УДК 004.93

СИСТЕМА КЛАССИФИКАЦИИ ЗВУКОВ ОКРУЖАЮЩЕЙ СРЕДЫ

И.Н. ЖУК

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь

Поступила в редакцию 19 января 2018

Аннотация. В статье представлена система классификации звуков окружающей среды и результаты сравнения производительности с другими системами на звуковой базе ESC 10. В представленной системе формирование признаков звукового сигнала осуществляется с помощью модели внутреннего уха и импульсов слухового нерва. Классификация звуков осуществляется с помощью различных конфигураций сверточных нейронных сетей. Доля правильных ответов классификации значительно выше результатов оригинальной статьи звуковой базы ESC 10.

Ключевые слова: внутреннее ухо, формирование признаков, классификация звуков, сверточные нейронные сети, ESC 10.

Abstract. This paper presents environmental sound classification system and performance comparison on ESC 10 dataset. The feature extraction method includes cochlea and auditory nerve models. Classification model includes classic convolutional neuron network architectures. Experiments based on different architectures of convolutional neural networks and proposed feature extraction method. The model outperforms baseline implementations and achieves results comparable to other state-of-the-art approaches.

Keywords: cochlea, auditory nerve spikes, feature extraction, sound classification, convolutional neuron networks, ESC 10.

Doklady BGUIR. 2018, Vol. 113, No. 3, pp. 54-58
Environmental sound classification system
I.N. Zhuk

Введение

Для наиболее эффективной работы со звуком в системах классификации и автоматического распознавания речи используются различные способы формирования признаков звукового сигнала. Сложность задачи заключается в том, что нам неизвестно, какие признаки наиболее информативные и репрезентативные. Помимо этого, не существует универсальных решений, поэтому в настоящее время в большинстве случаев используют несколько различных способов формирования признаков звукового сигнала.

Благодаря значительному росту производительности вычислительной техники и росту эффективности нейронных сетей появились системы, работающие напрямую с временной формой звукового сигнала [2]. Однако различные способы формирования признаков позволяют снизить требования к вычислительным ресурсам и получить более высокую производительность системы. Среди способов формирования признаков звукового сигнала наибольшей популярностью пользуются мел-кепстральные коэффициенты (МКК), способы, использующие взаимную информацию [3], *i*-векторы [4] и спектрограмма на основе преобразования Фурье [5]. Характеристические векторы вышеперечисленных признаков достаточно сильно отличаются. МКК подразумевает частотный анализ и сокращение значительной части входной информации, *i*-векторы представляют собой стохастический процесс сжатия информации с помощью линейной Гауссовской модели, спектрограмма, напротив, представляет собой подробный частотно-временной сигнал. Использование спектрограммы стало возможным благодаря значительным успехам в распознавании изображений и сверточным нейронным сетям [5].

С другой стороны, значительные успехи присутствуют в области моделирования физиологии и нервной системы [6, 7]. Большинство систем обработки и распознавания речи используют только частичное моделирование работы внутреннего уха. Наибольшей популярностью пользуются гамма-тоновые фильтры, моделирующие работу улитки. С помощью банка гамма-тоновых фильтров строятся кохлеаграммы и рассчитываются гамма-тоновые кепстральные коэффициенты [8]. Существует не так много исследований, посвященных системам, которые используют в качестве входных данных непосредственно нервные импульсы [9]. Их можно интерпретировать несколькими способами: можно использовать как бинарный вектор [9], использовать не сами импульсы, а частоту их появления [10] или преобразовать в изображение.

Целью этой работы является разработка системы классификации, которая использует непосредственно нервные импульсы в качестве характеристического вектора звукового сигнала, а также сравнение результата классификации данной системы с результатами других систем. Сравнение проводилось на звуковой базе Environmental Sound Classification 10 (ESC 10) [12]. Производительность сравнивалась со следующими системами: 1) классификатор – «случайный лес [15]», входные данные – мел-кепстральные коэффициенты [12]; 2) классификатор на основе многослойного персептрона, входные данные – мел-кепстральные коэффициенты [13]. Эксперименты проводились на звуковой базе ESC 10 с помощью метода перекрестной валидации с 10 блоками. Доля правильных ответов классификации составила 79,2 %, тем самым подтвердив эффективность данной системы.

Способ формирования признаков

Здесь способ формирования признаков описывается в краткой форме, более подробное описание используемых моделей можно найти в [1, 7, 10]. Для извлечения признаков из звукового сигнала используются модели внутреннего уха, волосковых клеток и аудиторного нерва. Человек воспринимает звук в диапазоне частот от 20 Гц до 20 кГц. В ушной улитке находится порядка 3500 внутренних волосковых клеток, которые отвечают за преобразование механических колебаний в электрический ток. Однако ввиду ограниченности частотного диапазона звуковых файлов и вычислительных ресурсов диапазон был сужен до 5,5 кГц, а количество волосковых клеток, используемых в модели, снижено до 200.

