УДК 004.934.2 DOI: 10.17587/mau.19.53-57

Д. А. Суворов, аспирант, dmitry.suvorov@skolkovotech.ru,
Р. А. Жуков, аспирант, roman.zhukov@skolkovotech.ru,
Д. О. Тетерюков, Ph. D, ст. преподаватель, d.tsetserukou@skoltech.ru,
Сколковский институт науки и технологий, г. Москва,
С. Л. Зенкевич, д-р физ. мат. наук, проф., zenkev@bmstu.ru,
Московский государственный технический университет имени Н. Э. Баумана, г. Москва

Аудиовизуальный детектор голосовой активности на базе глубокой сверточной сети и обобщенной взаимной корреляции[1]

Разработан алгоритм детектора голосовой активности, использующий данные с видеокамеры и массива микрофонов и благодаря этому обладающий высокой устойчивостью к внешним шумам. Обработка видеокадров заключается в поиске губ человека с помощью глубокой сверточной нейронной сети, обработка звука — в локализации источников звука с помощью обобщенной функции взаимной корреляции с весовой функцией преобразования фазы (GCC-PHAT). Решение об активации детектора голосовой активности принимается только в случае нахождения соответствия между направлением на губы и на активные источники звука. Разработанный детектор показал высокую устойчивость к шумам — шумы, производимые источниками звука вне видимости видеокамеры или целевого сектора для массива микрофонов, игнорировались в 100 % случаев во время эксперимента.

Ключевые слова: *детектор голосовой активности, массив микрофонов, сверточные сети, локализация источников звука, обработка звука*

Введение

В последнее время голосовые интерфейсы становятся неотъемлемой частью взаимодействия с робототехническими комплексами. В таких известных примерах сервисных роботов, как Amazon Echo, Google Home, Pepper, распознавание речи является ключевым компонентом системы. Детектор голосовой активности — это критически важный компонент системы распознавания речи, который значительно влияет на ее точность и производительность [1]. При правильном выделении участков звука, в которых присутствует целевой голосовой сигнал, детектор наличия голоса значительно уменьшает объем данных для обработки системой распознавания речи, что в итоге ускоряет ее работу и уменьшает вероятность ложных распознаваний. Особенно это свойство важно при работе систем распознавания речи на персональных устройствах пользователей — смартфонах, ноутбуках, смарт-телевизорах, у которых в отличие от специализированных серверов ограничена процессорная мощность.

Известно множество методов детектирования голосовой активности, основанных на: пороговых значениях энергии [2], спектральных признаках [3], нейронных сетях [4], аудиовизуальных признаках [5], моделях гауссовых смесей [6], данных от массива микрофонов [7]. Качество работы большинства известных методов быстро деградирует с ухудшением соотношения сигнал—шум [7]. Многие методы не способны работать в шумных местах при нахождении источника звука даже на небольшом расстоянии от микрофона. В данной статье предложен аудиовизуальный метод детектирования наличия голоса, использующий глубокие сверточные сети для анализа видеоданных, и алгоритм "взвешенный GCC-PHAT" [8] для анализа аудиоданных с массива микрофонов. Метод пригоден для работы в реальном времени, в том числе и на мобильных устройствах.

Целью данной работы является разработка детектора голосовой активности, способного игнорировать источники звука с направлений, где целевой источник речи не может присутствовать физически, и отсеивать источники звука, не являющиеся человеческой речью.

Постановка задачи детектирования голосовой активности

Имеется ограниченный дискретный сигнал, представленный во временной области, $x[lT]$, где $0 \leqslant l \leqslant L-1$, $x_{\min} \leqslant x[lT] \leqslant x_{\max}$, T — период дискретизации, L — число измерений в сигнале. Для звукового сигнала обычно

$$x_{\min} = -1;$$

$$x_{\max} = 1.$$

Сигнал $x[lT]$ является смесью двух ограниченных, дискретных и некоррелированных сигналов [9]:

$$x[lT] = s[lT] + n[lT], \quad (1)$$

где $n[lT]$ — это шум микрофона и окружающей среды, $s[lT]$ — это речевой сигнал.

