Методы аугментации аудио сигнала

Ю.М. Романовская, Е.А. Ильюшин

***Аннотация*—Задача распознавания звука с каждым годом становится всё более актуальной и востребованной. На примере задачи распознавания голосовых команд становится понятно, что требуется большое количество обучающих данных, так как модели должны учитывать разницу тембров, скорости речи, особенности дикции и многих других факторов. Реальный сбор этих данных представляется очень трудоемким, а по сути и невозможным. Вследствие чего активно ищутся способы автоматического обогащения обучающих наборов данных.**

Аугментация – это методика создания дополнительных данных на основе имеющихся. Есть два принципиально отличающихся подхода построения метода аугментации. В первом подходе на вход подают существующие данные, а возвращают те же данные, но с изменёнными характеристиками (например, ускоренные или более громкие сэмплы). Во втором предполагается использование синтетических данных, порождаемых моделью, обученной на реальных.

В рамках данной статьи выполнен обзор всего спектра существующих на сегодняшний день методов аугментации. Проведены эксперименты на базовых методах, сделаны выводы о применении и использовании представленных методов и об их влиянии на качество распознавания звука в рамках задачи распознавания голосовых команд.

***Ключевые слова*—аугментация, аудиосигнал, спектрограмма, GAN.**

I. Введение

В данной статье проводится полный обзор существующих методов аугментации звуковых данных. В задаче аугментации данных помимо метода преобразования важную роль играет представление аудиоданных. Аудиосигнал может по-разному храниться и обрабатываться в компьютере, и у каждого представления есть определенные особенности. *Аудиосигнал* – физический процесс, представляющий собой распространение акустической энергии в виде упругих волн механических колебаний в жидкой, газообразной и твердой среде. Чтобы работать со звуком посредством машинного обучения, нужно преобразовать его в числовые последовательности, что осуществляется измерения амплитуды сигнала в определенные интервалы времени [1]. Из числовых последовательностей мы можем получить представление звука в виде изображения (спектрограммы) [2] и мел-частотных коэффициентов [3], которые определены на частотах соответствующих голосу человека. Подробнее об их построении рассказывается в разделе II. В зависимости от выбранного представления звука, существуют различные способы аугментации. Обработка исходного аудиосигнала [4], [5], [6] интуитивно понятна, сюда входят стандартные способы увеличение/уменьшение громкости, высоты тона, ускорение темпа. Обработка спектрограмм включает способы работы с изображениями, но классические способы аугментации изображений мало того, что не улучшают модели, в некоторых экспериментах даже делают результаты хуже [7]. SpecAugment [8] – способ аугментации спектрограмм, маскирующий участки на частотно-временных представлениях, показывает отличные результаты. Наиболее трудоемким способом аугментации является использование порождающих моделей глубокого обучения [9], [10] для генерации новых звуковых данных (WaveGAN), а также новых спектрограмм (SpecGAN). Существуют модели, которые порождают мел-частотные коэффициенты [11]. Подробно о всех способах рассказывается в разделе III. В разделе IV отражены результаты экспериментов на базовых и наиболее распространённых методах на сегодняшний день, а также проведен сравнительный анализ. Результаты и подведение итогов представлены в разделе V.

Эта статья является продолжением серии публикаций, посвященных устойчивым моделям машинного обучения [12], [13]. Она подготовлена в рамках проекта кафедры Информационной безопасности факультета ВМК МГУ имени М.В. Ломоносова по созданию и развитию магистерской программы «Искусственный интеллект в кибербезопасности» [14].

II. Представление данных

A. Аудиосигнал

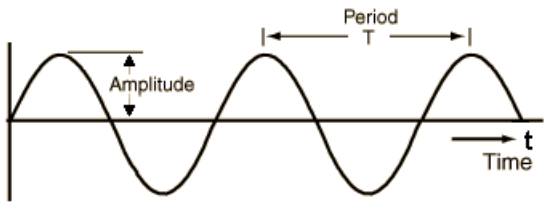

Рис. 1: Звуковой сигнал.

На рисунке 1 представлен звуковой сигнал, который характеризуется амплитудой, периодом и частотой. Амплитуда показывает интенсивность сигнала, а период – длину волны. Количество волн в сигнале в секунду называется частотой. Таким образом частота является обратной величиной периода. Чаще всего сигнал представляет собой композицию сигналов со сложными формой и периодом.

Статья получена 30 января 2023.

Юлия Михайловна Романовская, МГУ им. М.В. Ломоносова, (email: rom.ym@mail.ru).

Евгений Альбинович Ильюшин, МГУ им. М.В. Ломоносова, (email: eugene.ilyushin@gmail.com).

Исследование выполнено при поддержке Междисциплинарной научно-образовательной школы Московского университета «Мозг, когнитивные системы, искусственный интеллект».

