙРОННЫЕ АРХИТЕКТУРЫ ДЛЯ ОПРЕДЕЛЕНИЯ ПОЛА И ИДЕНТИФИКАЦИИ ГОВОРЯЩЕГО

Аннотация

В этой статье мы исследуем две нейронные архитектуры для задач определения пола и идентификации говорящего, используя функции мелкочастотных кепстральных коэффициентов (MFCC), которые не охватывают характеристики, связанные с голосом. Одна из наших целей – сравнить различные нейронные архитектуры, многослойный персептрон (MLP) и сверточные нейронные сети (CNNS) для обеих задач с различными настройками и автоматически изучить особенности, характерные для пола/ говорящего. Экспериментальные результаты показывают, что модели, использующие z-оценку и преобразование матрицы Грамиана, дают лучшие результаты, чем модели, использующие только максимальную-минимальную нормализацию MFCC. С точки зрения времени обучения, MLP требует больших периодов обучения для сходимости, чем CNN. Другие экспериментальные результаты показывают, что MLP превосходят CNN в решении обеих задач с точки зрения ошибок обобщения.

Ключевые слова: MLP, CNN, ASR; NN, определение пола; идентификация говорящего.

Н.О. Мекебаев[1], Д.К. Даркенбаев[2], А. Алтыбай[2]
[1]Қазақ Ұлттық қыздар педагогикалық университеті, Алматы қ., Қазақстан
[2]әл-Фараби атындағы Қазақ ұлттық университеті, Алматы қ., Қазақстан

СӨЙЛЕУШІНІ ЖӘНЕ ГЕНДЕРЛІК АНЫҚТАУҒА АРНАЛҒАН НЕЙРОНДЫҚ АРХИТЕКТУРА

Аңдатпа

Бұл мақалада біз дауыспен байланысты сипаттамаларды қамтымайтын кіші жиілікті кепстральды коэффициенттердің (MFCC) функцияларын қолдана отырып, сөйлеушінің жынысын анықтау және анықтау тапсырмаларына арналған екі нейрондық архитектураны қарастырамыз. Біздің мақсаттарымыздың бірі-әртүрлі нейрондық архитектураларды, көп қабатты перцептронды (MLP) және конволюциялық нейрондық желілерді (CNNS) екі тапсырма үшін де әртүрлі параметрлермен салыстыру және жынысқа/ сөйлеушіге тән ерекшеліктерді автоматты түрде зерттеу. Эксперименттік нәтижелер Z-баллды және Грамиан матрицасын түрлендіруді қолданатын модельдер тек mfcc максималды-минималды қалыпқа келтіруді қолданатын модельдерге қарағанда жақсы нәтиже беретіндігін көрсетеді. Оқу уақыты тұрғысынан MLP CNN-ге қарағанда конвергенция үшін үлкен оқу кезеңдерін қажет етеді. Басқа эксперименттік нәтижелер MLP жалпылау қателері тұрғысынан екі мәселені шешуде CNN-ден жоғары екенін көрсетеді.

Түйін сөздер: MLP, CNN, AS; NN, гендерлік анықтау; сөйлеушіні анықтау.

N. Mekebayev[1], D. Darkenbayev[2], A. Altybay[2]
[1]Kazakh National Women's Teacher Training University, Almaty, Kazakhstan
[2]al-Farabi Kazakh National University, Almaty, Kazakhstan

NEURAL ARCHITECTURES FOR GENDER DETERMINATION AND SPEAKER IDENTIFICATION

Abstract

In this article, we explore two neural architectures for gender determination and speaker identification tasks using functions of small-frequency cepstral coefficients (MFCC), which do not cover voice-related

characteristics. One of our goals is to compare different neural architectures, multilayer perceptron (MLP) and convolutional neural networks (CNNS) for both tasks with different settings and automatically study gender/speaker–specific features. Experimental results show that models using z-score and Gramian matrix transformation give better results than models using only maximum-minimum MFCC normalization. In terms of training time, MLP requires longer training periods for convergence than CNN. Other experimental results show that MLPs are superior to CNNs in solving both problems in terms of generalization errors.

Keywords: MLP, CNN, ASR; NN, gender determination; speaker identification.

Основные положения

Основная идея исследования заключается в создании модели и алгоритма определения гендерной идентичности и говорящего на основе нейронных сетей. В статье построены алгоритмы и модель определения мужского, женского голоса и говорящего с использованием нейронной сети. были созданы алгоритмы и модели обнаружения и обработки сигналов с использованием машинного обучения в задачах распознавания речи. В предварительной обработке речевого сигнала с использованием MFCC MFCC определена гендерная специфика. Был проведен сравнительный анализ архитектур нейронных сетей MLP и CNN для распознавания гендерной идентичности и звуковых характеристик говорящего, и было обнаружено, что CNN показал хорошие результаты.

Введение

Автоматическое определение пола и идентификация говорящих по голосу является важной задачей в области обработки аудиосигнала. Определение пола связано с определением того, принадлежит ли речь мужчине или женщине. Эта задача очень важна для автоматического распознавания речи в зависимости от пола (ASR), которое позволяет системе ASR быть более точной, чем системы, не зависящие от пола. Распознавание говорящего-это процесс автоматического распознавания говорящих на основе индивидуальной информации, содержащейся в речевой волне, которую можно разделить на идентификацию говорящего и верификацию говорящего. Верификация говорящего-это процесс принятия или отклонения заявления об идентификации говорящего. Идентификация говорящего, с другой стороны, это процесс определения того, какие зарегистрированные говорящие передают вводимую речь.

В целом, определение пола и идентификация говорящего могут рассматриваться как задачи классификации, в которых первая классифицирует входящий звук на две категории, а вторая классифицирует входной звук по количеству зарегистрированных говорящих. Для гендерной классификации было предложено множество подходов, наиболее часто используемыми методами являются дерево решений [1], машина опорных векторов (SVM) [2], байесовская сеть, K-ближайший сосед и сслучайный лес [3-5]. Как известно, для обработки аудиосигнала может быть использовано множество функций, а именно MFCC, промежуточные векторы (i-векторы) [6], энергетическая энтропия, кратковременная энергия и спектральный центроид и т.д. Модели гауссовой смеси (GMMS) с MFCC [7] - традиционный способ решения задачи идентификации говорящего. Эта структура была расширена для использования i-vector и совместного факторного анализа [8] для формирования компактного представления высказывания, имеющего специфические характеристики, связанные с голосом. В последнее десятилетие глубокое обучение начинает доминировать в различных областях искусственного интеллекта, таких как машинное обучение, обработка естественного языка [9], компьютерное зрение и обработка аудиосигнала и т.д.