[1] Исследование частично выполнено за счет гранта Фонда содействия инновациям (проект № 102ГРНТИС5/26071).

Анализ сигнала осуществляется покадрово с выделением кадра с помощью оконной функции. В частотной области можно записать:

$$X^n[k] = S^n[k] + N^n[k], \quad (2)$$

где k — индекс частоты ($0 \leqslant k \leqslant K-1$); K — число частот в дискретном преобразовании Фурье, n — номер анализируемого кадра.

Для каждого кадра n рассматриваются два вида гипотез:

- H_0: в сигнале X^n отсутствует речь, $X^n = N^n$;
- H_1: в сигнале X^n присутствует речь, $X^n = S^n + N^n$.

Задача детектирования голосовой активности заключается в отнесении кадра n к классу H_0 или H_1.

Алгоритм детектирования голосовой активности

Алгоритм детектирования наличия голоса, предложенный в данной статье, показан на рис. 1. Он состоит из следующих шагов:

1. С помощью микрофонной решетки осуществляется захват аппаратно-синхронизированного многоканального звука кадрами фиксированной длины, включающими в себе непрерывную последовательность измерений с каждого микрофона массива.

2. Вычислительный модуль, используя алгоритм "взвешенный GCC-PHAT", оценивает азимуты на активные источники звука в системе координат массива микрофонов.

3. Направление на активные источники звука уточняется с помощью калмановской фильтрации [10].

4. Параллельно с захватом звука захватывается видеоизображение с цифровой видеокамеры.

5. Графический вычислительный модуль с помощью сверточных глубоких нейронных сетей [11] находит лицо человека на изображении.

6. Вычислительный модуль находит положение губ внутри области, соответствующей лицу, с помощью классификатора, описанного в работе [12].

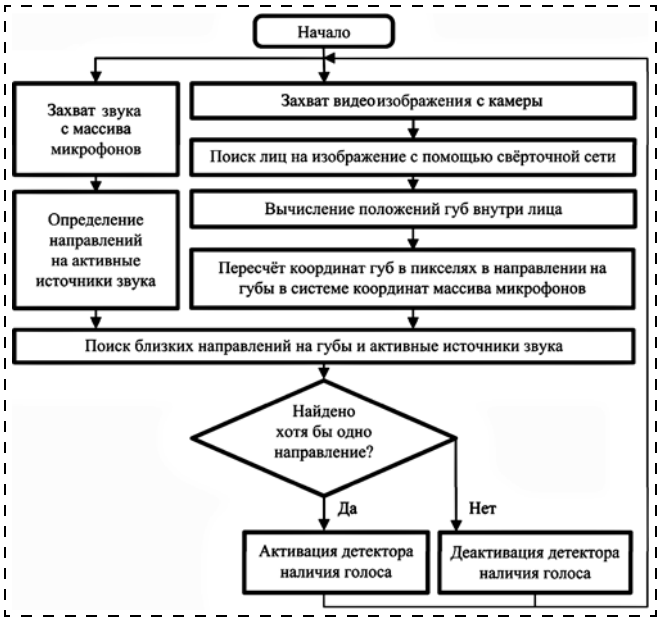

Рис. 1. Аудиовизуальный детектор наличия голоса

7. Вычислительный модуль, используя заранее известные оптические параметры камеры (фокусное расстояние, координату пересечения главной оптической оси с матрицей камеры, соотношение сторон одного пикселя матрицы камеры) и коэффициенты радиального и тангенциального искажений, высчитывает направление на обнаруженные губы в системе координат камеры. Параметры камеры предварительно определяются с помощью процедуры калибровки [13].

8. Решение об активации детектора голосовой активности принимается только в случае нахождения соответствующих направлений на губы и активный источник речи.

Длина звукового кадра для анализа подбирается так, чтобы статистические параметры сигнала можно было считать постоянными. На данный момент она составляет 256 измерений при частоте дискретизации 16 кГц. Обычно длина кадра составляет от одного до нескольких десятков миллисекунд.