Процесс обработки информации можно представить следующим образом. Сначала звуковой сигнал поступает на банк гамма-тоновых фильтров, состоящий из 200 фильтров, которые занимают область частот от 20 Гц до 5,5 кГц. Затем выход каждого фильтра подвергается выпрямлению и компрессии $\sqrt[3]{x}$. После этого каждый сигнал обрабатывается моделью аудиторного нерва, которая реализована с помощью метода «интегрировать и сработать» с утечками [11].

После этого на каждой временной диаграмме фиксируются моменты времени, в которые потенциал нейрона превышает порог срабатывания и испускается импульс. Затем все импульсы для каждого сигнала объединяются в общую матрицу. Элементы матрицы могут принимать значения либо 0 – импульса нет, либо 1 – импульс есть. Способ формирования признаков схематично представлен на рис. 1.

В центре рис. 1 представлена временная диаграмма потенциала одного нейрона. Также представлены параметры: потенциал срабатывания нейрона, по достижению которого нейрон испускает импульс и снижает потенциала до потенциал сброса. Параметр «рефрактерный период» характеризует период после импульса, в течение которого нейрон восстанавливает потенциал. Обычно $v(t)$ равняется потенциалу сброса в конце периода или на всем промежутке, а механизм утечки начинает работу после этого периода. Диапазон параметров достаточно широк, однако, руководствуясь известными биологическими параметрами, были выбраны следующие значения: потенциал срабатывания 0,7 мВ, потенциал сброса 0 В, рефрактерный период 5 мс, постоянная времени мембраны 2 мс. Эти параметры хорошо подходят для данной модели с точки зрения параметров сохранения информативности и фильтрации шума. Помимо этого, данные параметры близки к естественным биологическим. Это означает, что временная диаграмма нервных импульсов достаточно точно (хотя и не полностью) моделирует сигналы аудиторного нерва.

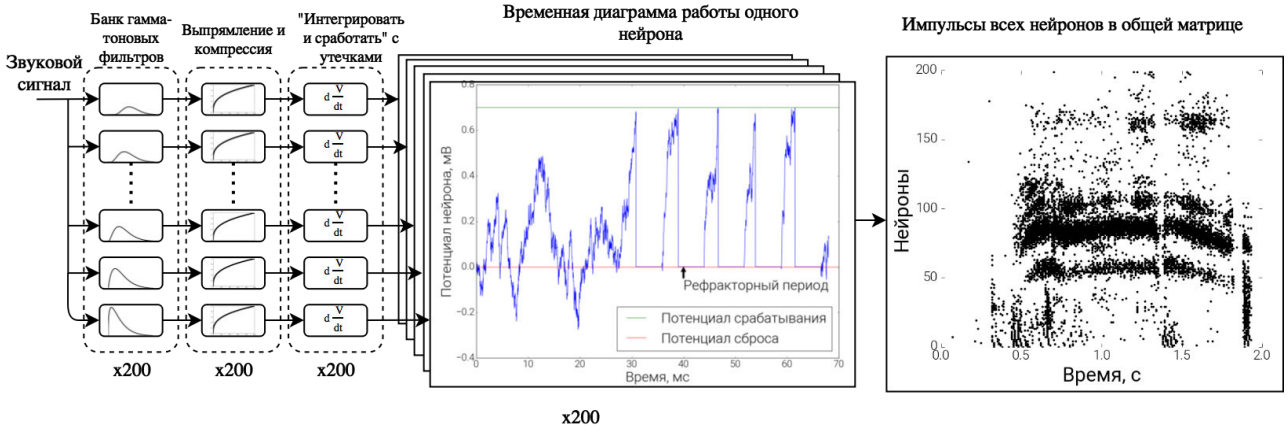

Рис. 1. Схематичное изображение способа формирования признаков

Вектор признаков представляет из себя 200 значений (1 значение для каждого нейрона) и соответствует временному отрезку 10 мс. Совокупность таких векторов для звукового сигнала длительностью 5 с является матрицей размером 500×200 и представляет собой диаграмму нервных импульсов. На рис. 2 представлены спектрограмма и диаграмма нервных импульсов для одного звукового файла, содержащего плач ребенка, из набора данных ESC 10 [12]. Частотный диапазон в обоих случаях составляет 5,5 кГц, а временной диапазон – 5 с. Однако частотные шкалы разные. При построении банка гамма-тоновых фильтров используется ERB частотная шкала, в то время как в спектрограмме используется линейная шкала. ERB шкала аппроксимирует особенность внутреннего уха нелинейно распределять частотные полосы. Эта особенность очень полезна в обработке речи.

