УДК 004.93'12 DOI: 10.18698/0236-3933-2019-1-35-49

ДЕТЕКТИРОВАНИЕ НЕИЗВЕСТНЫХ ЗВУКОВ ДЛЯ ЛЮДЕЙ С НАРУШЕННЫМ СЛУХОМ НА ОСНОВЕ ВАРИАЦИОННОГО АВТОЭНКОДЕРА

А.Х. Сарафасланян[1] artashes@bizb.ru
В.В. Чепраков[1] cpb@bizb.ru
Д.А. Суворов[1] suvorov@bizb.ru
М.В. Мозговой[2] mozgovoy@bmstu.ru
А.В. Волков[2] avv@bmstu.ru

[1] ООО «Бизнес Бюро», Москва, Российская Федерация
[2] МГТУ им. Н.Э. Баумана, Москва, Российская Федерация

Аннотация

Представлена система детектирования неизвестных звуков для людей с нарушенным слухом, разработанная на основе вариационного автоэнкодера. Проиллюстрирована архитектура созданного вариационного автоэнкодера, энкодерная и декодерная части которого состоят из полносвязных слоев. Описан процесс создания базы данных для обучения системы, приведено разбиение данной базы на блоки для обучения, тестирования и детектирования неизвестных звуков. Описана методика обучения системы и ее математическая основа, включающая в себя метод стохастической оптимизации *Adam* и вариационный нижний предел в качестве функции потерь. Проведено тестирование разработанной системы, установлено полное отсутствие ложно-отрицательных результатов детектирования неизвестных звуков и вероятность ложноположительного результата 14 %, что вполне приемлемо для ее практического использования. Приведены технологии, использовавшиеся для реализации системы, а также устройство, в которое система должна быть интегрирована. Рассмотрены дальнейшие возможности для улучшения системы

Ключевые слова

Вариационный автоэнкодер, глубокое обучение, распознавание звуков, цифровая обработка сигнала, детектирование, обучение

Поступила 24.07.2018

Работа выполнена при поддержке Фонда содействия инновациям (грант № 168ГРНТИС5/35848)

Введение. Невозможность слышать окружающие звуки людьми с нарушенным слухом может значительно снизить верность принятия правильного решения о дальнейших действиях [1]. В настоящее время известны методы классификации окружающих звуков с помощью сверточных нейронных сетей [2], машин опорных векторов [3], вычисления евклидова расстояния в пространстве высокой размерности [4], моделирования звуков с помощью смеси гауссиан [5]. Перечисленные методы подразумевают, что в процессе обучения им будут предъявлены все типы звуков, которые представляют интерес для пользователя. Однако людям с нарушенным слухом также необходимо реагировать и на не известные заранее звуки, например для распознавания аварийных ситуаций. Для этого используют методы детектирования аномалий в звуковых сигналах, например, одноклассовая машина опорных векторов [6], смесь гауссиан [7], глубокая автокодирующая смесь гауссиан [8] или автоэнкодер [9]. Эти методы допускают возможность недетектирования важного неизвестного звука, что может быть очень опасно в экстремальных ситуациях для людей с нарушенным слухом. Детектирование аномалий также возможно с помощью многоколоночной классификации, когда обучают нейросетевой классификатор различать необходимые классы звуковых сигналов. В случае если он не может определиться с принадлежностью звука к одному определенному классу, полагают, что звук аномальный, однако рассматриваемый звуковой сигнал может быть в действительности близок к двум классам, т. е. аномальным не являться, но при этом такая система определит его как аномальный.

Рис. 1. Прототип комплекса «СОМСИ», созданный с помощью 3D-принтера

Цель работы — создание системы детектирования потенциально важных для человека с нарушенным слухом звуков, которые могут прозвучать внутри жилого помещения или на улице. Система реализована с помощью методов глубокого машинного обучения с использованием вариационного автоэнкодера [10]. После обучения и тестирования система реализована на языке программирования *Python* в виде приложения под операционную систему (ОС) *Linux*. Система может функционировать в режиме реального времени. Разработанное приложение предназначено для интеграции в программно-аппаратный комплекс «Система обработки

мультиканальной сенсорной информации для лиц с нарушенным слухом и зрением (СОМСИ)» (рис. 1), который будет носиться с собой. Взаимодействие с комплексом осуществляется через тактильный интерфейс. Программно-аппаратный комплекс оснащен цифровым массивом микрофонов с *PDM*-интерфейсом [11], через который и будет захватываться звук для анализа.

Далее описан массив данных, использованный для обучения, архитектура разработанной нейронной сети, приведены процесс обучения системы и результаты ее тестирования.

Данные для обучения. Для разработки и проверки алгоритмов решения поставленной задачи использован существующий размеченный массив данных *UrbanSound8k* [12]. Массив данных содержит 8732 звуковые дорожки длительностью до 4 с. Звуки записаны в естественной среде, содержат инструментальный шум микрофонов, фоновые посторонние шумы и реверберации. Записи выполнены с разных моделей микрофонов с различными настройками чувствительности. Звуковые файлы разбиты на 10 классов:

1) air_conditioner (звук кондиционера);
2) car_horn (сигнал автомобиля);
3) children_playing (играющие дети);
4) dog_bark (собачий лай);
5) drilling (сверление);
6) enginge_idling (звук мотора);
7) gun_shot (выстрел);
8) jackhammer (звук перфоратора);
9) siren (сирена);
10) street_music (уличная музыка).

