«Сейфуллин оқулары – 18: « Жастар және ғылым – болашаққа көзқарас» халықаралық ғылыми -практикалық конференция материалдары = Материалы международной научно-практической конференции «Сейфуллинские чтения – 18: « Молодежь и наука – взгляд в будущее» - 2022.- Т.I, Ч.VI. – С.20-23

ЗАВИСИМОСТЬ ТОЧНОСТИ РАСПОЗНАВАНИЯ РЕЧЕВОЙ АКТИВНОСТИ ОТ КОЛИЧЕСТВА ДИКТОРОВ

Ертаев А.М., магистрант 2 курса,
Казахский агротехнический университет им. С. Сейфуллина, г. Нур-Султан

В телекоммуникационных системах связи одной из главных задач является фильтрация передаваемых сигналов с целью уменьшения их объема. Для осуществления этой задачи используются различные алгоритмы. Например, при передаче голосовых данных для отсечения звуков, не относящихся к человеческой речи, используются системы, которые называются детекторами активности речи, по английский - Voiceactivity detector (VAD).

Системы типа VAD основываются на различных алгоритмах. Далее приведены несколько таких алгоритмов.

Ключевые слова: диктор, VAD, алгоритмы, нейронные сети, шум, данные.

В работе [1] предлагается алгоритм, основанный на полиномиальной регрессии второго порядка, с аналогичной функцией в качестве VAD для систем проверки голоса, не зависящих от текста. Предлагаемый способ направлен на разделение областей устойчивого шума/тишины, областей устойчивой речи и областей начала/смещения речи. Регрессия применяется независимо к каждой полосе фильтра спектра mel, что позволяет алгоритму легко вписываться в обычный процесс извлечения коэффициентов кепстрала частоты mel (MFCC). Алгоритм k-средних также применяется для оценки средней энергии шума в каждой полосе для спектрального вычитания.

А в следующей работе [2] было показано, что в условиях сильного искажения звукового сигнала может быть достигнута точная сегментация целевой речи путем объединения нескольких потоков функций. В этой работе извлекаются четыре одномерных потока, каждый из которых пытается отделить речь от мешающего фона, используя различные характеристики, связанные с речью, т.е. (1) спектральную форму, (2) спектрально-временные модуляции, (3) структуру периодичности из-за присутствия гармоник высоты тона, и (4) профиль долгосрочной спектральной изменчивости.

В этой работе [3] алгоритм подавления шума для VAD и вычисления энергии шума реализован в цифровом контроллере сигналов для улучшения фонового белого шума, машинного шума и шума лепета. В этом способе частотный спектр речевого сигнала дальнего конца разделяется на множество

неоднородных полос. Энергия в каждой полосе вычисляется с использованием негерметичного интегрирования. Для каждой из полос частот отношение сигнал/шум (SNR) вычисляется с использованием энергии полосы сигнала, вычисленной на этапе вычисления энергии полосы, и оценки энергии полосы шума. Энергии полос текущего кадра и энергии полос шума сравниваются для того, чтобы классифицировать текущий кадр либо как шумовой кадр, либо как речевой кадр. Экспериментальные результаты были получены, когда в качестве фонового шума использовались шум лепета, шум автомобилей, уличный шум и шум аэропорта.

А в следующей работе [4]описывается алгоритм VAD, основанный на усиленных глубоких нейронных сетях (bDNNs). Предлагаемый алгоритм сначала генерирует несколько базовых прогнозов для одного кадра только из одного - DeepNeuralNetwork (DNN), а затем агрегирует базовые прогнозы для лучшего прогнозирования кадра. Кроме того, используется новая акустическая функция с несколькими разрешениями кохлеаграмма - Multi-Resolution Cochleagram (MRCG), которая объединяет признаки кохлеаграммы с несколькими спектрально-временными разрешениями и пофказывает превосходные результаты по разделения речи по многим акустическим признакам.