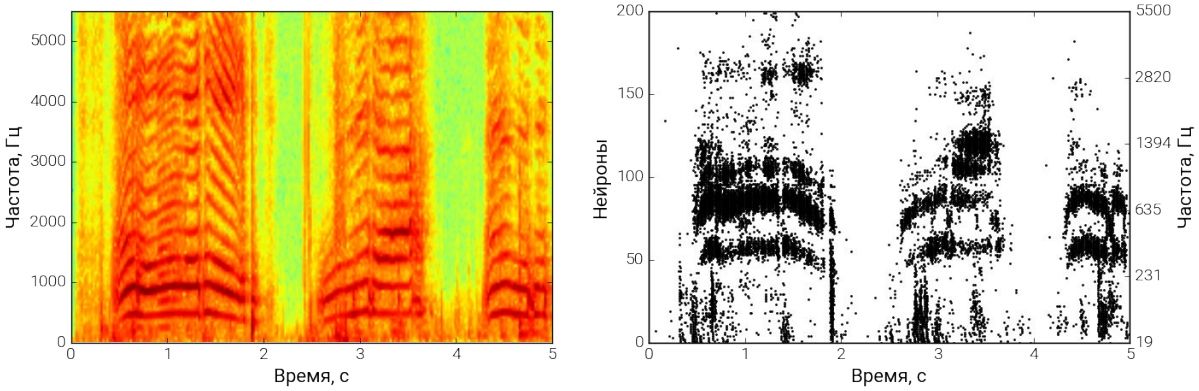

Рис. 2. Спектрограмма и диаграмма нервных импульсов звукового сигнала

Экспериментальные результаты

Для экспериментов был выбран набор данных ESC 10, разработанный для классификации звуков окружающей среды. Он включает в себя 10 классов, в каждом из которых 40 записей. В классы входят: собачий лай, шум дождя, морские волны, плач ребенка, тиканье часов, чихание, вертолет, бензопила, кукареканье петуха, потрескивание углей в огне.

Все записи – примерно одинаковой длительности 4–7 с, у большинства записей длина 5 с. Для простоты обработки файлы длительностью более 5 с обрезались после 5 с. Файлы длительностью до 5 с дополнялись нулевыми столбцами. В итоге входные данные представляют собой 10 классов по 40 объектов классификации, где каждый объект представлен матрицей размером 500×200, в которой строки характеризуют номера нейронов, а столбцы – ось времени. Каждый член матрицы принимает значения либо 0 (импульса нет), либо 1 (импульс есть).

Архитектура сети классификатора содержит в себе 3 блока 2D сверточной сети, функции активации Rectified Linear Unit (ReLU), слои MaxPool, Dropout регуляризацию и полносвязные слои. В качестве функции ошибки используется перекрестная энтропия. Оптимизация весовых коэффициентов сети выполняется с помощью алгоритма adadelta [14].

Используемая метрика оценки производительности – доля правильных ответов. Детальное описание архитектуры представлено в табл. 1.

Таблица 1. **Архитектура сети классификатора**

№	Тип слоя	Детали
1	2D свертка	Количество фильтров 16, размер ядра 3×3, шаг 2, функция активации «ReLU»
2	Maxpool 2D	Размер области 2×2
2	2D свертка	Количество фильтров 32, размер ядра 3×3, шаг 2, функция активации «ReLU»
3	Maxpool 2D	Размер области 2×2
	2D свертка	Количество фильтров 64, размер ядра 3×3, шаг 2, функция активации «ReLU»
4	Maxpool 2D	Размер области 2×2
5	Полносвязный слой	Размер 256 нейронов, функция активации «ReLU»
6	Полносвязный слой	Размер 256 нейронов, функция активации «ReLU»
7	Dropout	Вероятность пропуска 0.5
8	Полносвязный слой	Размер по количеству классов (10), функция активации «softmax»

Эксперименты проводились по схеме перекрестной валидации с 10 блоками. В сравнении представлены: результаты классификатора из оригинальной статьи [12] (случайный лес [15] с мел-кепстральными коэффициентами), результаты классификации многослойного персептрона с мел-кепстральными коэффициентами [13], результаты классификации сверточной нейронной сети с диаграммой нервных импульсов и измеренный результат у людей [12]. Сравнение результатов представлено в табл. 2.