Массив данных обладает относительно небольшим размером, но он достаточен для проверки работоспособности алгоритмов. Обучение системы на реальном массиве данных, содержащем значительно большее количество звуков, не потребует изменения алгоритмов, необходимо лишь значительно увеличить вычислительную мощность в процессе обучения.

Звуковые дорожки длительностью менее 1 с исключались. Частота дискретизации всех звуковых файлов 16 кГц. Многоканальный звук сведен к одноканальному. Разрешение звуковых измерений 16 бит. Состав результирующего массива данных приведен на рис. 2. Классы gun_shot (выстрел) и siren (сирена) не предъявлялись системе в процессе обучения — на них тестировалась способность системы обнаруживать наличие аномальных звуков.

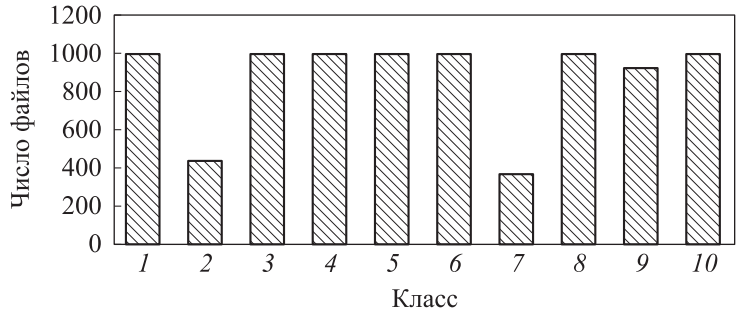

Рис. 2. Состав массива данных, использованного для разработки и тестирования решения:

1 — air_conditioner (звук кондиционера); *2* — car_horn (сигнал автомобиля); *3* — children_playing (играющие дети); *4* — dog_bark (собачий лай); *5* — drilling (сверление); *6* — enginge_idling (звук мотора); *7* — gun_shot (выстрел); *8* — jackhammer (звук перфоратора); *9* — siren (сирена); *10* — street_music (уличная музыка)

Обучение вариационного автоэнкодера. Весь одномерный звуковой сигнал был преобразован в двумерное представление путем вычисления его спектрограммы (рис. 3), так как анализ звуков с помощью глубоких нейронных сетей в частотной области обычно дает лучшие результаты [13]. Спектрограмма $X(t, \omega)$ вычисляется как квадрат модуля быстрого

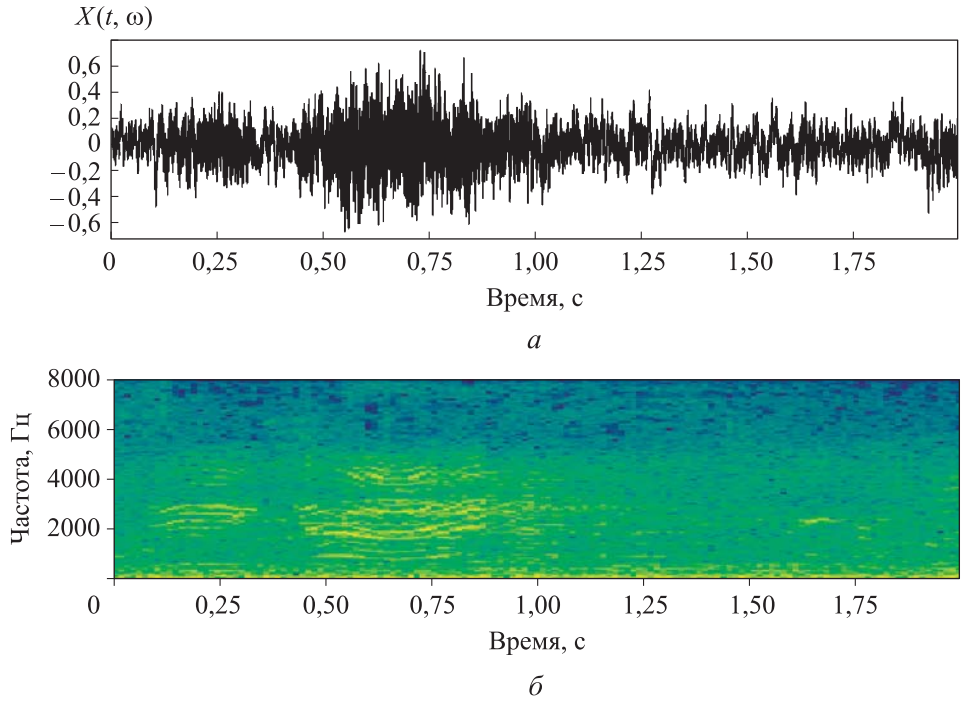

Рис. 3. Звуковой сигнал (*а*) и его представление в виде спектрограммы (*б*)