В этой статье мы используем различные нейронные архитектуры как для определения пола, так и для идентификации говорящего. Мы применяем многослойный персептрон (MLP) [9] и сверточные нейронные сети (CNN), чтобы иметь возможность изучать гендерные особенности / особенности говорящего по оригинальным векторам MFCC, которые не охватывают специфические характеристики речевого сигнала, связанные с голосом. Другой целью этой работы является сравнение производительности двух нейронных архитектур для обеих задач

с различными настройками / способами: 1) различное преобразование признаков, 2) разный размер модели, 3) добавление шумового сигнала к тестовому набору для измерения ошибки обобщения модели. Мы оцениваем наши методы на корпусе казахской [10]. Результаты эксперимента показывают, что MLP превосходит CNN с точки зрения ошибки обобщения, и для сходимости MLP требуется больше периодов обучения, чем CNN для обеих задач.

Методология исследования

Определение пола по голосу направлено на автоматическое определение пола автора по звуковым сигналам. Аналогично, идентификация говорящего заключается в установлении личности автора (имени или ID) путем анализа его / ее аудиозаписей.

Пусть $X = x_1, x_2, \dots x_n$ обозначает серию аудиосигналов в качестве входных данных. $G = g_1, g_2, \dots g_n$ - это двоичный вектор, равный 0/1 для гендерных категорий, соответствующих аудиосигналам X. Здесь мы используем 1 для обозначения женщины и 0 для мужчины. $S = s_1, s_2, \dots s_n$обозначает идентификатор говорящего, мы используем уникальный номер для различения говорящих. Пары обучающих элементов для двух задач могут быть определены следующим образом:

1) (X,G)=(x_1, g_1),…,(x_n, g_n) − для определения пола;

2) (X,S)=(x_1, s_1),…,(x_n, s_n) для идентификации говорящего.

Для этих двух задач мы используем X в качестве входных данных и извлекаем соответствующие признаки сигнала, затем используем различные нейронные сети для обучения моделей для определения пола и идентификации говорящего.

Извлечение признаков сигнала

Как и во многих задачах обработки речи (распознавание речи и т.д.), первым шагом является извлечение функций, которые могут быть использованы для идентификации лингвистического контента, содержащегося в аудиосигналах, и для отбрасывания информации о фоновом шуме. Частотные кепстральные коэффициенты Mel (MFCC) - это самые современные функции, широко используемые во многих приложениях для обработки речи. Прежде чем описывать MFCC, давайте покажем оригинальный аудиосигнал, показанный на рисунке 1. Исходный сигнал состоит из тысяч или миллионов чисел, его можно рассматривать как очень длинный вектор, который содержит лингвистический контент и шум. Оригинальный аудиосигнал, показанный на рисунке 1.

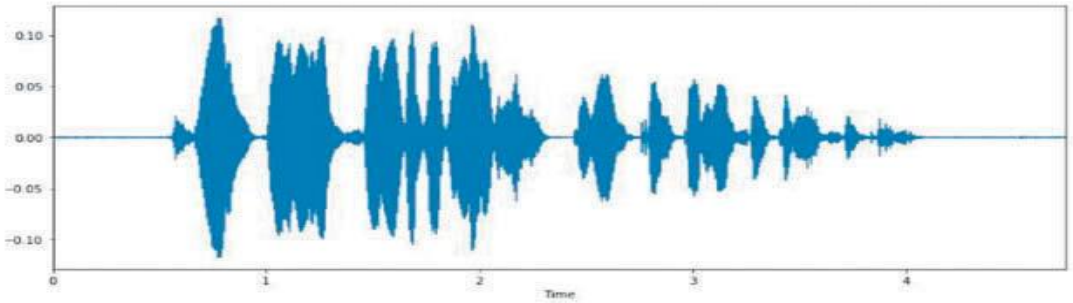

Рисунок 1. Оригинальный аудиосигнал

В этой работе мы используем MFCC для определения пола/говорящего, и способ извлечения функции MFCC не является предметом данной статьи. На практике мы применяем LibROSA, пакет python для анализа аудиосигнала. Его функция librosa.feature.mfcc была использована для извлечения MFCC. Извлеченные объекты показаны на рисунке 2.

На практике мы устанавливаем количество функций MFCC равным 40, тогда размерность MFCC для аудио равна $M \in R^{|40 \times n|}$ максимальная-минимальная нормализация вычисляется для каждой функции MFCC, и в дальнейшем она относится к оригиналу MFCC.

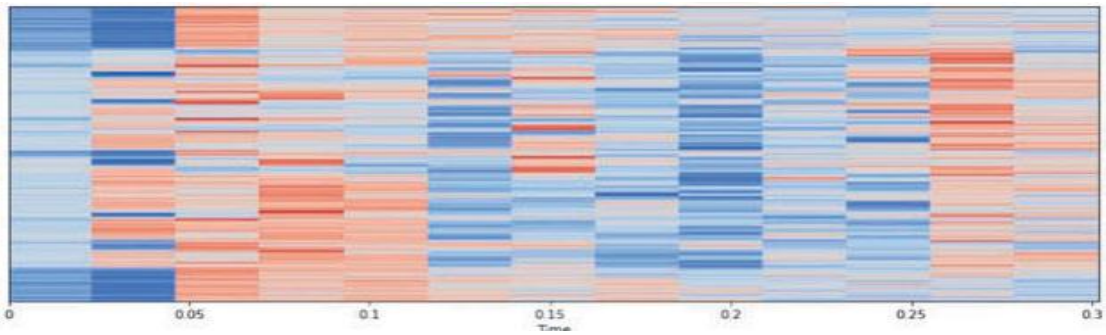

Рисунок 2. Функция MFCC аудиосигнала

Мы попробовали альтернативную нормализацию, z-баллы, для функций MFCC с помощью следующего расчета:

$$M^z = \frac{M-\mu}{std(M)} \quad (1)$$

где μ - среднее значение, а $std(M)$ - стандартное отклонение.