Для обучения моделей используются числовые последовательности, получаемые путем измерения значений амплитуды сигнала в фиксированные интервалы времени [1]. Такие замеры амплитуды называются сэмплами, а частота дискретизации – это количество таких замеров в секунду.

B. Спектрограмма

Спектрограмма – это представление сигнала в виде спектров [2], полученных на каждом временном отрезке анализируемого сигнала (Рис. 2).

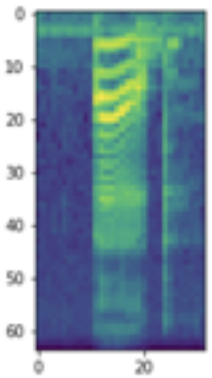

Рис. 2: Спектрограмма голосовой команды «nine» из данных, используемых в эксперименте.

Спектр – разложение аудиосигнала по гармоникам. Гармоника – это частоты кратные самой низкой частоте исходного сигнала. У спектрограммы на оси абсцисс отложено время, а на оси ординат частота, на самой спектрограмме откладывается спектр в разные моменты времени. Третье измерение, показывающее изменение амплитуды, отмечается интенсивностью цвета. С помощью спектрограмм задача обработки звука может быть представлена как задача обработки изображений.
Для построения спектрограммы используется преобразование Фурье (II-B.1).

$$X_a[k] = \sum_{i=0}^{n-1} x[n] e^{\frac{-2\pi i}{n} kn}, 0 \leq k \leq N-1 \qquad \text{(II-B.1)}$$

Преобразование Фурье разлагает сигнал на составляющие его частоты и отображает амплитуду каждой частоты, присутствующей в сигнале. На практике используется кратковременное преобразование Фурье, которое разбивает звуковой сигнал на кадры, так как в обычном преобразовании невозможно распознать в какой момент времени было конкретное распределение частот. Из временных кадров формируется спектрограмма. Если использовать спектрограммы, полученные таким образом, сигнал на них будет практически не заметен или сосредоточен в одной области. Это происходит из-за того, что человек иначе воспринимает звук. Проблема решается введением новой единицы измерения высоты звука – мел (II-B.2), логарифмически зависящая от частоты. Она посчитана эмпирически.

$$mel = 1127.010448 \ln(1 + \frac{freq}{700}) \qquad \text{(II-B.2)}$$

Громкость звука аналогичным образом воспринимается в логарифмической шкале, поэтому используют шкалу децибелов (дБ), единицы измерения которых возрастают экспоненциально. Например, 0 дБ – абсолютная тишина, 10 дБ – в 10 раз громче, 20 дБ – в 100 раз громче и т.д. Обычно под спектрограммой понимают мел-спектрограмму, где шкала частоты измеряется в мелах, а шкала громкости измеряется в децибелах.

C. Мел-частотные кепстральные коэффициенты

Для анализа голоса зачастую применяются мел-частотные кепстральные коэффициенты [3]. Для их построения вся запись сначала разбивается на кадры. Из-за того, что речевой сигнал не периодичен и конечен, в промежутках между фонемами возникает резкое падение амплитуды, что провоцирует появление большого количества шума. Для его устранения используются оконные функции (например, окна Хэмминга или Ханнинга). Далее получают спектр с помощью дискретного преобразования Фурье. Полученные спектральные коэффициенты пропускаются через мел-фильтры, сосредоточенные ближе к низким частотам. После вычисляют энергию каждого кадра, применяют дискретное косинусное преобразование, а на выходе получают мел-частотные кепстральные коэффициенты, которые также можно представить в виде изображений. (Рис. 3).

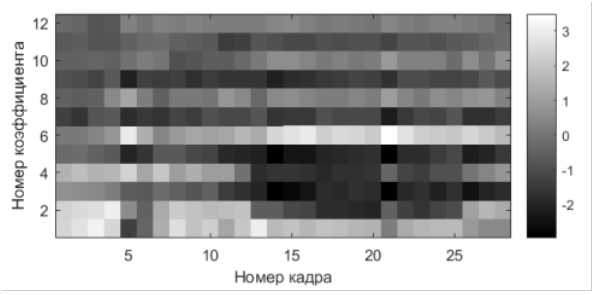

Рис. 3: Мел-частотные кепстральные коэффициенты.

III. Методы аугментации

Аугментация данных – это обычная стратегия, используемая для обогащения обучающей выборки. Есть различные способы аугментации, которые применяют к исходному сигналу [4], [5], [11], к спектрограммам [7], а также к мел-частотным кепстральным коэффициентам [9], [15]. Один из новых подходов – порождение звуковых данных или спектрограмм [8], [9].

Рассмотрим те из них, которые применяются при решении индустриальных задач.

A. Обработка звукового сигнала

Есть несколько основных способов аугментации, которые применяются к исходному аудиосигналу, они изменяют сигнал до этапа извлечения признаков:

1) Изменение громкости (увеличение или уменьшение).
2) Изменение скорости (ускорение или замедление).