преобразования Фурье неперекрывающихся областей звукового сигнала $x(t)$ [14]:

$$X(t,\omega)=\left|FFT(x(t))\right|^2.$$

Вариационный автоэнкодер — нейронная сеть, условно состоящая из энкодера $q_\Phi(\mathbf{z}|X)$ и декодера $p_\Theta(X|\mathbf{z})$ (рис. 4, *а*). Задача энкодера $q_\Phi(\mathbf{z}|X)$ генерировать сжатое представление в виде вектора случайных величин $\mathbf{z}$ входной спектрограммы X, а задача декодера $p_\Theta(X|\mathbf{z})$ максимально точно восстанавливать из сжатого представления $\mathbf{z}$ исходную спектрограмму X. В общем случае восстановить X из $\mathbf{z}$ невозможно, поэтому результат восстановления $\tilde{X}$ лишь с некоторой точностью соответствует спектрограмме X. Погрешность реконструирования можно вычислить как квадрат евклидова расстояния от оригинальной спектрограммы до ее восстановленного образа:

$$e\left(X,\tilde{X}\right)=\sum_i\sum_j\left(X_{ij}-\tilde{X}_{ij}\right)^2.$$

Полная архитектура вариационного автоэнкодера приведена на рис. 4, *б*. Энкодер реализован с помощью трех полносвязных слоев с нелинейностями ReLU [15]. Первый полносвязный слой реализует первичное извлечение признаков. Полносвязный слой описывается формулой $\mathbf{y}=W\mathbf{x}+\mathbf{b}$, где $\mathbf{x}$, $\mathbf{y}$ — входной и выходной векторы; W — матрица весовых коэффициентов слоя; $\mathbf{b}$ — вектор смещений. Нелинейная функция ReLU имеет вид $f(x)=\max(0,x)$.

Два второго слоя энкодер обучается на генерирование значений математического ожидания и стандартного отклонения нормального распределения вектора случайных величин $\mathbf{z}$.

Декодер реализован по аналогичной схеме, но состоит из двух параллельных одинаковых ветвей, результаты декодирования которых складываются с равными весовыми коэффициентами. Такая параллельность повышает точность декодирования [10]. При обучении весовые коэффициенты обеих ветвей всегда остаются равными.

Вариационный автоэнкодер не предназначен для детектирования аномальных сигналов, однако при реконструировании образа, который значительно отличается от образов, предъявляемых ему в процессе обучения, будет возникать высокая погрешность реконструирования [16]. Если погрешность реконструирования превышает некоторое пороговое значение, то системе предъявлен аномальный сигнал. Данный подход работает при-

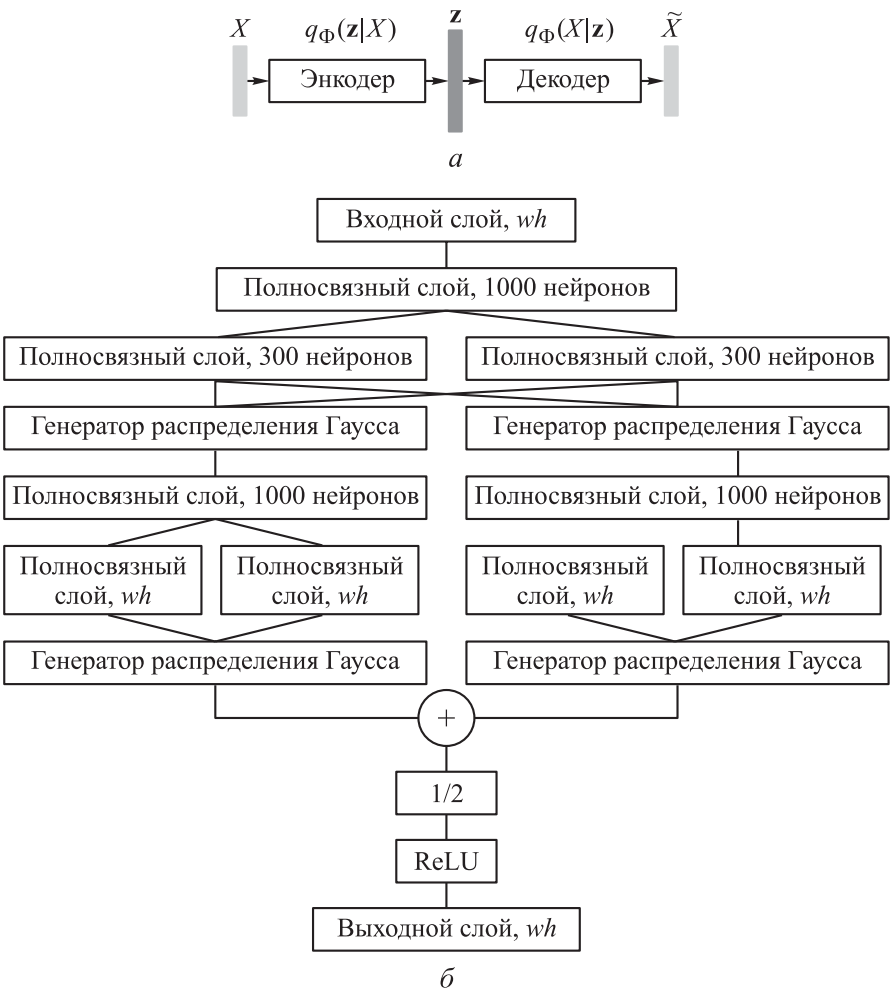

Рис. 4. Общая схема (*а*) и полная архитектура (*б*) вариационного автоэнкодера, использованного для детектирования неизвестных звуков

менительно к статистическим данным о поведении пользователей *web*-приложений [16]. Это позволяет предположить, что он может работать на звуковых сигналах, несмотря на другую их физическую природу.