Один из стандартных способов обработки переменной длины входных данных - найти максимальную длину аудиосигнала и заполнить его функции MFCC нулевым значением, если длина меньше максимальной длины. Один из эффективных способов решить проблему переменной длины путем следующего преобразования:

$$M^g = M^z \times M^{z^T} \quad (2)$$

Где $M^g \in R^{|40\times40|}$, 40 - количество функций MFCC. Тогда мы могли бы применить операцию atten к M^gили использовать ее двумерную форму (2-D).

Модели

Нейронные сети (NN) можно рассматривать как функцию-классификатор с параметрами, а нейронную сеть с несколькими слоями можно рассматривать как композицию функций, определенных следующим образом:

$$f_\theta(\cdot) = f_\theta^l(f_\theta^{(l-1)}(\dots f_\theta^1))$$

где θ обозначает параметры NN, а l - количество слоев. Далее мы опишем две архитектуры наших нейронных сетей для задач определения пола и говорящего: 1) прямой NN, это многослойный персептрон и относится к MLP; 2) сверточная NN, он относится к CNN;

Нейронные сети с прямой связью

Чтобы лучше описать модель, давайте начнем с простой нейронной сети. Как известно, однослойный персептрон [10] представляет собой NN без скрытых блоков, который содержит только входной слой и выходной слой. Нелинейного выделения признаков нет, что означает, что выходные данные вычисляются непосредственно из суммы произведения весов, соответствующих входным данным. Мы используем MLP, и это NNS, состоящая из множества восприятий, и MLP может изучать или извлекать нелинейные признаки. Вообще говоря, MLP состоит из входного слоя, некоторого количества скрытых слоев и выходного слоя.

Сверточные нейронные сети

Сверточные нейронные сети (CNN) представляют собой специализированный вид нейронной сети для обработки данных с 2-мерной сеточной топологией. CNN добилась огромного успеха в практических приложениях. В отличие от MLP, который использует

полностью подключенные слои для извлечения объектов, CNN использует две важные идеи, которые могут помочь улучшить модель: разреженные взаимодействия и совместное использование параметров. Первый представляет собой процесс извлечения объектов с ядром меньшего размера, чем входные данные. Например, при обработке аудио входные сигналы могут содержать тысячи или миллионы чисел, вместо того, чтобы передавать такой длинный вектор в NN, CNN может обнаруживать небольшие и значимые объекты, собирая локальную информацию. Совместное использование параметров относится к использованию одного и того же параметра для меньшего ядра, перемещающегося по двумерному входу. Типичный CNN состоит из трех этапов: - используйте слои свертки для выполнения набора линейных активаций; - каждая линейная активация выполняется с помощью функции нелинейной активации; - используйте функцию объединения для дальнейшего изменения выходных данных слоя.

Эксперименты

Мы провели серию экспериментов для оценки моделей MLP и CNN для задач определения пола и говорящего:

• первый эксперимент предназначен для анализа эффективности извлеченных функций из MFCC, которые тестируются для обеих моделей. Как описано в разделе 3, мы используем два типа объектов и сравниваем их: i) нормализованный-сглаженный исходный MFCC в виде длинного вектора признаков и ii) используя z-оценку для MFCC, затем преобразуем ее в матрицу Грама (уравнение 2).

• чтобы эффективно сравнить две модели, мы тестируем обученные модели двух типов 500 раз, добавляя различный шум (нормальное распределение с нулевым средним значением и одной дисперсией) к тестовому набору.

• визуализация аудиозаписей после обучения модели для обеих задач.

Для оценки модели мы сообщаем о результатах точности, отзыва, оценки F1 при различных настройках модели.

Наборы данных

Мы используем набор данных из исследования [9]. В таблицах 1 и 2 представлена статистика наборов данных по признаку пола и говорящего. Можно видеть, что в обучающем и тестовом наборах всего 855 и 570 аудиозаписей для задачи определения пола. Количество аудиозаписей для мужчин и женщин в обучающем наборе равно 448 и 407. Остальные 570 аудиозаписей для мужчин и женщин в качестве тестового набора.

Таблица 1. Наборы данных для определения пола

Наборы данных	*Мужчина*	*Женщина*	*Всего*
Обучение (training)	*488*	*407*	*855*
Тестирование (testing)	*308*	*268*	*570*

В таблице 2 показаны обучающие и тестовые наборы для идентификации говорящих. Видно, что всего имеется 19 динамиков, и у каждого из них есть около 60 аудиозаписей для обучения. Это довольно маленький обучающий набор, и он не способен хорошо обучать какие-либо модели глубокого обучения, требующие больших данных. Но в этой работе, для идентификации говорящего, мы используем этот набор данных для обучения моделей MLP и CNN и проведения оценок.

Архитектуры малых и больших моделей обобщены в таблицах 3 и 4. Основным гиперпараметром является количество скрытых блоков, которое мы устанавливаем равным 128, 64, 32 и 512, 256, 128 для малых и больших моделей соответственно. Мы используем Relu в качестве функции активации для всех слоев, а значение отсева установлено равным 0.15.