3) Сдвиг высоты тона на случайное число в полутонах (повышение или понижение) при сохранении неизменной длительности [16].
4) Добавление случайного шума, тишины, смешивание сигнала с фоновыми звуками из разных типов акустических сцен.
5) Сдвиг по времени (вправо или влево).
6) Добавление реверберации – наложение эхо-эффекта [17], [18]. *Реверберация* – уменьшение интенсивности звука путём отражения звукового сигнала. Используют три способа:
   a) На вход CNN приходят чистые данные, в процессе свёртки применяется искусственный звук реверберации, и на выходе получают новые данные (Рис. 4).

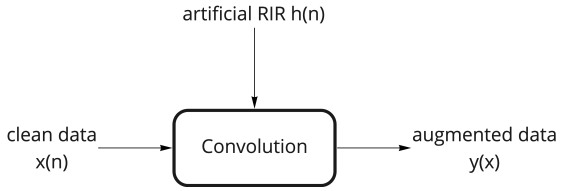

Рис. 4: Схема 1 добавления реверберации.

   b) Обучают на чистых данных в качестве входных данных и на выходных в виде аудиосигналов с эффектом реверберации. Затем уже получают реальные аугментации, применяя обученную CNN модель (Рис. 5).

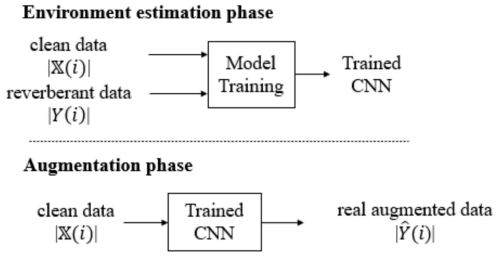

Рис. 5: Схема 2 добавления реверберации.

   c) Используется алгоритм генерации шума для создания эффекта реверберации. В основе алгоритма симуляции фонового шума лежит уравнение (III-A.1).

$$x_r[t] = x[t] * h_s[t] + \sum n_i[t] * h_i[t] + d[t], \quad \text{(III-A.1)}$$

   где x_r – результат применения реверберации; x – исходный сигнал; h_s – источник реверберации, соответствующий положению динамика; n_i – шум точечного источника; d – другие источники аддитивного шума.
7) Модификация некоторых частот в сигнале случайным образом.
8) Сжатие динамического диапазона (увеличение громкости громких звуков, уменьшение громкости тихих звуков).
9) **WSOLA** (Waveform similarity overlap-add) [19] – задача увеличения темпа при сохранении тембра, голоса, высоты тона и качества звука. WSOLA достигается путем разложения аудио сегмента временной области x(t) на короткие блоки, а затем перемещения этих блоков вдоль временной оси для построения выходного аудиосигнала.
10) Смешивание исходных сигналов [20]. Есть несколько способов создания новых данных при использовании совокупности старых. Различают три вида: Mixup, SamplePairing, Mixup with label preserving.
   a) **Mixup** – генерирует образцы с использованием линейной интерполяции, получая из (X_i, Y_i) и (X_j, Y_j) новый сигнал X_n (III-A.2) и новую метку Y_n (III-A.3), где $\lambda \sim \beta(\alpha, \alpha)$, а $\alpha \sim (0, \infty)$.
   Новые метки используют взвешенную вероятность старых меток. Y_i, Y_j подаются в формате one-hot векторов, на выходе получают вектор вероятностей Y_n.

$$X_n = * X_i + (1 - \lambda) * X_j \quad \text{(III-A.2)}$$

$$Y_n = \lambda * Y_i + (1 - \lambda) * Y_j \quad \text{(III-A.3)}$$

   b) **SamplePairing** – берется среднее значение двух входных векторов (III-A.4), метка выбирается такой же, как и у первого вектора (III-A.5).

$$X_n = 0.5 * X_i + 0.5 * X_j \quad \text{(III-A.4)}$$

$$Y_n = Y_i \quad \text{(III-A.5)}$$

   c) **Mixup with label preserving** – является комбинацией двух предыдущих методов, выборка расширяется линейной интерполяцией (III-A.6), но метки не вычисляются заново (III-A.7).

$$X_n = (1 + \lambda) * X_i - \lambda * X_j \quad \text{(III-A.6)}$$

$$Y_n = Y_i \quad \text{(III-A.7)}$$

B. Обработка спектрограмм

Представление аудиосигнала в виде спектрограмм позволяет применять методы аугментаций изображений. Классические методы аугментации изображений, такие как отражения, повороты, добавление шума только ухудшают результаты распознавания аудио, поэтому были разработаны новые способы применимые к спектрограммам.