Обучающий массив данных разбит на две части: 1) 90 % образцов предъявлялись для обучения; 2) для 10 % образцов вычислялось значение функции потерь в целях определения момента начала переобучения.

Обучение осуществлялось с помощью метода *Adam* [17]:

$$w[t+1] = w[t] - \alpha \frac{1}{\sqrt{g[t+1]+\varepsilon}} v[t+1];$$

$$g[t+1] = \mu g[t] + (1-\mu)\nabla(L, w[t])\nabla(L, w[t]);$$

$$v[t+1]=\beta v[t]+(1-\beta)\nabla(L,w[t]),$$

где t — номер итерации; L — функция потерь; w — набор обучаемых параметров нейронной сети; ε, μ, β — параметры алгоритма.

В качестве функции потерь использовался вариационный нижний предел [18]:

$$L=-D_{KL}\left(q_\Phi(\mathbf{z}\,|\,X)\,||\,p_\Theta(\mathbf{z})\right)+\log\left(p_\Theta(X\,|\,\mathbf{z})\right),$$

где D_{KL} — расстояние Кульбака — Лейблера [19],

$$D_{KL}=\frac{1}{2}\sum_{i=1}^{\dim Z}\left(1+\log\left(\sigma_i^2\right)-\mu_i^2-\sigma_i^2\right);$$

$\log(p_\Theta(X\,|\,\mathbf{z}))$ — логарифмическое правдоподобие выхода,

$$\log\left(p_\Theta(X\,|\,\mathbf{z})\right)=\sum_{i=1}^{\dim X}\log\left(\frac{1}{\sigma_i\sqrt{2\pi}}e^{-\frac{(\mu_i-X_i)^2}{2\sigma_i^2}}\right).$$

Обучение проводилось в течение 200 эпох на виртуальной машине с ОС *Ubuntu 16.04* с доступом к видеокарте *Nvidia Tesla M60*. Система реализована с помощью библиотек *Theano* и *Lasagne*. Процесс обучения показан на рис. 5. После 175 эпох началось переобучение, так как значение функции потерь на тестовой выборке перестало уменьшаться, а на обучающей — нет.

Тестирование точности детектирования. После окончания обучения вариационного автоэнкодера ему были предъявлены звуки нормальных и аномальных классов, чтобы вычислить пороговое значение погрешности реконструирования, после которого звук можно полагать аномальным. Гистограмма распределения погрешности реконструирования приведена на рис. 6. Согласно полученной гистограмме, пороговое значение составило $5{,}7788 \cdot 10^{-7}$.

Проведено тестирование точности детектирования аномальных звуков. Система со 100 % вероятностью определяет незнакомые звуки, но с 14 % вероятностью может принять знакомый звук за неизвестный.

Для анализа результатов обучения вариационного автоэнкодера вычислены результаты кодирования только с помощью энкодера $q_\Phi(\mathbf{z}\,|\,X)$ спектрограмм X. Над полученными векторами проведена операция снижения размерности *t-SNE* [20] до трехмерного пространства. Результаты *t-SNE*-визуализации приведены на рис. 7. Вариационный автоэнкодер не научился выделять четкие кластеры известных и неизвестных звуков, однако это не мешает детектированию аномальных звуков по погрешности реконструирования.