Архитектура сверточной сети, использующейся для поиска губ, представлена на рис. 2. Нейронная сеть принимает на вход изображение и с помощью пирамиды Гаусса представляет его в разных масштабах, создавая новое расширенное изображение. Все дальнейшие операции осуществляются над новым изображением. Такое преобразование увеличивает вычислительную сложность, но обеспечивает инвариантность к масштабу.

Нейронная сеть была обучена на "DLib face detection dataset", в состав которого входит 7213 изображений человеческих лиц.

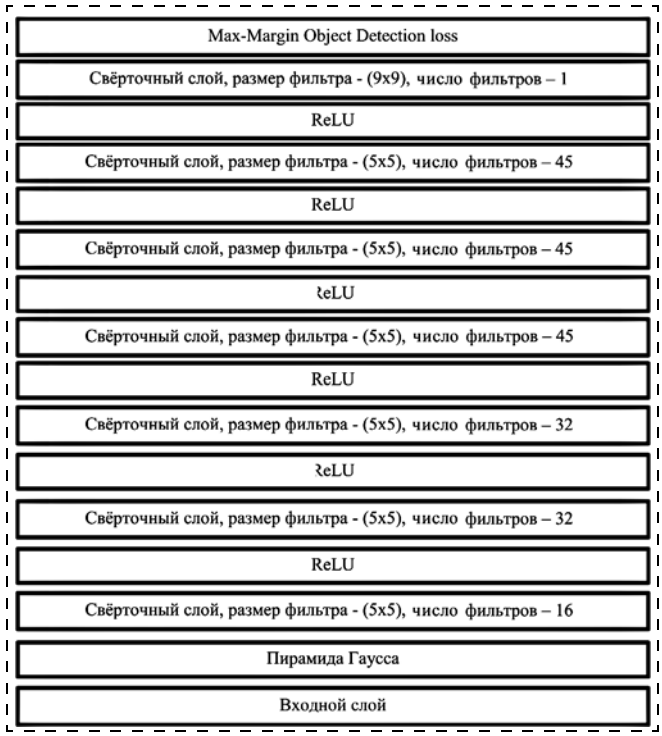

Рис. 2. Архитектура нейронной сети, используемой для поиска губ на изображениях

Определение направлений на активные источники звука с помощью алгоритма "взвешенный GCC-PHAT" продемонстрировано ниже.

Вначале проводится умножение каждого канала захваченного кадра на окно Ханна [14] и его дискретное преобразование Фурье:

$$X_m^l[k] = \sum_{n=0}^{N-1} \omega[n] x_m[l\Delta N + n]\mathbf{e}^{-j2\pi kn/N}, \quad (3)$$

где ω — вектор, содержащий коэффициенты окна Ханна; x_m — вектор измерений звукового сигнала; k — индекс частоты.

Взвешенный GCC-PHAT вычисляется следующим способом:

$$R_{pq}^l[n] = \frac{1}{N}\sum_{k=0}^{N-1} \frac{\zeta_{pq}^l[k] X_p^l[k] X_q^l[k]^*}{\left|X_p^l[k]\right|\left|X_q^l[k]\right| + \varepsilon} \mathbf{e}^{j2\pi kn/N}, \quad (4)$$

где $R_{pq}^l[n]$ — результат кросс-корреляции; $\zeta_{pq}^l[k]$ — частотная маска, необходимая для подавления влияния шумов; ε — коэффициент для избегания деления на ноль.

Далее осуществляется отображение значения $R_{pq}^l[n]$ из интервала $[0, N-1]$ в интервал $[-N/2 + 1, N/2]$:

$$\hat{R}_{pq}^l[n] = R_{pq}^l[n \bmod N], \quad (5)$$

затем вычисляется индекс, соответствующий максимальному значению кросскорреляции, который однозначно пересчитывается в направление на источник звука табличным способом:

$$\tau_{pq}^l = \operatorname{argmax}(\hat{R}_{pq}^l[n]), -n_{pq}^{\max} \leqslant n \leqslant n_{pq}^{\max}; \quad (6)$$

$$n_{pq}^{\max} = \frac{f_s}{c}\|x_p - x_q\|, \quad (7)$$

где f_s — частота дискретизации; c — скорость звука; x_p и x_q — положения микрофонов p и q.