1) **SpecAugment** [7].
   Аугментация применяется на входных данных нейронной сети и подразумевает маскирование по временной и частотной осях мел-спектограмм.
   Чтобы сеть извлекала полезные признаки и была устойчива к искажениям во времени, частичной потере частотной информации и частичной потере небольших сегментов речи, была разработана следующая политика аугментации:
   a) **TimeWarp** – искажение времени. Выбирается случайная точка вдоль горизонтальной полосы, проходящей через центр изображения в пределах временных шагов (W, t-W), и сдвигается влево или вправо вдоль этой линии на расстояние w, выбранное из равномерного распределения $[0, W]$, где W – параметр.

Маскирование данных – наложение случайной маски на спектрограмму и обнуление замаскированных коэффициентов.

b) **Частотное маскирование** – маскируются f последовательных каналов частоты $[f_0, f_0+f)$, где f берется из равномерного распределения $[0, F]$, а f_0 выбирается из $[0, v-f)$, где v – количество частотных каналов mel (Рис. 6).

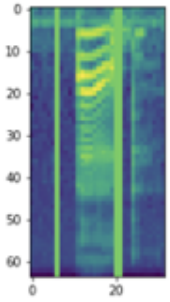

Рис. 6: Частотное маскирование голосовой команды «nine» из данных, используемых в эксперименте.

c) **Временное маскирование** – t последовательных временных шагов $[t_0, t_0+f)$ маскируются, где t берется из равномерного распределения $[0, T]$, а t_0 из $[0, T-t)$ (Рис. 7).

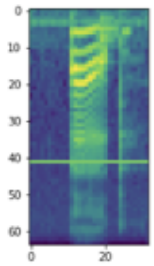

Рис. 7: Временная маскировка голосовой команды «nine» из данных, используемых в эксперименте.

2) Создание новой спектрограммы путем суммирования двух случайных спектрограмм принадлежащих одному и тому же классу [4]. Возможно применение техник линейной интерполяции спектрограмм и получения взвешенных меток, как в случае обработки исходного сигнала [20].
3) Случайным образом применяется сдвиг высоты тона и сдвиг по времени.
4) **VTLP** (Vocal Tract Length Perturbation) [21] – метод преобразования спектрограмм с использованием случайного линейного искажения по частотному измерению. Основная идея заключается в применении нормализации не для того, чтобы убрать различия, а наоборот, добавить вариации в аудио. Это может быть получено нормализацией к произвольной цели вместо нормализации к каноническому среднему. Добавляет вариабельности к речевым данным, имитируя различную длину голосового тракта.

   Для VTLP генерируется коэффициент деформации α для каждого примера, и деформируется ось частот так, чтобы частота f отображалась на новую частоту f' (4).

$$f' = \begin{cases} f\alpha, f \leq F_{hi}\frac{min(\alpha,1)}{\alpha} \\ \frac{S}{2} - \frac{\frac{S}{2} - F_{hi}min(\alpha,1)}{\frac{S}{2} - F_{hi}\frac{min(\alpha,1)}{\alpha}}(\frac{S}{2} - f) \text{ иначе,} \end{cases}$$

   где S – частота дискретизации, F_{hi} – граница частот. Процедуру деформации можно применить не к спектрограммам, а к самим фильтрам (этап получения MFCC).
5) **Композиция аугментаций**. Как пример используется суммирование спектрограмм, на одну из которых применяются временные сдвиги, и на обе спектрограммы при суммировании применяется VTLP (2.9).

$$saug(t) = \alpha\varphi(s_1(t), \varphi_1) + (1-\alpha)\cdot\varphi(s_2(t-), \varphi_2), \quad \text{(III-B.1)}$$

   где α, β – рандомные значения из $[0, 1]$, T — временной сдвиг из промежутка $[1, M]$, M — ширина спектрограммы и функция выравнивания P_{hi}, параметризованная вектором $\varphi = (f_0, g, Q)$. f_0 – центральная частота случайно выбранная из заданного промежутка, g – коэффициент усиления и Q – Q-фактор.
6) Две спектрограммы, разрезанные в точке T, меняются местами, образуя новую спектрограмму [4]. Такой вид аугментации подходит далеко не для всех задач распознавания звука. Данная аугментация применялась в задаче распознавания звуков животных.

C. Порождение данных

До этого рассматривались методы, которые каким-либо образом изменяли существующие данные. В качестве методов аугментации также могут быть рассмотрены и методы порождения синтетических данных [8], [9].