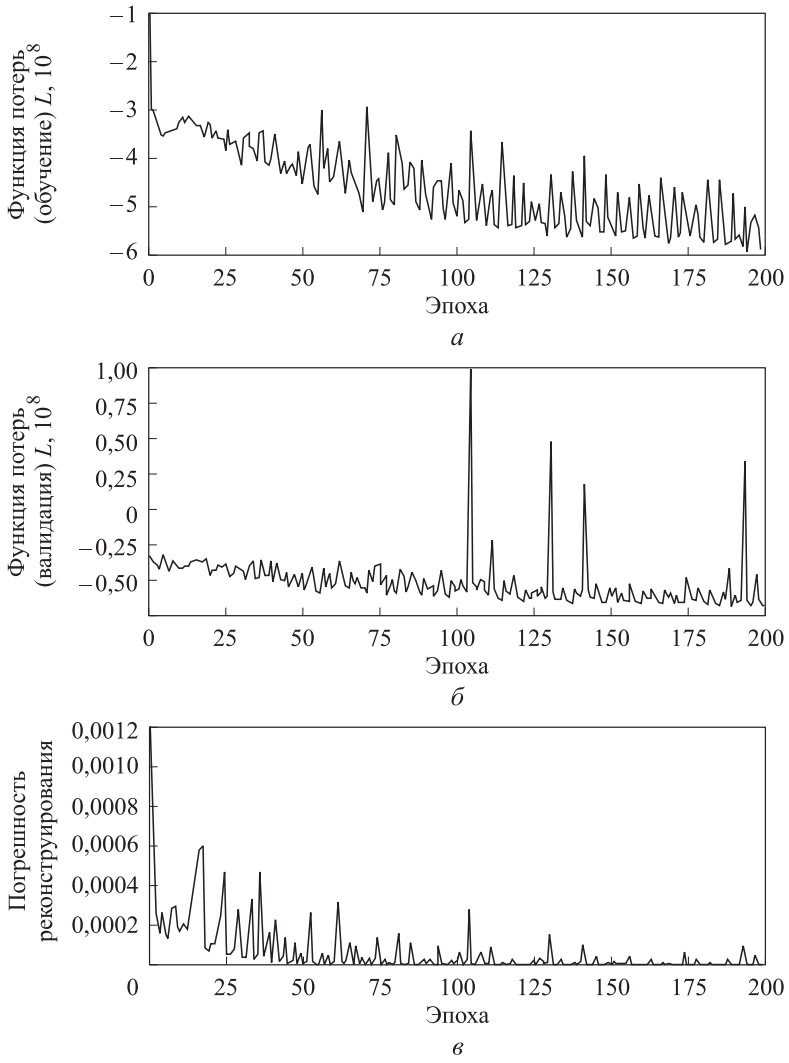

Рис. 5. Процесс обучения вариационного автоэнкодера:
а — функция потерь на обучающей выборке; *б* — функция потерь на тестовой выборке; *в* — погрешность реконструирования на тестовой выборке

Рис. 6. Распределение погрешности реконструирования нормальных (*1*) и аномальных (*2*) звуков

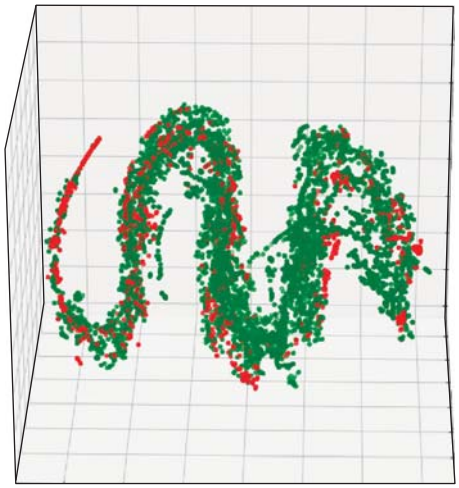

Рис. 7. *t-SNE*-визуализация результатов кодирования спектрограмм X энкодерной частью $q_{\Phi}(\mathbf{z}|X)$:

● — аномальные звуки; ● — обычные звуки

Применение разработанной системы. После окончания разработки и обучения системы она была интегрирована в «СОМСИ». Программная архитектура полученного решения для детектирования аномальных звуков приведена на рис. 8. Программное обеспечение работает под управлением ОС *Linux* (дистрибутив *Ubuntu FriendlyCore*) на четырехядерном процессоре *Samsung S5P4418* архитектуры *ARM Cortex-A9* с загрузкой всех четырех ядер приблизительно на 80 %. Входящий звуковой сигнал обрабатывается кадрами длительностью 1 с, перекрытие между кадрами 200 мс. Длительность обработки одного кадра около 500 мс. Таким образом, максимальная задержка детектирования аномального звука составляет не более 1 с, но так как для исключения ложных срабатываний программное обеспечение сигнализирует о наличии аномального звука только после трех срабатываний подряд детектора аномальных звуков, фактически задержка детектирования равна 2600 мс.

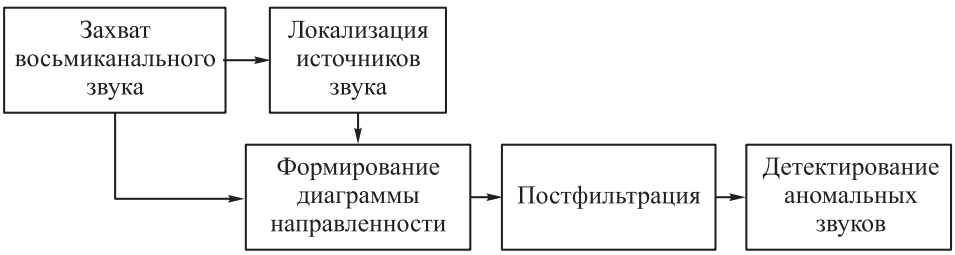

Рис. 8. Программная архитектура подсистемы детектирования аномальных звуков в «СОМСИ»

Восьмиканальный звук с массива микрофонов проходит несколько этапов предобработки и превращается в одноканальный звук перед тем, как поступить в систему детектирования аномальных звуков. Такая архитектура выбрана для снижения влияния фоновых шумов, реверберации и инструментальных шумов микрофонов на качество детектирования [21]. На этапе локализации источников звуков с помощью глубоких сверточных нейронных сетей определяются направления на активные источники звука во входящем многоканальном звуковом сигнале [22]. Этап формирования диаграммы направленности сводит с помощью лучеформирователя Кейпона многоканальный звук к одноканальному, одновременно подавляя все шумы вне целевого направления, определенного на этапе локализации источников звука [23]. Постфильтрация с помощью постфильтра Зелинского дополнительно снижает уровень фоновых и инструментальных шумов [24].