В нашей работе мы предполагаем, что наилучшим решением для создания эффективной системы VAD может быть только использование технологии искусственных нейронных сетей (ИНС). Это связано с тем, что проблема обнаружения речевой активности в звуковых данных является трудно-алгоритмизируемой задачей. Однако главная проблема использования ИНС заключается в том, что она предварительно должна хорошо обучиться к решению поставленных перед ней задач. Зачастую для обучения ИНС требуется довольно большой объем данных. Очевидно, что чем больше можно использовать обучающих данных, тем сложнее становится структура нейронной сети, и соответствующим образом повышается эффективность ее работы

Для того, чтобы попытаться ответить на выше поставленный вопрос, требуется провести определенные экспериментальные исследования. В данном исследовании, которые мы провели, главным образом ставилась задача определения зависимости эффективности работы нейронной сети от количества дикторов, используемых при обучении этой сети. Поскольку система обнаружения голосовой активности должна быть диктора-независимой, соответствующая нейронная сеть должна обучаться на массиве данных, включающих речевые примеры от множества дикторов. Таким образом, была поставлена задача по экспериментальному исследованию влияния количества дикторов, используемых в базе обучающих данных, на эффективность нейронной сети для детектирования речевой активности в звуковых данных.

Для осуществления экспериментального исследования согласно поставленной задаче, нами была использована модуль NNTools среды программирования MatLab. Там была спроектирована структура

многослойного, полносвязного персептрона. Были выбраны следующие параметры для данной нейронной сети:

- нейронная сеть имеет 36 входов и один выход;
- нейронная сеть состоит из трех скрытых слоев по 20 нейронов в каждом;
- 80% обучающих данных используется для тренировки сети и 20% для валидации;
- данные для обучения и валидации распределяются случайным образом;
- производительность сети (она же величина ошибки сети) определяется перекрестной энтропией (сама функция тренировки выбирает данный способ определения ошибки);
- функция активации: сигмоидная в скрытых слоях и в выходном многопеременная логистическая функция;
- количество итераций обучения сети (эпох) равно 1000;
- тренировочная функция: метод обратного распространения с использованием масштабируемых сопряженных градиентов (scaledconjugategradients);
- обучение сети производится в параллельном режиме на центральном процессоре, где количество использованных потоков равно 4.

В качестве обучающих данных использованы данные со следующими характеристиками:

- общее количество дикторов дикторов 190;
- в тесте участвовало 16 дикторов;
- объем данных для тестирования сети не меняется в течении всего эксперимента;
- дикторы, участвующие в тестировании, не входят в обучающую выборку;
- при обучении на каждой последующей итерации добавляются новые дикторы к уже существующим.

Таким образом, сеть каждый раз обучается заново на увеличивающемся объеме данных. Приэто на каждом шаге добавляются данные по новым 10-и дикторам. Примеры речевых сигналов от каждого диктора хранятся в отдельных аудиофайлах формата WAV. Помимо этого, к каждому аудиофайлу прилагается Excel – файл. Разделение участков звуковых данных аудиофайла на речевые и неречевые участки выполняется вручную посредством прослушивания этих аудиофайлов. После того, как были проведены эти подготовительные работы, была осуществлены обработка аудиофайлов в следующей последовательности:

1. считывается аудиофайл и считывается excel-файл данного аудиофайла с метками границ и соответствующими этим границам метками «не речь» или «речь» (0 или 1);
1. производится нормализация внутри каждого блока (не зависимо от метки речь\не речь), при которой вначале вычитается

среднее значение по блоку, а затем делится на максимальное значение блока;

2. внутри каждого блока вычисляются 36 мел частотных кепстральных коэффициентов на каждом участке сигнала длительностью 20 миллисекунд;

3. формируется вектор целей (targets) для каждого блока в соответствии с его меткой «речь» или «не речь», длина которого равна количеству фреймов в соответствующем блоке сигнала.