1) Модели, основанные на порождающих состязательных сетях (GAN) [19]. GAN работает по следующему принципу: объединены две конкурирующих сети, которые улучшают друг друга. Одна сеть – генератор, учится преобразовывать любой вектор, который следует заданной функции распределения, в совершенно новый. Во второй сети дискриминатор учится различать реальные данные и синтетические, созданные генератором.
   a) **WaveGAN** применяет модель GAN для синтеза необработанного звука в неконтролируемых условиях. WaveGAN может создавать слова из небольшого запаса человеческой речи, а также синтезировать аудио из различных областей: вокализация птиц, фортепиано и т.д. Основа модели – DCGAN с модифицированной операцией свёртки (1x25 вместо 5x5).
   b) **SpecGAN** применяет модель GAN, используя её на изображениях, для генерации спектрограмм. Этот способ работает хуже, чем WaveGAN. Существенный недостаток моделей GAN – случайная генерация выходных данных. Это приводит к неконтролируемому увеличению обучающих данных, которое может не оказать никакого влияния на обучение

классификатора или даже ослабить его при работе с небольшими наборами данных. Поэтому был предложен новый способ.

c) **Эволюционно-генеративный подход**

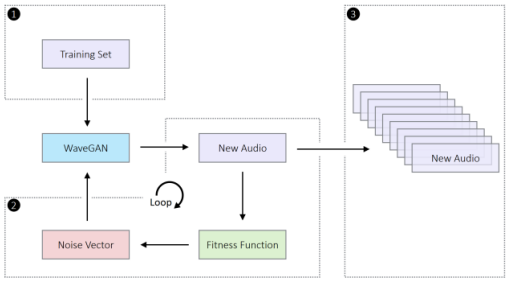

Рис. 8: Эволюционно-генеративный алгоритм [19].

На первом этапе используется GAN для создания реалистичных аудиоданных, на втором – применяется эволюционный алгоритм для поиска в исходном пространстве порождающей модели векторов, которые приводят к выборкам с предопределенными характеристиками (рис. 8). Вводится понятие эволюционной функции, которая инициализирует вектор шума z и подает в алгоритм G, который получает звук. Результат подставляется в функцию f(G(z)), которая оценивает, насколько он совпал с желаемым t. Плохие результаты отбрасываются, а самые лучше рекомбинируются [20] и снова подаются алгоритму.

$$\text{fitness}(z) = \frac{1}{(|f(G(z)) - t|)} \qquad \text{(III-C.1)}$$

2) **Генерация MFCC коэффициентов**
RNN char [10].. Символы в данном случае это коэффициенты. RNN учится моделировать распределение вероятностей следующего символа в последовательности после наблюдения за предыдущими символами, также созданными RNN.

IV. Эксперименты

В данном разделе проведены эксперименты на задаче распознавания голосовых команд. При исследовании существующих методов аугментации были выявлены наиболее широко используемые методы [16]. На них была проверена задача со следующей постановкой:

Всего 30 классов, для наглядности разбитые на группы:

- «on», «two», «three», «four», «five», «six», «seven», «eight», «nine»;
- «up», «down», «left», «right»;
- «yes», «no»;
- «on», «off»;
- «bird», «cat», «dog»;
- «Marvin», «Sheila»;
- «bed», «house», «tree»;
- «go», «happy», «stop», «wow».

В каждом классе находится от 1700 до 2400 аудиофайлов, где различные дикторы произносят слова (названия классов). Набор данных состоит из 64727 файлов. Размер обучающей выборки составляет 51094. Остальные файлы поделены на валидационную и тестовую выборки поровну.

Таблица I: Результаты применения методов аугментации.

Data	Epoch	1	2	3	4	5	6
NA	**Acc(%)**	83.2	87.2	88.5	89.3	89.3	89.3
	AvgL	0.55	0.41	0.39	0.36	0.37	0.37
SA	**Acc(%)**	86.8	88.8	90.3	90.2	90.4	91.2
	AvgL	0.43	0.39	0.35	0.37	0.36	0.35
PS AGN	**Acc(%)**	**87.1**	**90.3**	**91.0**	**90.7**	**91.3**	**91.5**
	AvgL	**0.43**	**0.32**	**0.31**	**0.31**	**0.32**	**0.33**
TSt	**Acc(%)**	86.1	88.4	89.0	90.5	90.3	90.4
	AvgL	**0.42**	0.33	0.32	0.31	0.32	0.33
TSh	**Acc(%)**	86.5	87.3	89.0	90.3	89.3	90.3
	AvgL	0.44	0.34	0.34	0.33	0.34	0.33

Каждая аугментация применялась ко всем файлам в обучающей выборке, что приводило к расширению обучающих файлов в два раза. Частота дискретизации каждого файла – 16000, каждый приводился к длине в одну секунду. Если файл оказывался длиннее, лишнее отрезалось, а если короче – недостающие данные заполнялись нулями (тишиной). Перед подачей аудиофайла в модель классификатора (Resnet18) по нему строилось трехканальное изображение мел-спектрограммы. В качестве функции потерь использовали кросс-энтропию, в роли оптимизатора – алгоритм Адам. В статье сравнивали на описанном ваше наборе данных 4 метода аугментации:

1) SpecAugment.
2) Добавление гауссовского шума (AddGaussianNoise) с повышением, понижением тона (PitchShift).
3) Растяжение времени (TimeStretch).
4) Сдвиг времени (TimeShift).