Заключение. Разработана и успешно протестирована система детектирования аномальных звуков. Система основана на вариационном автоэнкодере и технологиях глубокого обучения, поэтому не привязана к конкретным звукам, на которых обучалась и тестировалась. В дальнейшем система может быть обучена на массиве данных большего объема, что значительно повысит точность ее работы. Большую практическую ценность имеет тот факт, что система была реализована в виде приложения на языке *Python* под ОС *Linux*, которое способно обрабатывать звуковой сигнал в реальном времени. Это означает, что уже в настоящее время ее можно использовать в разрабатываемых носимых устройствах для людей с нарушенным слухом. Далее для повышения точности алгоритмическим способом в состав энкодера и декодера могут быть внедрены сверточные слои.

ЛИТЕРАТУРА

[1] Hersh M. Deafblind people, communication, independence, and isolation. *J. Deaf. Stud. Deaf. Educ*., 2013, vol. 18, iss. 4, pp. 446–463. DOI: 10.1093/deafed/ent022

[2] Sainath T.N., Parada C. Convolutional neural networks for small-footprint keyword spotting. *INTERSPEECH*, 2015, pp. 1478–1482.

[3] Tzanetakis G., Cook P. Musical genre classification of audio signals. *IEEE Trans. Speech Audio Process*., 2002, vol. 10, iss. 5, pp. 293–302. DOI: 10.1109/TSA.2002.800560

[4] Tavares T.F., Foleiss J.H. Automatic music genre classification in small and ethnic datasets. *Proc. 13th CMMR Int. Symp*., 2017, pp. 25–28.

[5] Bragg D., Huynh N., Ladner R.E. A personalizable mobile sound detector app design for deaf and hard-of-hearing users. *Proc. 18th Int. ACM SIGACCESS Conf. Computers Accessibility*, 2016, pp. 3–13. DOI: 10.1145/2982142.2982171

[6] Lecomte S., Lengellé R., Richard C., et al. Abnormal events detection using unsupervised one-class SVM — Application to audio surveillance and evaluation. *8th IEEE AVSS Int. Conf.*, 2011, pp. 124–129.

[7] Bishop C.M. Pattern recognition and machine learning. Springer, 2006.

[8] Zong B., Song Q., Min M.R., et al. Deep autoencoding Gaussian mixture model for unsupervised anomaly detection. *ICLR*, 2018.
URL: https://openreview.net/pdf?id=BJJLHbb0- (дата обращения: 09.07.2018).

[9] Oh D.Y., Yun I.D. Residual error based anomaly detection using auto-encoder in SMD machine sound. *Sensors*, 2018, vol. 18, no. 5, art. 1308. DOI: 10.3390/s18051308

[10] Kingma D.P., Welling M. Auto-encoding variational Bayes. ICLR, 2014.
URL: https://arxiv.org/pdf/1312.6114.pdf (дата обращения: 09.07.2018).

[11] Жуков Р.А., Суворов Д.А., Тетерюков Д.О. и др. Конструирование подсистемы ввода сигнала на основе массива микрофонов с цифровым интерфейсом. *Вестник МГТУ им. Н.Э. Баумана. Сер. Приборостроение*, 2018, № 3, с. 70–82.
DOI: 10.18698/0236-3933-2018-3-70-82

[12] Salamon J., Jacoby C., Bello J.P. A dataset and taxonomy for urban sound research. *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 1041–1044.
DOI: 10.1145/2647868.2655045

[13] Hertel L., Phan H., Mertins A. Comparing time and frequency domain for audio event recognition using deep learning. *IEEE IJCNN*, 2016, pp. 3407–3411.
DOI: 10.1109/IJCNN.2016.7727635

[14] Fulop S.A., Fitz K. Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications. *J. Acoust. Soc. Am.*, 2006, vol. 119, iss. 1, pp. 360–371. DOI: 10.1121/1.2133000

[15] Maas A., Hannun A., Ng A. Rectifier nonlinearities improve neural network acoustic models. *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013. URL: https://ai.stanford.edu/~amaas/papers/relu_hybrid_icml2013_final.pdf (дата обращения: 09.07.2018).

[16] Xu H., Chen W., Zhao N., et al. Unsupervised anomaly detection via variational auto-encoder for seasonal KPIs in web applications. *Proc. 2018 World Wide Web Conf.*, 2018, pp. 187–196. DOI: 0.1145/3178876.3185996

[17] Kingma D., Ba J. Adam: a method for stochastic optimization. *ICLR*, 2015.
URL: https://arxiv.org/pdf/1412.6980.pdf (дата обращения: 09.07.2018).

[18] Yang X. Understanding the variational lower bound.
URL: http://legacydirs.umiacs.umd.edu/~xyang35/files/
understanding-variational-lower.pdf (дата обращения: 09.07.2018).

[19] Press W.H., Teukolsky S.A., Vetterling W.T., et al. The art of scientific computing. Cambridge Univ. Press, 2007.

[20] Maaten L., Hinton G. Visualizing data using t-SNE. *JMLR*, 2008, no. 9, no. 1, pp. 2579–2605.

