ISSN 1816-0301 (Print)
ISSN 2617-6963 (Online)

УДК 004.934
https://doi.org/10.37661/1816-0301-2020-17-2-36-43

Поступила в редакцию 06.12.2019
Received 06.12.2019

Принята к публикации 08.04.2020
Accepted 08.04.2020

Выделение речевой активности на фоне шумов при помощи компактной сверточной нейронной сети

Г. С. Вашкевич✉, И. С. Азаров

Белорусский государственный университет информатики и радиоэлектроники, Минск, Беларусь
✉E-mail: ryhorv@gmail.com

Аннотация. Исследуется задача выделения речевой активности из зашумленного звукового сигнала. Предлагается компактная модель сверточной нейронной сети, которая имеет всего 385 параметров. Модель нетребовательна к вычислительным ресурсам, что позволяет использовать ее в рамках концепции Интернета вещей для портативных устройств с низким энергопотреблением. В то же время эта модель обеспечивает высокую точность определения речевой активности на уровне лучших современных аналогов. Указанные полезные свойства достигаются путем применения специального сверточного слоя, учитывающего гармоническую структуру вокализованной речи и устраняющего избыточность модели за счет инвариантности к изменениям частоты основного тона. В рамках экспериментов производительность модели оценивалась в различных шумовых условиях для разных соотношений сигнала и шума. Результаты экспериментов показали, что предложенная модель обеспечивает более высокую точность определения речевой активности по сравнению с моделью, представленной компанией Google в фреймворке WebRTC.

Ключевые слова: детектор речевой активности, гармонический сигнал, сверточная нейронная сеть, частота основного тона, обработка речи

Для цитирования. Вашкевич, Г. С. Выделение речевой активности на фоне шумов при помощи компактной сверточной нейронной сети / Г. С. Вашкевич, И. С. Азаров // Информатика. – 2020. − Т. 17, № 2. – С. 36–43. https://doi.org/10.37661/1816-0301-2020-17-2-36-43

Voice activity detection in noisy conditions using tiny convolutional neural network

Ryhor S. Vashkevich✉, Elias S. Azarov

Belarusian State University of Informatics and Radioelectronics, Minsk, Belarus
✉E-mail: ryhorv@gmail.com

Abstract. The paper investigates the problem of voice activity detection from a noisy sound signal. An extremely compact convolutional neural network is proposed. The model has only 385 trainable parameters. Proposed model doesn't require a lot of computational resources that allows to use it as part of the "internet of things" concept for compact low power devices. At the same time the model provides state of the art results in voice activity detection in terms of detection accuracy. The properties of the model are achieved by using a special convolutional layer that considers the harmonic structure of vocal speech. This layer also eliminates redundancy of the model because it has invariance to changes of fundamental frequency. The model performance is evaluated in various noise conditions with different signal-to-noise ratios. The results show that the proposed model provides higher accuracy compared to voice activity detection model from the WebRTC framework by Google.

Keywords: voice activity detector, harmonic signal, convolutional neural network, pitch, speech processing

For citation. Vashkevich R. S., Azarov E. S. Voice activity detection in noisy conditions using tiny convolutional neural network. *Informatics*, 2020, vol. 17, no. 2, pp. 36−43 (in Russian). https://doi.org/10.37661/1816-0301-2020-17-2-36-43

Введение. Выделение речевой активности из звукового сигнала является актуальной задачей, которая часто возникает при построении различных систем обработки речи. Разработка детектора речи в реальных условиях эксплуатации дополнительно усложняется из-за наличия в речи посторонних звуков (шума). При высоком уровне зашумления выделить речь становится и вовсе невозможно.

В большинстве научных публикаций предлагаются решения для выделения речевой активности на основе методов машинного обучения, которые обеспечивают высокую точность выделения, но вместе с тем являются вычислительно затратными. С целью сокращения вычислений рассматриваются альтернативные подходы, в том числе и аналитическая модель определения речевой активности на основе анализа формант гармонического сигнала [1]. К недостаткам данного метода можно отнести допущение о том, что речевой сигнал всегда имеет гармоническую структуру. В работе [2] предлагается отказаться от применения технологий машинного обучения путем анализа спектрограммы звукового сигнала, разделенной на две части. Допускается, что информация о речевом сигнале всегда содержится в нижней полосе частот, а в верхней располагается шум. Недостатком такого подхода является чувствительность к низкочастотным шумам.