Для реализации методов аугментации исходного аудиосигнала применялась библиотека audiomentations на базе библиотеки librosa, язык Python [19]. Из неё были использованы функции PitchShift (PS), AddGaussianNoise (AGN), TimeStretch (TSt), TimeShift (TSh) и Compose – для композиции методов аугментации.

Для реализации метода SpecAugment, изменяющего спектрограммы, применялись функции временного и частотного маскирования [22], предложенные в качестве реализации к основной статье [7] по алгоритму SpecAugment, язык Python. В таблице I представлены результаты применения выбранных методов аугментации, продемонстрированы метрики Accuracy (Acc) и average loss (AvgL) для каждой из 6 эпох.

На матрице confusion matrix (Рис. 9) для методов добавления гауссовского шума и изменения высоты тона видно, что классификация слов проходит равномерно без выделения определенного класса.

По результатам экспериментов лучше всего себя показали аугментации добавления гауссовского шума и изменения высоты тона при одновременном использовании,что говорит о перспективе получения лучших результатов при использовании композиции аугментаций.

V. Заключение

В данной статье была рассмотрена задача аугментации звуковых данных. Был проведен детальный обзор

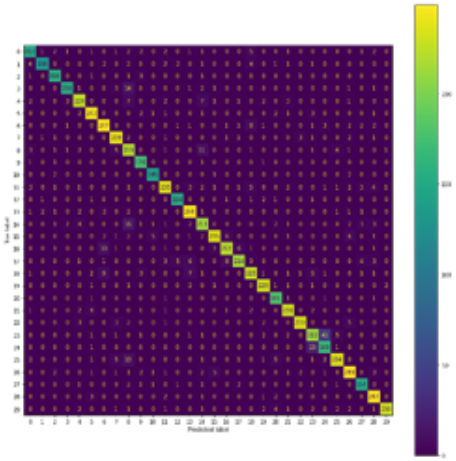

Рис. 9: Confusion matrix тона для методов AddGaussianNoise() и PitchShift().

существующих и используемых методов аугментации как исходного сигнала, так и спектрограмм. Отдельное внимание было уделено методам порождения синтетических данных, которые в текущей обстановке развития нейронных сетей, набирают всё большую популярность. Были поставлены эксперименты на базовых методах, таких как добавление шума, сдвиг тона, растяжение и сдвиг времени и на самом новом из них, применяемом на спектрограммах – SpecAugment. Код экспериментов доступен на Github [23].

Список литературы

[1] Chauhan Nagesh. Audio data analysis using deep learning with python (part 1). — URL: https://www.kdnuggets.com/2020/02/audio-data-analysis-deep-learning-python-part-1.html.

[2] Oppenheim Alan V. Speech spectrograms using the fast fourier transform. — Vol. 7, no. 8. — P. 57–62. — URL: http://ieeexplore.ieee.org/document/5213512/ (online; accessed: 2023-01-13).

[3] Muda Lindasalwa, Begam Mumtaj, Elamvazuthi I. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. — arxiv : 1003.4083 [cs].

[4] Nanni Loris, Maguolo Gianluca, Paci Michelangelo. Data augmentation approaches for improving animal audio classification. — arxiv : 1912.07756 [cs, eess, stat].

[5] Investigation of data augmentation techniques for disordered speech recognition / Mengzhe Geng, Xurong Xie, Shansong Liu et al. // Interspeech 2020. — P. 696–700. — arxiv : 2201.05562 [cs, eess].

[6] Audio augmentation for speech recognition / Tom Ko, Vijayaditya Peddinti, Daniel Povey, Sanjeev Khudanpur // Interspeech 2015. — ISCA. — P. 3586–3589. — URL: https://www.isca-speech.org/archive/interspeech_2015/ko15_interspeech.html (online; accessed: 2023-01-13).

[7] SpecAugment: A simple data augmentation method for automatic speech recognition / Daniel S. Park, William Chan, Yu Zhang et al. // Interspeech 2019. — P. 2613–2617. — arxiv : 1904.08779 [cs, eess, stat].

[8] Yang Jeong Hyeon, Kim Nam Kyun, Kim Hong Kook. Se-resnet with gan-based data augmentation applied to acoustic scene classification // DCASE 2018 workshop. — 2018.

[9] Donahue Chris, McAuley Julian, Puckette Miller. Adversarial audio synthesis. — arxiv : 1802.04208 [cs].

[10] Overcoming data scarcity in speaker identification: Dataset augmentation with synthetic MFCCs via character-level RNN / Jordan J. Bird, Diego R. Faria, Cristiano Premebida et al. // 2020 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC). — IEEE. — P. 146–151. — URL: https://ieeexplore.ieee.org/document/9096166/ (online; accessed: 2023-01-13).