[21] Kumatani K., McDonough J., Raj B. Microphone array processing for distant speech recognition: from close-talking microphones to far-field sensors. *IEEE Signal Process. Mag.*, 2012, vol. 29, iss. 6, pp. 127–140. DOI: 10.1109/MSP.2012.2205285

[22] Suvorov D.A., Ge D., Zhukov R.A. Deep residual network for sound source localization in the time domain. *JEAS*, 2018, vol. 13, no. 13, pp. 5096–5104.

[23] Tashev I. Sound capture and processing: practical approaches. John Wiley & Sons, 2009.

[24] Aleinik S. Acceleration of Zelinski post-filtering calculation. *J. Sign. Process Syst.*, 2017, vol. 88, iss. 3, pp. 463–468. DOI: 10.1007/s11265-016-1191-9

Сарафасланян Арташес Хачатурович — канд. физ.-мат. наук, технический директор ООО «Бизнес Бюро» (Российская Федерация, 125438, Москва, ул. Автомоторная, д. 4а, стр. 21).

Чепраков Вячеслав Валерьевич — инженер ООО «Бизнес Бюро» (Российская Федерация, 125438, Москва, ул. Автомоторная, д. 4а, стр. 21).

Суворов Дмитрий Андреевич — программист ООО «Бизнес Бюро» (Российская Федерация, 125438, Москва, ул. Автомоторная, д. 4а, стр. 21).

Мозговой Михаил Владимирович — заместитель директора по методической работе Головного учебно-исследовательского и методического центра профессиональной реабилитации лиц с ограниченными возможностями здоровья (инвалидов) МГТУ им. Н.Э. Баумана (Российская Федерация, 105005, Москва, 2-я Бауманская ул., д. 5, стр. 1).

Волков Алексей Васильевич — директор Ресурсного учебно-методического центра по обучению инвалидов и лиц с ограниченными возможностями здоровья МГТУ им. Н.Э. Баумана (Российская Федерация, 105005, Москва, 2-я Бауманская ул., д. 5, стр. 1).

Просьба ссылаться на эту статью следующим образом:

Сарафасланян А.Х., Чепраков В.В., Суворов Д.А. и др. Детектирование неизвестных звуков для людей с нарушенным слухом на основе вариационного автоэнкодера. *Вестник МГТУ им. Н.Э. Баумана. Сер. Приборостроение*, 2019, № 1, с. 35–49. DOI: 10.18698/0236-3933-2019-1-35-49

EMPLOYING A VARIATIONAL AUTO-ENCODER TO DETECT UNKNOWN SOUNDS FOR HEARING-IMPAIRED PEOPLE

A.Kh. Sarafaslanyan[1] artashes@bizb.ru
V.V. Cheprakov[1] cpb@bizb.ru
D.A. Suvorov[1] suvorov@bizb.ru
M.V. Mozgovoi[2] mozgovoy@bmstu.ru
A.V. Volkov[2] avv@bmstu.ru

[1] **Business Bureau Company Limited, Moscow, Russian Federation**
[2] **Bauman Moscow State Technical University, Moscow, Russian Federation**

Abstract

The paper presents a system of detecting unknown sounds for hearing-impaired people built upon a variational auto-encoder. We define the architecture of our variational autoencoder, the encoder and decoder in which both consist of fully connected layers. We describe the process of creating the dataset and splitting it into training, test and unknown sound detection subsets. We then describe the method of training the system and the mathematics behind it, including the *Adam* stochastic optimization method and a variational lower bound as a loss function. We tested our system and established that there are no false negative detection results for unknown sounds and that the false positive result probability is 14 %, which is quite acceptable in practice. We provide the technology we used to implement the system and the device that should house it. We consider possible ways of further improving the system

Keywords

Variational autoencoder, deep learning, sound recognition, digital signal processing, detection, learning

Received 24.07.2018

This work was supported by the Innovation Promotion Foundation (grant no. 168GRNTIS5/35848)

REFERENCES

[1] Hersh M. Deafblind people, communication, independence, and isolation. *J. Deaf. Stud. Deaf. Educ.*, 2013, vol. 18, iss. 4, pp. 446–463. DOI: 10.1093/deafed/ent022

[2] Sainath T.N., Parada C. Convolutional neural networks for small-footprint keyword spotting. *INTERSPEECH*, 2015, pp. 1478–1482.

[3] Tzanetakis G., Cook P. Musical genre classification of audio signals. *IEEE Trans. Speech Audio Process.,* 2002, vol. 10, iss. 5, pp. 293–302. DOI: 10.1109/TSA.2002.800560

[4] Tavares T.F., Foleiss J.H. Automatic music genre classification in small and ethnic datasets. *Proc. 13th CMMR Int. Symp.,* 2017, pp. 25–28.

[5] Bragg D., Huynh N., Ladner R.E. A personalizable mobile sound detector app design for deaf and hard-of-hearing users. *Proc. 18th Int. ACM SIGACCESS Conf. Computers Accessibility*, 2016, pp. 3–13. DOI: 10.1145/2982142.2982171

[6] Lecomte S., Lengellé R., Richard C., et al. Abnormal events detection using unsupervised one-class SVM — Application to audio surveillance and evaluation. *8th IEEE AVSS Int. Conf.*, 2011, pp. 124–129.

[7] Bishop C.M. Pattern recognition and machine learning. Springer, 2006.