Решения на основе методов машинного обучения в целом более устойчивы к шумам, поскольку способны учитывать различную природу шумов и большую вариацию человеческого голоса. Один из самых простых примеров применения машинного обучения представлен в работе [3], где используется метод опорных векторов для определения двух классов характеристических признаков речевого сигнала: голоса и шума. Во многих решениях применяются искусственные нейронные сети, которые являются более мощным инструментом для классификации характеристических признаков [4–9]. Высокая вычислительная сложность методов на основе искусственных нейронных сетей связана с большим числом параметров. В частности, в работе [4] при помощи сверточных нейронных сетей предлагается моделировать частотные вариации входного сигнала, а при помощи слоев долгой краткосрочной памяти (long short-term memory, LSTM) сети учитывать его временные вариации. Модель имеет примерно 100 000 параметров. В работах [5, 9] представлены нейронные сети, состоящие только из полносвязных слоев с очень большим числом параметров (модель с 1 млн параметров в [9]), которые требуют большой обучающей выборки и склонны к переобучению. В решениях [5, 8] предлагается сначала удалить шумы из исходного сигнала и оставить только голос. Таким образом, первые несколько слоев моделей выполняют очистку сигнала от шума, а последующие слои используются в качестве классификатора очищенного сигнала.

В публикациях [4, 6, 7] принимается во внимание тот факт, что речевой сигнал является протяженным во времени, и выделяются протяженные во времени признаки речи при помощи рекуррентных нейронных сетей. При этом модель, предложенная в работе [6], имеет достаточно малое количество настраиваемых параметров (около 350), что выгодно выделяет ее на фоне других. Однако рекуррентные нейронные сети сложны в обучении и требуют большой обучающей выборки для надежной работы, поскольку имеют очень большое число возможных состояний.

Во многих задачах обработки речи, в том числе и в задачах детектирования речевой активности, в качестве характеристических признаков используются мел-кепстральные коэффициенты (mel-frequency cepstral coefficients, MFCC). Эти признаки в работах [3, 5, 9] успешно применяются для детектирования речевой активности. В статьях [1, 2] выделяют признаки, основанные на спектрограмме сигнала, принимая во внимание тот факт, что речевой сигнал в большинстве случаев имеет гармоническую структуру.

В последнее время начали набирать популярность end-to-end подходы к построению моделей глубокого обучения. Их суть заключается в том, что на вход глубокой нейросетевой модели подаются данные, которые не прошли предварительную обработку и из которых не извлечены

характеристические признаки. Эти данные сами по себе являются признаками. Например, в работе [4] авторы на вход нейронной сети подают такую аудиозапись, что помогает избежать дополнительных вычислительных затрат на предобработку входных данных. Однако подобные модели тяжело обучать, поскольку для этого требуется огромное количество данных.

В настоящей работе предлагается ультракомпактная модель сверточной нейронной сети для определения речевой активности, которая имеет всего 385 параметров. Предложенная модель нетребовательна к вычислительным ресурсам, что позволяет использовать ее в рамках концепции Интернета вещей для портативных устройств с низким энергопотреблением. Точность определения речевой активности предложенной модели находится на уровне лучших современных аналогов и по результатам экспериментов превосходит наиболее популярный в настоящее время детектор речи, представленный компанией Google в фреймворке WebRTC. Полезные свойства предложенной модели обеспечиваются использованием специального сверточного слоя, учитывающего гармоническую структуру вокализованной речи и устраняющего избыточность за счет инвариантности к изменениям частоты основного тона. В рамках выполненных экспериментов производительность модели оценивалась в различных шумовых условиях с разными соотношениями сигнала и шума. Реализация предложенной модели для выделения речевой активности доступна по адресу https://github.com/gvashkevich/vad.

Сверточные нейронные сети для обработки речи. В задачах обработки речи, таких как распознавание, синтез и детектирование, применяют сверточные нейронные сети, состоящие из последовательности сверточных слоев (convolution layers) и слоев объединения (pooling layers). Общая структура одной из таких сетей показана на рис. 1. Входными признаками являются кратковременные амплитудные спектры речевого сигнала.