Таблица 2. Наборы данных для идентификации говорящего. Ниже указаны номера аудиозаписей каждого говорящего, соответствующие каждому идентификатору говорящего

Наборы данных	*Идентификатор говорящего* 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19	*Всего*
Обучение (training)	63 63 56 57 58 60 63 66 59 59 63 61 65 61 60 57 54 59 56	*1140*
Тестирование (testing)	12 12 19 18 17 15 12 9 16 16 12 14 10 14 18 21 16 19	*285*

Таблица 3. Гиперпараметры MLP, используемые для обеих задач

MLP	**Small**			**Large**		
Layers	**Layer-1**	**Layer-2**	**Layer-3**	**Layer-1**	**Layer-2**	**Layer-3**
Hidden units	128	64	32	512	256	128
Dropout	0.15	0.15	0.15	0.15	0.15	0.15
Activation function	Relu	Relu	Relu	Relu	Relu	Relu

Таблица 4. Гиперпараметры CNN, используемые для обеих задач

CNN	**Small**			**Large**		
Layers	**Conv. maxP**	**Conv. maxP**	**Dense**	**Conv. maxP**	**Conv. maxP**	**Dense**
Hidden units	128	64	32	512	256	128
Dropout	0.15	0.15	0.15	0.15	0.15	0.15
Activation function	Relu	Relu	Relu	Relu	Relu	Relu

Настройка модели

В экспериментах тестируются нейронные архитектуры MLP и CNN с малыми и большими моделями. Мы обучаем две версии наших моделей, чтобы оценить компромисс между производительностью и размером.

Результаты исследования

На рисунке 3 показано распределение аудиосигналов с функциями MFCC и функциями после нормализации z-балла и преобразования матрицы Грамиана. На рисунке 3 (a, d ,c и e) показано распределение аудио с исходным MFCC, после нормализации z-балла и преобразования матрицы Грамиана для набора данных гендера соответственно. Аналогично, на рисунке 3 (b, d и e, f) показано распределение звука для набора данных динамика. Можно видеть, что распределение оригинальных MFCC по полу и аудиозаписям диктора не разделено по полу или динамикам, или мы можем сказать, что они смешаны вместе и трудно различить пол или аудиозаписи диктора / После преобразования z-score распределение аудиозаписей становится более плотным, чем исходные. Что еще более неожиданно, аудиозаписи после преобразований z-score и матрицы Грамиана, распределение аудиозаписей по полу и говорящим, изменяются. Из рисунка 3 (e, f) мы можем наблюдать, что аудиосигнал для обозначения пола и говорящего более различим, а несколько аудиозаписей смешаны. Оказывается, что если мы обучим/протестируем модели на этом наборе данных с таким распределением, то легко достигнем 100% точности. На рисунке 3 показано распределение аудио по полу и говорящим.

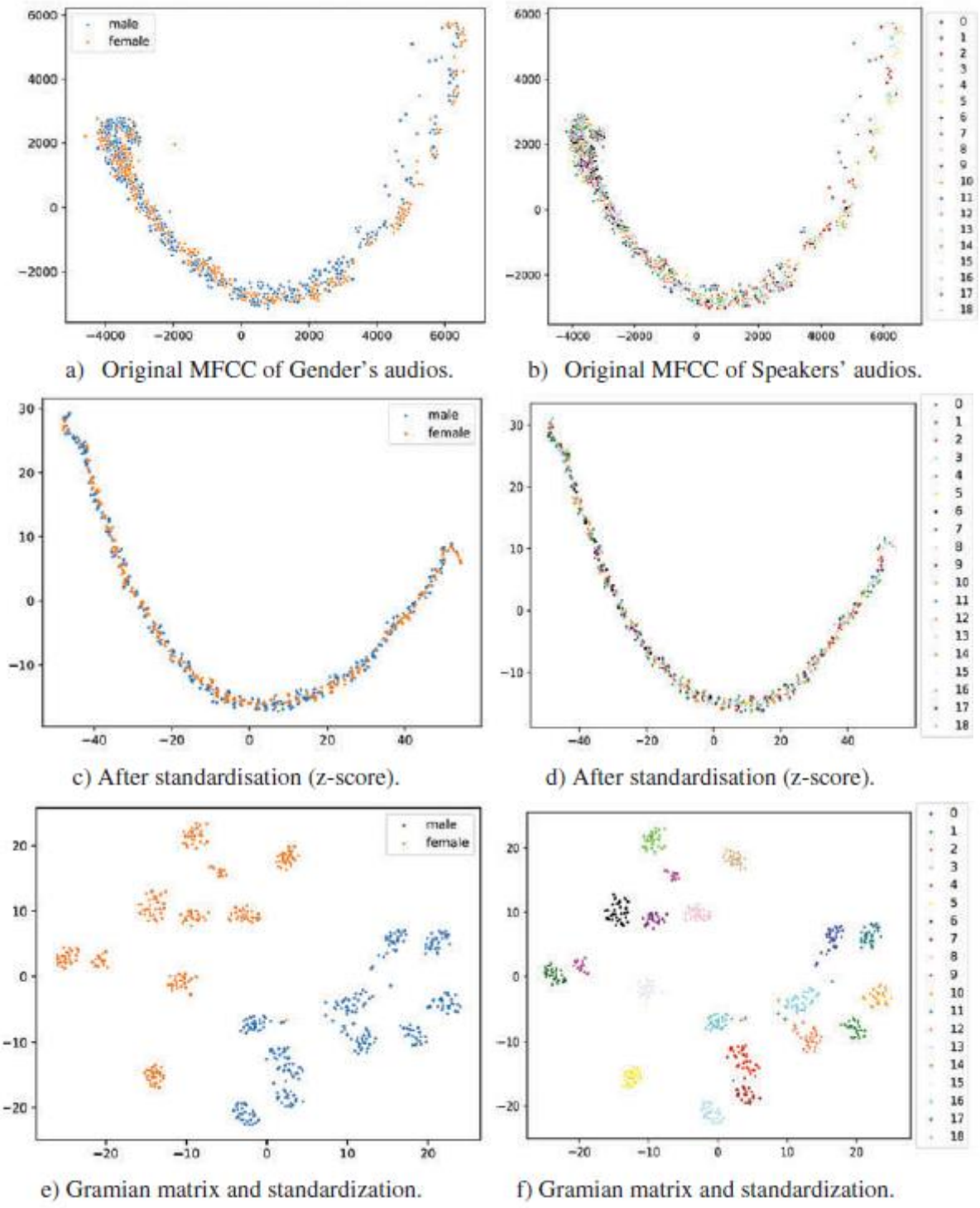

Рисунок 3. Распределение аудио по полу и говорящим

В таблице 5 приведены результаты распознавания пола и говорящего с различными функциями (L и G) и настройками модели (маленькая и большая). Можно заметить, что модель, обученная/протестированная после преобразования Грамиана (обозначенная в таблице как G), дала 100% оценку F1, что подтверждает результаты, упомянутые выше. Сравнивая результаты моделей, использующих длинный сглаженный вектор MFCC с нормализацией max-min (обозначенный как L в таблице 5), мы можем видеть, что полностью подключенный MLP превосходит CNN по определению пола.