4. после того как собрали данные от 10-ти дикторов, отправляем их на вход нейронной сети (inputs) и цели (targets) нейронной сети.

После того, как были подготовлены обучающие данные, проводилось обучение сети. А эффективность обученной сети определяем через вычисления ее ошибки. Очевидно, что чем выше величина ошибки сети, тем ниже ее эффективность. Ошибка сети определялась по следующему алгоритму:

1. на вход обученной сети подаются данные для тестирования и вычисляется выход сети (outputs);

1. в нашем случае выход сети может принимать любые значения от 0 до 1;

2. далее в зависимости от порога мы присваиваем каждому значению выхода либо 0 (если меньше порога), либо 1 (если больше порога);

3. ошибка в таком случае равна: error = targets-outputs, и таким образом формируется вектор ошибок.

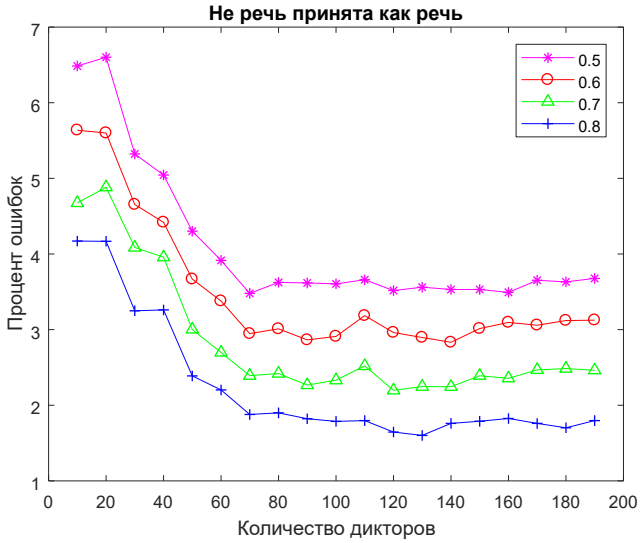

Рисунок 1 – графики зависимости ошибок первого рода от количества дикторов. Синий – для порога 0.8, Зеленый – для порога 0.7, Красный – для порога 0.6 и Фиолетовый – для порога 0.5

На основе полученных результатов можно сделать вывод о том, что при увеличении объема обучающей выборки минимально возможное

значение ошибки обобщения уменьшается, причем характер зависимости ошибки обобщения от мощности обучающей выборки совпадает с теоретическими предположениями. Однако при этом увеличиваются временные затраты на обучение нейронной сети, а также увеличивается отклонение ошибки обобщения от установившегося значения. Таким образом, была реализована модель нейронной сети для решения задачи детектирования речи в зависимости от количества дикторов. При анализе поведения нейронной сети было определено, что существующего размера обучающей выборки недостаточно для достижения нулевой ошибки обобщения нейронной сети. Однако, построенная сеть показала способность к обучению, подтвержденную экспериментальными данными.

Список используемой литературы

1. Gökay Dişken, Zekeriya Tüfekc, Ulus Çevik. A robust polynomial regression-based voice activity detector for speaker verification // J AUDIO SPEECH MUSIC PROC. 2017**,** 23 (2017).
2. Maarten Van Segbroeck, Andreas Tsiartas and Shrikanth S. Narayanan. A robust frontend for VAD: exploiting contextual, discriminative and spectral cues of human voice// Academic Press; 1st edition. 2013.
3. Charu Singh, Maarten Venter, Rajesh Kumar Muth, David Brown. A Real-Time DSP-Based System for Voice Activity Detection and Background Noise Reduction // Intelligent Speech Signal Processing, Chapter 3, Pages 39-54. 2019.
4. Xiao-Lei Zhang, DeLiang Wang. Boosted Deep Neural Networks and Multi-resolution Cochleagram Features for Voice Activity Detection // Proc. Interspeech, Pages 1534-1538. 2014.