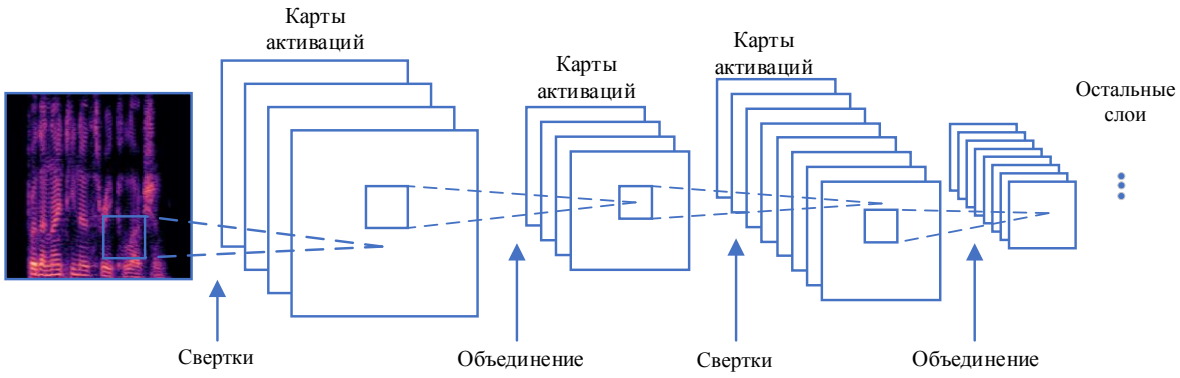


Рис. 1. Сверточная нейронная сеть

Свертка выполняется с фиксированным окном и небольшим шагом. Такой способ обработки звуковых сигналов заимствован из задач компьютерного зрения и обработки изображений, где соседние пикселы равноценны по своему значению. Однако носителем речи служит спектр со специфической структурой. Большая часть речи является вокализованной (от 80 до 95 % общей продолжительности речевого сигнала) и образуется при помощи голосовых связок, создающих периодические колебания. По этой причине спектр речи имеет преимущественно гармоническую структуру и почти вся значимая информация, необходимая для решения задачи, сконцентрирована на частотах, пропорциональных частоте основного тона ($F0$) (рис. 2).

Таким образом, весь амплитудный спектр в качестве характеристического вектора является избыточным, причем избыточность можно устранить, выбирая информационно важные компоненты спектра, кратно соответствующие частоте основного тона. Между тем необходимо учитывать, что частота основного тона изменяется во времени в достаточно широких пределах. Считается, что для речевых приложений ее диапазон варьируется от 50 до 450 Гц.

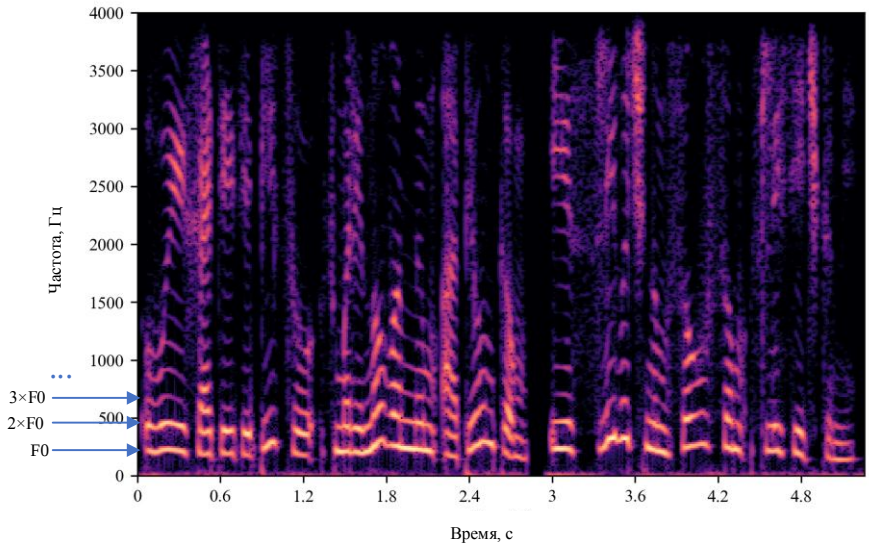


Рис. 2. Амплитудный спектр речевого сигнала

Для устранения избыточности нейронной сети предлагается применять свертки амплитудного спектра, инвариантные к изменению основного тона. Зададим N гипотез о том, что $F0$ принимает значения в допустимом диапазоне от $F0_{min}$ до $F0_{max}$. Для каждой гипотезы выберем M информационно важных составляющих спектра, соответствующих гармоникам речевого сигнала, частоты которых пропорциональны частоте основного тона. Из выбранных компонент спектра сформируем матрицу признаков размером $N \times M$, где вдоль первой оси будут располагаться гипотезы, а вдоль второй – гармоники, соответствующие гипотезам. Полученную матрицу признаков подадим на вход классической сверточной нейронной сети для классификации речи и шума (рис. 3).