Таблица 5. Результаты распознавания пола с различными признаками: L обозначает использование нормализованного откормленного длинного вектора mfcc; G обозначает использование z-оценки и преобразования матрицы Gramian. P - точность, R - отзыв, а F1 – оценка

Models		Male			Female			Macro, avg		
		P	R	F1	P	R	F1	P	R	F1
L	MLP-small	79.56	85.09	82.24	81.78	75.37	78.44	80.67	80.23	80.34
	MLP-large	77.22	92.05	83.98	88.57	69.40	77.82	82.89	80.72	80.90
G	MLP-small	100	100	100	100	100	100	100	100	100
	MLP-large	100	100	100	100	100	100	100	100	100
L	CNN-small	79.68	83.11	81.36	80	76.11	78.01	80.55	79.07	79.21
	CNN-large	76.57	88.74	82.20	84.54	69.40	76.22	79.84	79.61	79.68
G	CNN-small	100	100	100	100	100	100	100	100	100
	CNN-large	100	100	100	100	100	100	100	100	100

На рисунке 4 показана кривая обучения этих моделей для определения пола, и можно видеть, что MLP, как правило, требуется больше периодов обучения для сходимости, чем CNN (рис. 4 (a, b)). Здесь мы показываем только модели малого размера, и ситуация для большой аналогична. Из рисунка видно, что MLP и CNN с матрицей Грамиана (G) занимают относительно меньше времени обучения, чем L. На рисунок 4 показано кривая обучения MLP и CNN для определения пола.

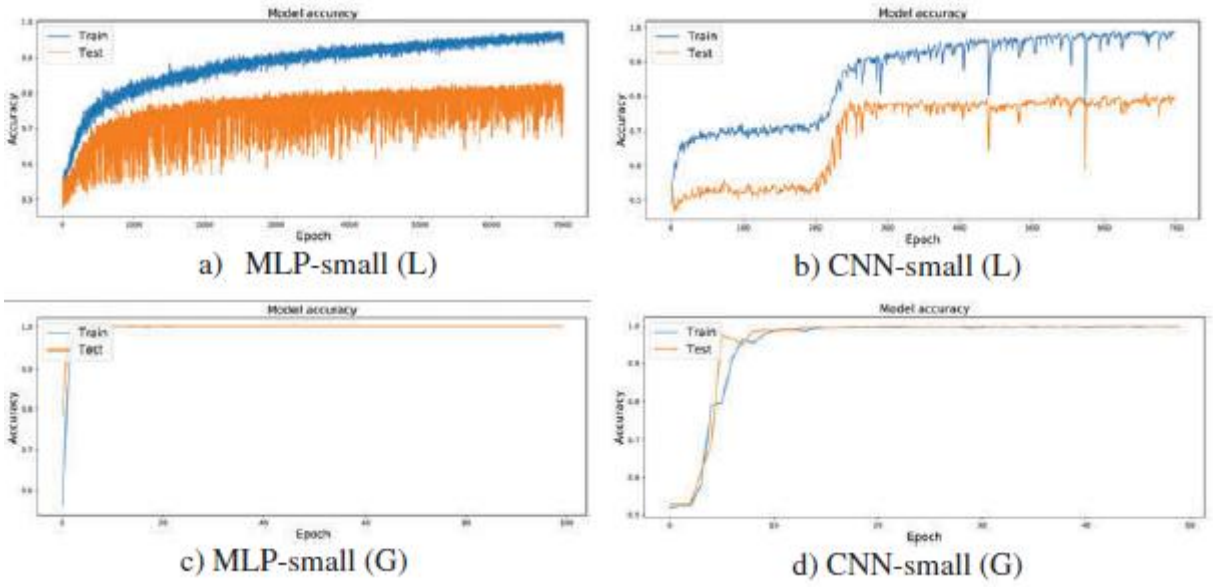

Рисунок 4. Кривая обучения MLP и CNN для определения пола

Давайте перейдем к результатам идентификации динамиков, в которых мы также использовали различные настройки функций и размер модели. В таблице 6 приведены результаты, и мы видим, что модель, использующая функцию L, получила относительно худшие результаты, независимо от того, какие модели мы используем. F-оценка моделей, использующих L, в диапазоне от 19% до 36%. Преобразование матрицы Грамиана, по-видимому, заметно повысило производительность модели. F-оценка обеих моделей почти достигает 99%, что выше, чем у модели, обученной с помощью L.

Как мы можем видеть, для сходимости модели с L требуются большие периоды обучения; напротив, модель с G требует меньше шагов обучения.

Таблица 6. Результаты идентификации говорящего

Models		Macro, avg.		
		P	R	F1
L	MLP-small	20.76	20.36	19.94
	MLP-large	21.76	21.92	21.50
G	MLP-small	98.66	98.57	98.58
	MLP-large	99.36	99.28	99.30
L	CNN-small	36.78	35.68	35.25
	CNN-large	36.77	36.51	36.16
G	CNN-small	99.17	99.09	99.12
	CNN-large	99.17	99.47	99.47

Другая проблема, которую можно обнаружить на этих рисунках a, b, заключается в том, что точность обучения модели с L постепенно достигает 99%, результаты проверки остаются на уровне 20%. На рисунке 5 показана кривая обучения идентификации говорящего.