[11] A comparison on data augmentation methods based on deep learning for audio classification / Shengyun Wei, Shun Zou, Feifan Liao, weimin lang. — Vol. 1453, no. 1. — P. 012085. — URL: https://iopscience.iop.org/article/10.1088/1742-6596/1453/1/012085 (online; accessed: 2023-01-13).

[12] Ilyushin Eugene, Namiot Dmitry, Chizhov Ivan. Attacks on machine learning systems-common problems and methods. — Vol. 10, no. 3. — P. 17–22.

[13] Stroeva Ekaterina, Tonkikh Aleksey. Methods for formal verification of artificial neural networks: A review of existing approaches. — Vol. 10, no. 10. — P. 21–29.

[14] Artificial intelligence in cybersecurity. — URL: https://cs.msu.ru/node/3732 (online; accessed: 2022-12).

[15] Rejaibi Emna, Komaty Ali, Meriaudeau Fabrice et al. MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech. — arxiv : 1909.07208 [cs, eess].

[16] Rahman Aamer Abdul, Angel Arul Jothi J. Classification of Urban-Sound8k: A study using convolutional neural network and multiple data augmentation techniques // Soft Computing and its Engineering Applications / Ed. by Kanubhai K. Patel, Deepak Garg, Atul Patel, Pawan Lingras. — Springer Singapore. — Vol. 1374. — P. 52–64. — Series Title: Communications in Computer and Information Science. URL: https://link.springer.com/10.1007/978-981-16-0708-0_5 (online; accessed: 2023-01-13).

[17] A study on data augmentation of reverberant speech for robust speech recognition / Tom Ko, Vijayaditya Peddinti, Daniel Povey et al. // 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). — IEEE. — P. 5220–5224. — URL: http://ieeexplore.ieee.org/document/7953152/ (online; accessed: 2023-01-13).

[18] Yun Deokgyu, Choi Seung Ho. Deep learning-based estimation of reverberant environment for audio data augmentation. — Vol. 22, no. 2. — P. 592. — URL: https://www.mdpi.com/1424-8220/22/2/592 (online; accessed: 2023-01-13).

[19] An evolutionary-based generative approach for audio data augmentation / Silvan Mertes, Alice Baird, Dominik Schiller et al. // 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP). — IEEE. — P. 1–6. — URL: https://ieeexplore.ieee.org/document/9287156/ (online; accessed: 2023-01-13).

[20] Wei Shengyun, Xu Kele, Wang Dezhi et al. Sample mixed-based data augmentation for domestic audio tagging. — arxiv : 1808.03883 [cs, eess].

[21] Jaitly Navdeep, Hinton Geoffrey E. Vocal tract length perturbation (vtlp) improves speech recognition // Proc. ICML Workshop on Deep Learning for Audio, Speech and Language. — Vol. 117. — 2013. — P. 21.

[22] Caceres Zach. Library for SpecAugment realization. — URL: https://github.com/zcaceres/spec_augment (online; accessed: 2023-01-30).

[23] Romanovskaya Yulia. Experiments. — URL: https://github.com/fabuloudy/augmentations (online; accessed: 2023-01-30).

Sound augmentation methods

Y.M. Romanovskaya, E.A. Ilyushin

***Abstract*—The problem of sound recognition is becoming more relevant and in demand every year. Considering the recognition voice commands task, it becomes clear that a large amount of training data is required, since models must take the difference in timbres, speed, diction features, and many other factors into account. The actual collection of this data is to be very time-consuming, but in fact, impossible. As a result, the search for algorithms for the automatic creation of training synthetic datasets is actively underway.**

Augmentation is a method of creating additional data based on existing ones. There are two fundamentally different approaches. The first approach takes existing data as input and returns the same data, but with changed characteristics (i.e., accelerated or louder samples). The second method uses the original data only for training the model, and generates new data independently.

This article provides an overview of the entire spectrum of existing augmentation methods. We try several methods in our experiments and make conclusions about the application and usage of the presented approaches as well as their impact on the quality of sound recognition an example of a voice recognition task.

***Keywords*—augmentation, audio signal, spectrogram, GNN**

References

[1] Chauhan Nagesh. Audio data analysis using deep learning with python (part 1). — URL: https://www.kdnuggets.com/2020/02/audio-data-analysis-deep-learning-python-part-1.html.

[2] Oppenheim Alan V. Speech spectrograms using the fast fourier transform. — Vol. 7, no. 8. — P. 57–62. — URL: http://ieeexplore.ieee.org/document/5213512/ (online; accessed: 2023-01-13).

[3] Muda Lindasalwa, Begam Mumtaj, Elamvazuthi I. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. — arxiv : 1003.4083 [cs].