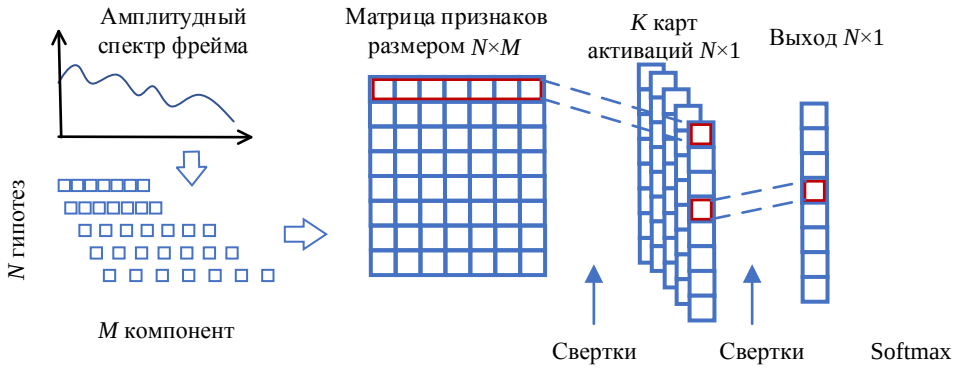


Рис. 3. Предлагаемая модель сверточной нейронной сети для детектирования речевого сигнала

Полученная модель состоит из одного сверточного и одного полносвязного слоя. В свою очередь, сверточный слой состоит из K фильтров с ядром $1 \times M$ и функций активации ReLU. Ядро такого размера позволяет осуществлять операцию свертки только вдоль оси с гармониками исходного сигнала, обрабатывая каждую гипотезу независимо друг от друга. Результат работы сверточного слоя агрегируется при помощи полносвязного слоя с одним нейроном. Для получения распределения вероятностей того, какая из N гипотез содержит в себе гармоническую составляющую, выход второго слоя поступает на функцию активации softmax.

Формирование характеристических признаков. Как было описано в предыдущем разделе, формирование характеристических признаков происходит путем выбора частотных компонент спектрограммы сигнала, кратных заданной частоте основного тона $F0$. Гипотетические

значения $F0_i$ получаются путем равномерного разбиения частотного интервала $F0_{\min}$ ... $F0_{\max}$ на N значений:

$$F0_i = F0_{\min} + i \cdot \frac{F0_{\max} - F0_{\min}}{N}, \; i = 0, ..., N-1.$$

Далее для каждой гипотезы $F0_i$ из исходной спектрограммы выбираются по M гармоник. Индексы I_i всех M гармоник в спектрограмме с частотным разрешением f_r для заданной $F0_i$ вычисляются при помощи выражения

$$I_i^j = \text{round}\left(\frac{1+j}{2} \cdot \frac{F0_i}{f_r} + 1\right), \; j = 0, ..., M-1,$$

где round$(*)$ – операция округления.

Частотное разрешение спектрограммы определяется по формуле

$$f_r = \frac{f_s}{N_{fft}},$$

где f_s – частота дискретизации исходного сигнала, N_{fft} – размер быстрого преобразования Фурье.

Для получения результирующего характеристического вектора признаков X для одного фрейма звукового сигнала s необходимо вычислить дискретное преобразование Фурье, а затем из амплитудного спектра выбрать только те компоненты, которые соответствуют вычисленным индексам I_i^j:

$$S = \log_{10}|FFT(s)|,$$

$$X(i,j) = S\left(I_i^J\right), \; i = 0, ..., M-1, \; j = 0, ..., M-1.$$

Таким образом, для каждого входного фрейма формируется матрица признаков X размером $N \times M$, состоящая из N гипотез по M компонент. Общая идея выделения признаков показана на рис. 4, где черным цветом отмечены элементы матрицы признаков X, а красным – элементы матрицы признаков X с частотой $F0$ для каждого кандидата.