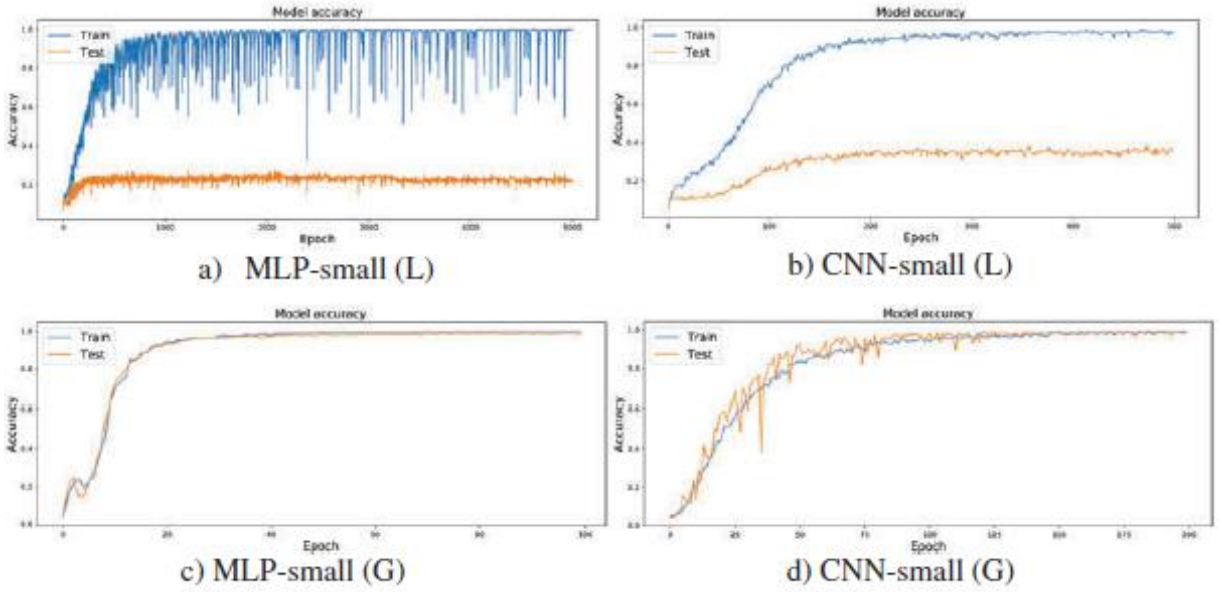

Рисунок 5. Кривая обучения MLP и CNN для идентификации говорящего

Способность к обобщению

Чтобы сравнить ошибку обобщения моделей, мы провели еще один эксперимент: добавили шумовой сигнал к тестовому набору, каждый из них был нормально распределен с нулевым средним значением и одной дисперсией, а уже обученные модели с G были протестированы 500 раз. На рисунках 6 и 7 показано, как распределяются F-баллы режимов.

Видно, что для распознавания пола и говорящего CNN не могут превзойти MLP при добавлении к тестовому набору различных нормально распределенных шумовых сигналов. Одна из возможных причин того, что каждый уровень MLP полностью подключен, и входные данные взаимодействуют друг с другом в большем количестве измерений.

В отличие от этого, CNN имеют слой свертки, который обычно рассматривается как средство извлечения объектов региона, которое использует слайдовое двумерное окно с заданным шагом и общим весом на входе для извлечения объектов региона, и входные данные взаимодействуют только в регионах.

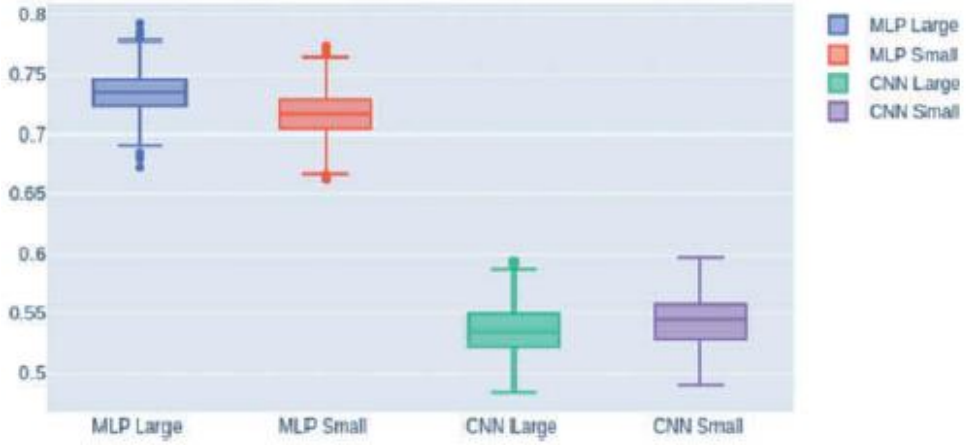

Рисунок 6. Результаты определения пола после добавления шумового сигнала к тестовому набору

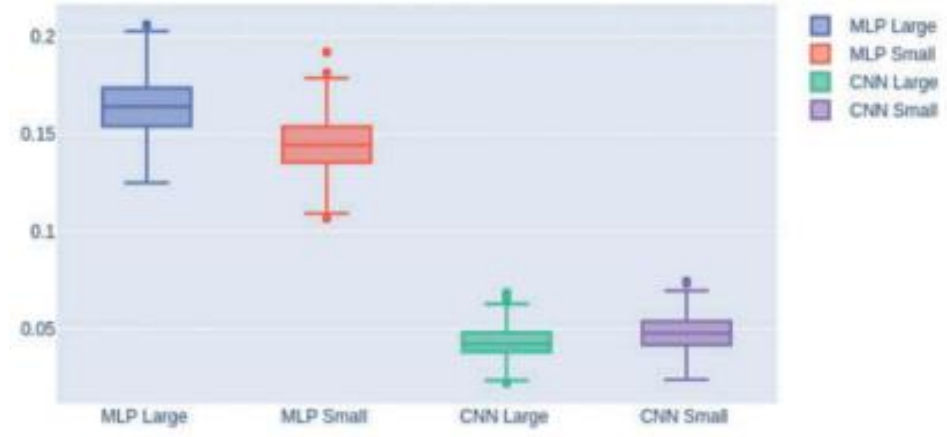

Рисунок 7. Результаты идентификации динамика после добавления шумового сигнала к тестовому набору

Интеграция между одной входной областью и другой не фиксируется, за исключением случаев, когда мы выбираем достаточно маленький размер шага. Другая возможная причина заключается в том, что MLP имеют больше обучаемых параметров, чем CNN, поскольку MLP имеют полностью связанные слои, а CNN имеют общие веса для слоев свертки. В результате можно видеть, что CNN дает большую ошибку обобщения, чем MLP, для идентификации пола и говорящего.