[4] Nanni Loris, Maguolo Gianluca, Paci Michelangelo. Data augmentation approaches for improving animal audio classification. — arxiv : 1912.07756 [cs, eess, stat].

[5] Investigation of data augmentation techniques for disordered speech recognition / Mengzhe Geng, Xurong Xie, Shansong Liu et al. // Interspeech 2020. — P. 696–700. — arxiv : 2201.05562 [cs, eess].

[6] Audio augmentation for speech recognition / Tom Ko, Vijayaditya Peddinti, Daniel Povey, Sanjeev Khudanpur // Interspeech 2015. — ISCA. — P. 3586–3589. — URL: https://www.isca-speech.org/archive/interspeech_2015/ko15_interspeech.html (online; accessed: 2023-01-13).

[7] SpecAugment: A simple data augmentation method for automatic speech recognition / Daniel S. Park, William Chan, Yu Zhang et al. // Interspeech 2019. — P. 2613–2617. — arxiv : 1904.08779 [cs, eess, stat].

[8] Yang Jeong Hyeon, Kim Nam Kyun, Kim Hong Kook. Se-resnet with gan-based data augmentation applied to acoustic scene classification // DCASE 2018 workshop. — 2018.

[9] Donahue Chris, McAuley Julian, Puckette Miller. Adversarial audio synthesis. — arxiv : 1802.04208 [cs].

[10] Overcoming data scarcity in speaker identification: Dataset augmentation with synthetic MFCCs via character-level RNN / Jordan J. Bird, Diego R. Faria, Cristiano Premebida et al. // 2020 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC). — IEEE. — P. 146–151. — URL: https://ieeexplore.ieee.org/document/9096166/ (online; accessed: 2023-01-13).

[11] A comparison on data augmentation methods based on deep learning for audio classification / Shengyun Wei, Shun Zou, Feifan Liao, weimin lang. — Vol. 1453, no. 1. — P. 012085. — URL: https://iopscience.iop.org/article/10.1088/1742-6596/1453/1/012085 (online; accessed: 2023-01-13).

[12] Ilyushin Eugene, Namiot Dmitry, Chizhov Ivan. Attacks on machine learning systems-common problems and methods. — Vol. 10, no. 3. — P. 17–22.

[13] Stroeva Ekaterina, Tonkikh Aleksey. Methods for formal verification of artificial neural networks: A review of existing approaches. — Vol. 10, no. 10. — P. 21–29.

[14] Artificial intelligence in cybersecurity. — URL: https://cs.msu.ru/node/3732 (online; accessed: 2022-12).

[15] Rejaibi Emna, Komaty Ali, Meriaudeau Fabrice et al. MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech. — arxiv : 1909.07208 [cs, eess].

[16] Rahman Aamer Abdul, Angel Arul Jothi J. Classification of Urban-Sound8k: A study using convolutional neural network and multiple data augmentation techniques // Soft Computing and its Engineering Applications / Ed. by Kanubhai K. Patel, Deepak Garg, Atul Patel, Pawan Lingras. — Springer Singapore. — Vol. 1374. — P. 52–64. — Series Title: Communications in Computer and Information Science. URL: https://link.springer.com/10.1007/978-981-16-0708-0_5 (online; accessed: 2023-01-13).

[17] A study on data augmentation of reverberant speech for robust speech recognition / Tom Ko, Vijayaditya Peddinti, Daniel Povey et al. // 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). — IEEE. — P. 5220–5224. — URL: http://ieeexplore.ieee.org/document/7953152/ (online; accessed: 2023-01-13).

[18] Yun Deokgyu, Choi Seung Ho. Deep learning-based estimation of reverberant environment for audio data augmentation. — Vol. 22, no. 2. — P. 592. — URL: https://www.mdpi.com/1424-8220/22/2/592 (online; accessed: 2023-01-13).

[19] An evolutionary-based generative approach for audio data augmentation / Silvan Mertes, Alice Baird, Dominik Schiller et al. // 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP). — IEEE. — P. 1–6. — URL: https://ieeexplore.ieee.org/document/9287156/ (online; accessed: 2023-01-13).

[20] Wei Shengyun, Xu Kele, Wang Dezhi et al. Sample mixed-based data augmentation for domestic audio tagging. — arxiv : 1808.03883 [cs, eess].

[21] Jaitly Navdeep, Hinton Geoffrey E. Vocal tract length perturbation (vtlp) improves speech recognition // Proc. ICML Workshop on Deep Learning for Audio, Speech and Language. — Vol. 117. — 2013. — P. 21.

[22] Caceres Zach. Library for SpecAugment realization. — URL: https://github.com/zcaceres/spec_augment (online; accessed: 2023-01-30).

[23] Romanovskaya Yulia. Experiments. — URL: https://github.com/fabuloudy/augmentations (online; accessed: 2023-01-30).