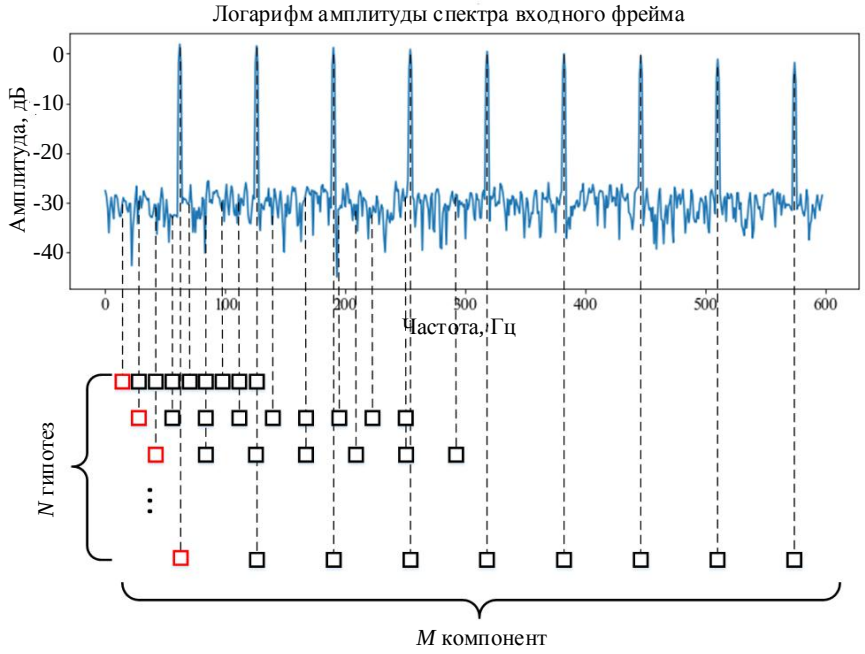


Рис. 4. Процесс формирования матрицы признаков

Задача нейронной сети состоит в определении существования среди предложенных гипотез такой, которая содержит амплитуды гармоник исходного сигнала. Если модель сможет обнаружить эту гипотезу, то текущий входной пример соответствует речевому сигналу, в противном случае данный пример классифицируется как шум.

Экспериментальные исследования. Для тестирования и обучения модели применялся открытый набор данных Musan [10]. Набор состоит из записей трех типов: речи, шума и музыки, однако в экспериментах использовались только записи речи и шума. Обучение производилось на речевых данных с различными шумами и соотношениями сигнала и шума (signal-to-noise ratio, SNR). Для каждой речевой записи случайно выбирались три шумовые записи. Их содержимое по отдельности прибавлялось к речевой записи с заданным коэффициентом k, определяющим значение SNR зашумленного сигнала:

$$\mathrm{SNR} = 20 \cdot \log_{10}\left(\frac{A}{k \cdot A_n}\right),$$

где A и A_n – средние мощности речевого и шумового сигналов соответственно,

$$A = \frac{1}{len(s)}\sum s^2.$$

Значение коэффициента определялось по формуле

$$k = 10^{\frac{\log_{10}\left(\frac{A}{A_n}\right) - \frac{\mathrm{SNR}}{20}}{2}}.$$

Следует учитывать, что значения мощностей A и A_n – квадратичные величины, поэтому зашумленный аддитивным шумом сигнал вычислялся согласно выражению

$$s_{add} = s + \sqrt{k} \cdot s_n.$$

В исследовании значение SNR выбиралось случайным образом в диапазоне от 10 до 20 дБ. Зашумленный входной сигнал разбивался на фреймы длиной 50 мс с шагом 12,5 мс, после чего из полученных фреймов формировались характеристические признаки с диапазоном от 75 до 350 Гц, которые подавались на вход классификатора. Следовательно, каждый фрейм классифицировался независимо от остальных.

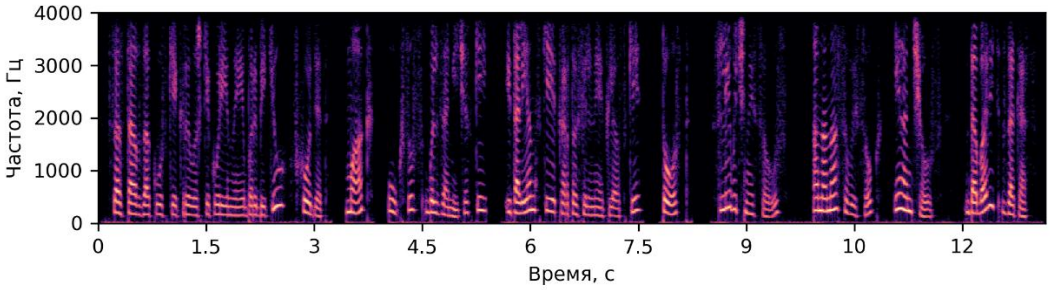
Если хотя бы один из выходов сети принимал значение активации больше порогового, то считалось, что входной фрейм относится к речи, в противном случае – к шуму. Использованное пороговое значение составляло 0,15.