Визуализация

На рисунке 8 показана визуализация тестового набора после обучения модели для идентификации пола и говорящего обеих моделей с различными формами признаков (L и G). Мы используем обученные модели в тестовом наборе и получаем выходные данные промежуточного слой в качестве представления аудио, затем используйте для визуализации. На рис. 8 (a - c) и d показаны результаты визуализации для определения пола, видно, что MLP, по-видимому, лучше классифицирует аудиозаписи на два класса: мужские и женские. На рисунке 8 показана визуализация тестового набора после обучения модели для идентификации пола и говорящего.

Одним из критериев оценки кластеризованных результатов является отображение двух расстояний: внутриклассового и межклассового. Первое заключается в измерении расстояния между элементами в классе, и меньшее расстояние указывает на лучшие результаты. Последнее предназначено для измерения расстояния между разными классами, и большое расстояние указывает на лучшие результаты.

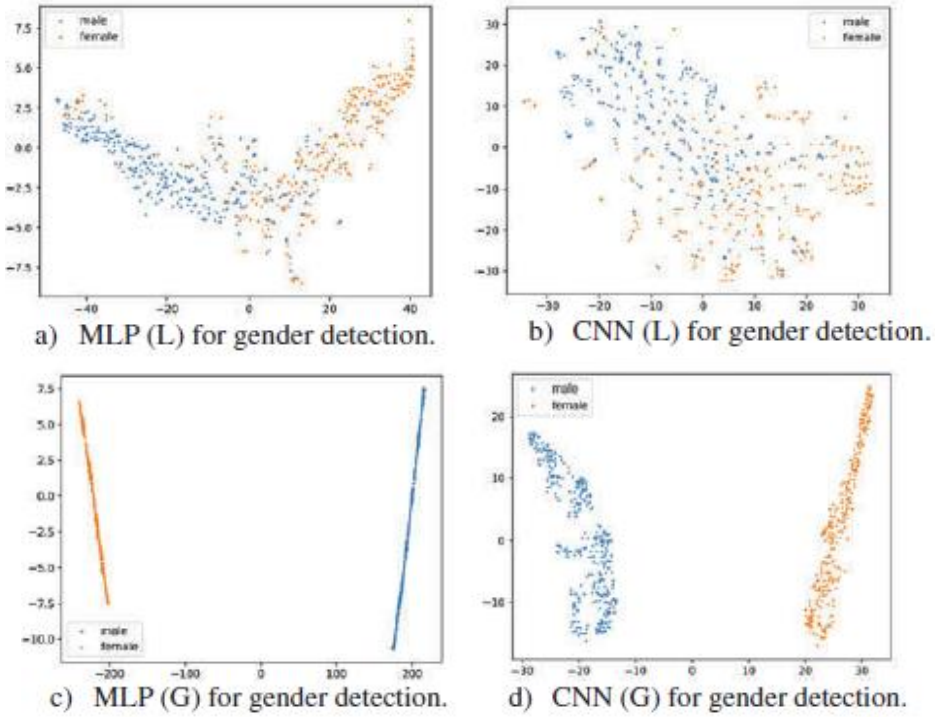

Рисунок 8 (a - c) и d результаты визуализации для определения пола

Сравнение рисунков 8(c, d) показывает, что MLP может классифицировать аудио в более плотный класс, чем CNN, видно, что расстояние внутри класса меньше, чем у CNN. На рисунках 8(e - g) и h показаны результаты визуализации для идентификации говорящего. На рисунке 8 (g, h) показаны результаты для модели с G, и, как мы видим, MLP для идентификации говорящего имеет большее межклассовое расстояние, чем CNN. На рисунке 8 (e - g) и h показано результаты визуализации для идентификации говорящего.

Дискуссия

Создана модель и алгоритм определения гендерной специфики и говорящего на основе нейронных сетей. В этой статье был проведен сравнительный анализ архитектур нейронных сетей в определении гендерной идентичности и говорящего. Там CNN показал лучшие результаты, чем MLP.

Эксперимент включает в себя две нейронные архитектуры: MLP и CNN, которые представлены различными модельными регуляциями. Для двух идентичных вычислений модель CNN опережает MLP, где относительная рациональность в определении гендерной идентичности варьируется в вариантах от 10% до примерно 20%. Относительная рациональность в определении говорящего колеблется от 2% до 6%. Этот результат отмечен несколькими аспектами.

MLP и CNN делают большую часть первого шага при сравнении тренировочного процесса, а второй-значительно меньше. Результаты визуализации показывают, что MLP помещает аудиозаписи одного и того же диктора в область зеркалирования, в то время как CNN находится в относительно широкой области, поэтому эти результаты показывают, что модель CNN лучше, чем MLP.