Таблица 2. **Сравнение производительности систем классификации звуков окружающей среды**

№	Способ формирования характеристического вектора	Способ классификации	Доля правильных ответов
1	Мел-кепстральные коэффициенты	«Случайный лес»	73,7 %
2	Мел-кепстральные коэффициенты	Многослйный персептрон	74,5 %
3	Нервные импульсы	Сверточная нейронная сеть	79,2 %
4	–	Человек	95,0 %

Заключение

В этой статье рассмотрен способ использования модели нервных импульсов в качестве формирования признаков для звукового сигнала. Этот способ позволяет получить временные последовательности нервных импульсов слухового нерва. Эти последовательности обладают интересными свойствами. Во-первых, они обладают нелинейным частотным разрешением, что позволяет более детально представить наиболее интересные частотные полосы средних и низких частот. Во-вторых, нервные импульсы имеют бинарное представление. Это позволяет избавиться от нормализации входных данных, которая имеет критическое значение в случае использование других методов формирования признаков. Также бинарное представление является более оптимальным с точки зрения обучения нейронной сети. Благодаря этим особенностям рассмотренный способ формирования признаков доказал свою эффективность в задаче классификации звуковых сигналов. Из минусов следует выделить высокую (относительно других методов) вычислительную сложность. В будущем планируется провести сравнение с другими популярными методами формирования признаков и провести поиск оптимальных параметров модели.

Список литературы / References

1. Brian hears: online auditory processing using vectorization over channels / B. Fontaine [et al.] // Front. Neuroinform. 5:9. 2011. doi: 10.3389/fninf.2011.00009.
2. Palaz D., Magimai M. Convolutional Neural Networks-based Continuous Speech Recognition using Raw Speech Signal. Doss, Ronan Collobert. Idiap-RR-18-2014.
3. Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI/ D. Povey [et al.] // Proc. Interspeech. 2016. P. 2751–2755.
4. Speaker adaptation of neural network acoustic models using i-vectors / G. Saon [et al.] // in ASRU. 2013. P. 55–59.
5. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin / Dario A. [et al.] // arXiv:1512.02595 [cs.CL]. December 2015.
6. Goodman D.F., Brette R. The Brian simulator // Front. Neurosci. 3,2:192-197. doi: 10.3389/neuro.01.026.2009.

7. Equation-oriented specification of neural models for simulations / Stimberg M. [et al.] // Frontiers Neuroinf. 2014. doi:10.3389/fninf.2014.00006.
8. An auditory-based feature for robust speech recognition / Y. Shao [et al.] // Acoustics, Speech and Signal Processing. April 2009. P. 4625–4628.
9. Automatic Speech Recognition with Neural Spike Trains / M.H. Holmberg [et al.] // Interspeech. Lisbon, Portugal, September 4–8, 2006.
10. Ivanov A.V., Likhachov D.S., Petrovsky A.A. Spiking neuron auditory model for speech processing systems // 9th International Workshop on Systems, Signals and Image Processing IWSSIP. Manchester, United Kingdom, 2002.
11. Gerstner W., Kistler W. Spiking Neuron Models: Single Neurons, Populations, Plasticity. Cambridge University Press, 2002.
12. Piczak K.J. ESC: Dataset for Environmental Sound Classification // Proceedings of the 23rd ACM international conference on Multimedia. 2015. P. 1015–1018.
13. A real-time environmental sound recognition system for the Android OS / Pillos A. [et al.] // Detection and Classification of Acoustic Scenes and Events. 2016.
14. Matthew D.Z. ADADELTA: An Adaptive Learning Rate Method. arXiv:1212.5701v1 [cs.LG]. December 2012.
15. Breiman L. Machine Learning // Kluwer Academic Publishers, 45: 5. 2001. doi.org/10.1023/A:1010933404324.

Сведения об авторе

Жук И.Н., аспирант кафедры электронных вычислительных средств Белорусского государственного университета информатики и радиоэлектроники.

Information about the author

Zhuk I.N., PG student of computer-engineering department of Belarusian state university of informatics and radioelectronics.

Адрес для корреспонденции

220013, Республика Беларусь,
г. Минск, ул П. Бровки 6,
Белорусский государственный университет информатики и радиоэлектроники
тел. +375-44-532-15-17;
e-mail: ivan.nikolaevich.zhuk@gmail.com
Жук Иван Николаевич

Address for correspondence

220013, Republic of Belarus,
Minsk, P. Brovka st., 6,
Belarusian state university
of informatics and radioelectronics
tel. +375-44-532-15-17;
e-mail: ivan.nikolaevich.zhuk@gmail.com
Zhuk Ivan Nikolaevich