В качестве целевых значений для тренировочных данных применялось значение частоты основного тона речевых записей, полученное с помощью алгоритма YAPT [11]. Полученные значения квантовались таким образом, чтобы количество уровней квантования соответствовало числу гипотез. При этом невокализованные участки относились к нулевому уровню.

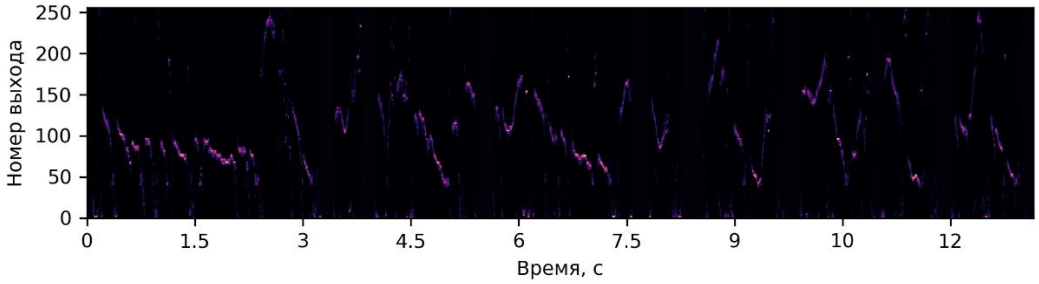
В ходе экспериментов были эмпирически подобраны оптимальные значения для количественных параметров гармоник $M = 22$, гипотез $N = 100$ и фильтров сверточного слоя $K = 16$.

Обучение сети осуществлялось методом Adam [12] с шагом обучения 0,001. Обучение продолжалось 50 000 итераций с размером минимальной партии в 256 элементов.

Анализ работы модели. Визуализация активаций выходного слоя сети позволяет оценить, какая именно гипотеза во входном векторе признаков представляет собой гармоническую структуру входного сигнала. На рис. 5, *б* представлены активации выхода нейронной сети для участка аудиозаписи, спектрограмма которого показана на рис. 5, *а*. Видно, что активации сети имеют высокое значение на тех участках изображения, которые соответствуют вокализованным участкам речи на спектрограмме. При этом изменение номера гипотезы с максимальной активацией повторяет изменение частоты основного тона в исходном сигнале, что обеспечивает возможность применения предложенной модели в качестве оценщика тона.



а)



б)

Рис. 5. Визуализация работы сверточной сети

Предложенная модель имеет всего 385 обучаемых коэффициентов, что позволяет ей обучаться на небольшой выборке данных и выделять речевую активность небольшим объемом вычислений. При этом анализ других нейросетевых моделей, описанных в первом разделе, показывает, что они имеют значительно больше обучаемых параметров и требуют значительно больших вычислительных ресурсов как для обучения, так и для применения.

Чтобы получить численные значения, характеризующие качество работы предложенной модели относительно других современных моделей, используется метрика AUC. Предложенная модель сравнивается с моделью, разработанной компанией Google для фреймворка WebRTC (URL: https://github.com/wiseman/py-webrtcvad). В настоящее время детектор речевой активности из WebRTC является одной из наиболее популярных открытых моделей для решения этой задачи. Точность предложенной модели (0,8821) превосходит точность модели WebRTC (0,8755) на тестовом наборе данных.

При решении ряда задач, связанных с выделением речевой активности, бывает недостаточно только пофреймовой обработки входного сигнала. Поэтому дальнейшая работа будет посвящена выделению речевой активности на более высоком уровне, где будут учитываться слова или фразы речи. Поскольку предложенная модель способна эффективно выделять речевые признаки, связанные с интонацией речи, данную задачу, предположительно, можно решить, анализируя последовательности обработанных на основе предложенной модели фреймов путем добавления надстройки из дополнительных слоев нейронной сети.