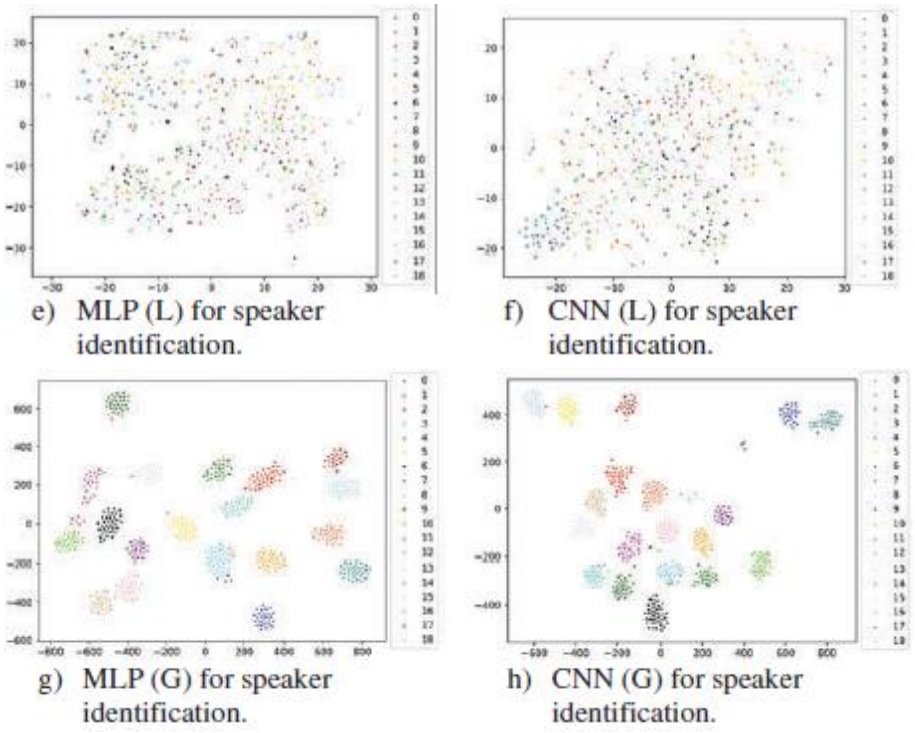

Рисунок 8 (e - g) и h результаты визуализации для идентификации говорящего

Заключение

В этой статье мы применили различные архитектуры нейронных сетей для определения пола и идентификации говорящего. Сравнения двух нейронных сетей были проведены разными способами: 1) различные преобразования признаков (L и G), 2) различные размеры модели (малый и большой) и 3) добавление шумового сигнала к тестовому набору для измерения ошибок обобщения модели (протестировано 500 раз). Результаты показывают, что для обеих задач два типа нейронных сетей получают относительно лучшие результаты после применения z-оценки и преобразования матрицы Грамиана. С точки зрения времени обучения, MLP требует большего количества периодов обучения для сходимости, когда используется только нормализация max-min для функций MFCC. Размер модели не оказывает существенного влияния на характеристики модели, а большие модели дают лишь незначительное улучшение. Другой результат сравнения показывает, что MLP превосходят CNN в этих экспериментах с точки зрения ошибки обобщения. Результаты визуализации показывают, что MLP могут классифицировать аудио в более плотные классы, чем CNN, для обеих задач.

Список использованных источников

[1] Auer, P., Burgsteiner, H., & Maass, W (2019). A learning rule for very simple universal approximators consisting of a single layer of perceptrons. Neural Networks: The Official Journal of the International Neural Network Society, 21(5), 786–795.

[2] Rabiner L (2019). A tutorial on hidden markov models and selected applications in speech recognition. – P. 257–286.

[3] Orken Mamyrbayev, Nurbapa Mekebayev, Mussa Turdalyuly, Nurzhamal Oshanova,Tolga Ihsan Medeni and Aigerim Yessentay (2019). Voice Identification Using Classification Algorithms//We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists. London– 1 – 14 p.

[4] Cunningham, P., & Delany, S. (2019). k-nearest neighbour classifiers. Multiple Classifier System, 1 – 17.

[5] Mermelstein P (2020). Distance measures for speech recognition, psychological and instrumental // Pattern recognition and artificial intelligence. – Vol. 116. –P. 374–388.

[6] Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2019). Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech and Language Processing. 19(4), 1–11

[7] Toleu, A., Tolegen, G., & Makazhanov, A. (2021). Character-aware neural morphological disambiguation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics 2, 666–671. Association for Computational Linguistics, Vancouver, Canada.

[8] Kalimoldayev, M., Mamyrbayev, O., Mekebayev, N., Kydyrbekova, A (2020). Algorithms for detection gender using neural networks // International Journal of Circuits, Systems and Signal Processing. 2020, 154-159.

[9] Mekebayev N., Tuyebaev Ch., Sabrayev K., Yerkebay A. Researchof acoustic and linguistic modeling based on repetitive neural networks for speech recognition of children // Bulletin of physics & mathematical sciences. No1(77), 2022, https://doi.org/10.51889/2022-1.1728-7901.16 , No1(77), 2022, 119-126

[10] Freund, Y., & Schapire, R. E. (1999, Dec). Large margin classification using the perceptron algorithm. Machine Learning, 37(3), 277–296. doi:10.1023/ A:1007662407062

References

[1] Auer, P., Burgsteiner, H., & Maass, W (2019). A learning rule for very simple universal approximators consisting of a single layer of perceptrons. Neural Networks: The Official Journal of the International Neural Network Society, 21(5), 786–795.

[2] Rabiner L (2019). A tutorial on hidden markov models and selected applications in speech recognition, 257–286.

[3] Orken Mamyrbayev, Nurbapa Mekebayev, Mussa Turdalyuly, Nurzhamal Oshanova,Tolga Ihsan Medeni and Aigerim Yessentay (2019). Voice Identification Using Classification Algorithms. We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists. London, 1 – 14 p.

[4] Cunningham, P., & Delany, S. (2019). k-nearest neighbour classifiers. Multiple Classifier System, 1–17.

[5] Mermelstein P (2020). Distance measures for speech recognition, psychological and instrumental. Pattern recognition and artificial intelligence. Vol. 116. 374–388.

[6] Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2019). Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech and Language Processing. 19(4), 1–11

[7] Toleu, A., Tolegen, G., & Makazhanov, A. (2021). Character-aware neural morphological disambiguation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics 2, 666–671. Association for Computational Linguistics, Vancouver, Canada.

[8] Kalimoldayev, M., Mamyrbayev, O., Mekebayev, N., Kydyrbekova, A (2020). Algorithms for detection gender using neural networks// International Journal of Circuits, Systems and Signal Processing. 2020, 154-159

[9] Mekebayev N., Tuyebaev Ch., Sabrayev K., Yerkebay A. Researchof acoustic and linguistic modeling based on repetitive neural networks for speech recognition of children // Bulletin of physics & mathematical sciences. No1(77), 2022, https://doi.org/10.51889/2022-1.1728-7901.16 , No1(77), 2022, 119-126

[10] Freund, Y., & Schapire, R. E. (1999, Dec). Large margin classification using the perceptron algorithm. Machine Learning, 37(3), 277–296. doi:10.1023/ A:1007662407062