[8] Zong B., Song Q., Min M.R., et al. Deep autoencoding Gaussian mixture model for unsupervised anomaly detection. *ICLR*, 2018.
Available at: https://openreview.net/pdf?id=BJJLHbb0- (accessed: 09.07.2018).

[9] Oh D.Y., Yun I.D. Residual error based anomaly detection using autoencoder in SMD machine sound. *Sensors*, 2018, vol. 18, no. 5, art. 1308. DOI: 10.3390/s18051308

[10] Kingma D.P., Welling M. Auto-encoding variational Bayes. *ICLR*, 2014. Available at: https://arxiv.org/pdf/1312.6114.pdf (accessed: 09.07.2018).

[11] Zhukov R.A., Suvorov D.A., Teteryukov D.O. et al. Designing a signal input subsystem based on a digitally interfaced microphone array. *Vestn. Mosk. Gos. Tekh. Univ. im. N.E. Baumana, Priborostr.* [Herald of the Bauman Moscow State Tech. Univ., Instrum. Eng.], 2018, no. 3, pp. 70–82 (in Russ.). DOI: 10.18698/0236-3933-2018-3-70-82

[12] Salamon J., Jacoby C., Bello J.P. A dataset and taxonomy for urban sound research. *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 1041–1044. DOI:10.1145/2647868.2655045

[13] Hertel L., Phan H., Mertins A. Comparing time and frequency domain for audio event recognition using deep learning. *IEEE IJCNN*, 2016, pp. 3407–3411. DOI: 10.1109/IJCNN.2016.7727635

[14] Fulop S.A., Fitz K. Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications. *J. Acoust. Soc. Am.*, 2006, vol. 119, iss. 1, pp. 360–371. DOI: 10.1121/1.2133000

[15] Maas A., Hannun A., Ng A. Rectifier nonlinearities improve neural network acoustic models. *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013. Available at: https://ai.stanford.edu/~amaas/papers/relu_hybrid_icml2013_final.pdf (accessed: 09.07.2018).

[16] Xu H., Chen W., Zhao N., et al. Unsupervised anomaly detection via variational auto-encoder for seasonal KPIs in web applications. *Proc. 2018 World Wide Web Conf.,* 2018, pp. 187–196. DOI: 0.1145/3178876.3185996

[17] Kingma D., Ba J. *Adam*: a method for stochastic optimization. *ICLR*, 2015. Available at: https://arxiv.org/pdf/1412.6980.pdf (accessed: 09.07.2018).

[18] Yang X. Understanding the variational lower bound. Available at: http://legacydirs.umiacs.umd.edu/~xyang35/files/understanding-variational-lower.pdf (accessed: 09.07.2018).

[19] Press W.H., Teukolsky S.A., Vetterling W.T., et al. The art of scientific computing. Cambridge Univ. Press, 2007.

[20] Maaten L., Hinton G. Visualizing data using t-SNE. *JMLR*, 2008, no. 9, no. 1, pp. 2579–2605.

[21] Kumatani K., McDonough J., Raj B. Microphone array processing for distant speech recognition: from close-talking microphones to far-field sensors. *IEEE Signal Process. Mag.*, 2012, vol. 29, iss. 6, pp. 127–140. DOI: 10.1109/MSP.2012.2205285

[22] Suvorov D.A., Ge D., Zhukov R.A. Deep residual network for sound source localization in the time domain. *JEAS*, 2018, vol. 13, no. 13, pp. 5096–5104.

[23] Tashev I. Sound capture and processing: practical approaches. John Wiley & Sons, 2009.

[24] Aleinik S. Acceleration of Zelinski post-filtering calculation. *J. Sign. Process Syst.*, 2017, vol. 88, iss. 3, pp. 463–468. DOI: 10.1007/s11265-016-1191-9

Sarafaslanyan A.Kh. — Cand. Sc. (Phys.-Math.), Chief Technology Officer, Business Bureau Company Limited (Avtomotornaya ul. 4a, str. 21, Moscow, 125438 Russian Federation).

Cheprakov V.V. — Engineer, Business Bureau Company Limited (Avtomotornaya ul. 4a, str. 21, Moscow, 125438 Russian Federation).

Suvorov D.A. — Programmer, Business Bureau Company Limited (Avtomotornaya ul. 4a, str. 21, Moscow, 125438 Russian Federation).

Mozgovoi M.V. — Deputy Director in Methodological Work, Head Training, Research and Methodological Centre for Vocational Rehabilitation of the Health-Impaired (the Disabled), Bauman Moscow State Technical University (2-ya Baumanskaya ul. 5, str. 1, Moscow, 105005 Russian Federation).

Volkov A.V. — Director, Resource Training and Methodological Centre for the Education of the Disabled and Health-Impaired, Bauman Moscow State Technical University (2-ya Baumanskaya ul. 5, str. 1, Moscow, 105005 Russian Federation).

Please cite this article in English as:

Sarafaslanyan A.Kh., Cheprakov V.V., Suvorov D.A., et al. Employing a Variational Auto-Encoder to Detect Unknown Sounds for Hearing-Impaired People. *Herald of the Bauman Moscow State Technical University*, *Series Instrument Engineering*, 2019, no. 1, pp. 35–49 (in Russ.). DOI: 10.18698/0236-3933-2019-1-35-49