Заключение. Основой предложенной модели выделения речевой активности из зашумленного звукового сигнала на базе компактной сверточной нейронной сети служит специальный сверточный слой, который учитывает гармоническую структуру вокализованных участков речевого сигнала. Добавление такого слоя позволяет значительно сократить количество настраиваемых параметров нейросетевой модели, что обусловливает ее низкую требовательность к объему обучающих данных и вычислительным ресурсам. Данный факт делает предложенную модель идеальной для реализации во встраиваемых системах и мобильных устройствах с низким энергопотреблением.

Качество работы модели подтверждено сравнением с популярным современным решением для выделения речевой активности из фреймворка WebRTC от компании Google: предложенная модель по точности детектирования речи близка к модели WebRTC.

References

1. Yoo I.-C., Lim H., Yook D. Formant-based robust voice activity detection. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 2015, vol. 23, no. 12, pp. 2238–2245. https://doi.org/10.1109/TASLP.2015.2476762

2. Pang J. Spectrum energy based voice activity detection. *The 7th IEEE Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, 9–11 January 2017*. Las Vegas, 2017, pp. 1–5. https://doi.org/10.1109/CCWC.2017.7868454

3. Kinnunen T., Chernenko E., Tuononen M., Fränti P., Li H. Voice activity detection using MFCC features and support vector machine. *The 12th International Conference on Speech and Computer (SPECOM07), Moscow, Russia, 15–18 October 2007*. Moscow, 2007, vol. 2, pp. 556–561.

4. Zazo R., Sainath T. N., Simko G., Parada C. Feature learning with raw-waveform CLDNNs for voice activity detection. *17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, 8–12 September 2016*. San Francisco, 2016, pp. 3668–3672. https://doi.org/10.21437/Interspeech.2016-268

5. Zhang X., Wu J. Denoising deep neural networks based voice activity detection. *International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013*. Vancouver, 2013, pp. 853–857. https://doi.org/10.1109/ICASSP.2013.6637769

6. Hughes T., Mierle K. Recurrent neural networks for voice activity detection. *International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013*. Vancouver, 2013, pp. 7378–7382. https://doi.org/10.1109/ICASSP.2013.6639096

7. Eyben F., Weninger F., Squartini S., Schuller B. Real-life voice activity detection with LSTM Recurrent Neural Networks and an application to Hollywood movies. *International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013*. Vancouver, 2013, pp. 483–487. https://doi.org/10.1109/ICASSP.2013.6637694

8. Wang Q., Du J., Bao X., Wang Z.-R., Dai L.-R., Lee C.-H. A universal VAD based on jointly trained deep neural networks. *16th Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015*. Dresden, 2015, pp. 2282–2286.

9. Ryant N., Liberman M., Yuan J. Speech activity detection on youtube using deep neural networks. *14thAnnual Conference of the International Speech Communication Association, Lyon, France, 25–29 August 2013*. Lyon, 2013, pp. 728–731.

10. Snyder D., Chen G., Povey D. *Musan: a Music, Speech, and Noise Corpus*, 2015. Available at: https://arxiv.org/abs/1510.08484 (accessed 20.10.2019).

11. Kasi K., Zahorian S. A. Yet another algorithm for pitch tracking. *International Conference on Acoustics, Speech, and Signal Processing, Orlando, 13–17 May 2002*. Orlando, 2002, vol. 1, pp. 361–364. https://doi.org/10.1109/ICASSP.2002.5743729

12. Kingma D. P., Ba J. *Adam: a Method for Stochastic Optimization*, 2014. Available at: https://arxiv.org/abs/1412.6980 (accessed 20.10.2019).

Информация об авторах

Вашкевич Григорий Сергеевич, магистр технических наук, аспирант кафедры ЭВС, Белорусский государственный университет информатики и радиоэлектроники, Минск, Беларусь.
E-mail: ryhorv@gmail.com

Азаров Илья Сергеевич, доктор технических наук, доцент, заведующий кафедрой ЭВС, Белорусский государственный университет информатики и радиоэлектроники, Минск, Беларусь.

Information about the authors

Ryhor S. Vashkevich, M. Sci. (Eng.), Postgraduate Student of the Department of EMU, Belarusian State University of Informatics and Radioelectronics Minsk, Belarus.
E-mail: ryhorv@gmail.com

Elias S. Azarov, Dr. Sci. (Eng.), Associate Professor, Head of the Department of EMU, Belarusian State University of Informatics and Radioelectronics Minsk, Belarus.