В случае наличия нескольких активных источников звука направление на каждый из них будет соответствовать локальному максимуму кросс-корреляции.

Частотная маска $\zeta_{pq}^l[k]$ вычисляется на основе соотношения сигнал-шум в каждом канале:

$$|X_{pq}^l[k]|^2 = |X_p^l[k]|^2 |X_q^l[k]|^2; \quad (8)$$

$$Y_{pq}^l[k] = \frac{1}{2W+1}\sum_{\Delta k=-W}^{W} \log(|X_{pq}^l[k+\Delta k]|^2 + \varepsilon), \quad (9)$$

где $(2W + 1)$ — это размер прямоугольного окна, используемого для оценки шума, ε — коэффициент для избегания деления на ноль.

Затем вычисляется разность между текущим уровнем мощности и минимальным уровнем мощности ΔA в предыдущих кадрах в буфере:

$$A_{pq}^l[k] = Y_{pq}^l[k] - \min\{Y_{pq}^{l-\Delta A}[k], \ldots, Y_{pq}^l[k]\}. \quad (10)$$

После этого вычисляется разность между текущим уровнем мощности и минимальным уровнем мощности ΔB в будущих кадрах:

$$B_{pq}^l[k] = Y_{pq}^l[k] - \min\{Y_{pq}^l[k], \ldots, Y_{pq}^{l+\Delta B-1}[k]\}. \quad (11)$$

Эта операция добавляет небольшую задержку локализации, так как необходимо знать будущие значения звуковых измерений:

$$D_{pq}^l[k] = \begin{cases} 1\ (A_{pq}^l[k] > \Theta_A) \wedge (B_{pq}^l[k] > \Theta_B); \\ 0 \text{ иначе}, \end{cases} \quad (12)$$

где Θ_A и Θ_B — заранее определенные пороговые значения.

Затем вычисляется частотная маска:

$$\zeta_{pq}^l[k] = \begin{cases} 1 \sum_{l'=l}^{l+\Delta D} D_{pq}^{l'}[k] > \Theta_D; \\ 0 \text{ иначе.} \end{cases} \quad (13)$$

Оптическая модель камеры [12], используемая для пересчета пиксельных координат губ в углы в системе координат камеры, приведена ниже:

$$\begin{pmatrix} x \\ y \\ \omega \end{pmatrix} = \begin{pmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}; \quad (14)$$

$$\begin{cases} x' = x(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) + 2p_1 xy + p_2(r^2 + 2x^2); \\ y' = y(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) + p_1(r^2 + 2y^2) + 2p_2 xy, \end{cases} \quad (15)$$

где $(x, y)^{\mathrm{T}}$ — координаты губ в пикселях на матрице камеры для идеальной модели линзы;

$(x', y')^{\mathrm{T}}$ — координаты губ в пикселях на матрице камеры для реальной модели линзы, учитывающей радиальные и тангенциальные искажения;

$(X, Y, Z)^{\mathrm{T}}$ — координаты губ в системе координат камеры;

f_x, f_y — фокусные расстояния;

c_x, c_y — координаты оптического центра линзы в пикселях в системе координат матрицы;

$(k_1, k_2, k_3)^{\mathrm{T}}$ — коэффициенты радиального искажения;

$(p_1, p_2)^{\mathrm{T}}$ — коэффициенты тангенциального искажения.

Когда алгоритм определяет губы с пиксельными координатами $(x', y')^{\mathrm{T}}$, им соответствует бесконечное число положений губ в системе координат камеры $(X, Y, Z, 1)^{\mathrm{T}}$, которые являются решением системы (15). Все указанные решения лежат на одной прямой L. Направляющий вектор прямой L указывает направление на губы. Прямая L проходит через оптический центр камеры с координатами $P_0 = (0, 0, 0, 1)^{\mathrm{T}}$ в системе координат камеры. Другая точка $P_1 = (X, Y, Z, 1)$ может быть найдена из уравнения (14) и предположения, что $Z = 1$.

Аппаратная реализация

На базе ранее разработанного авторами линейного восьмиканального массива MEMS микрофо-

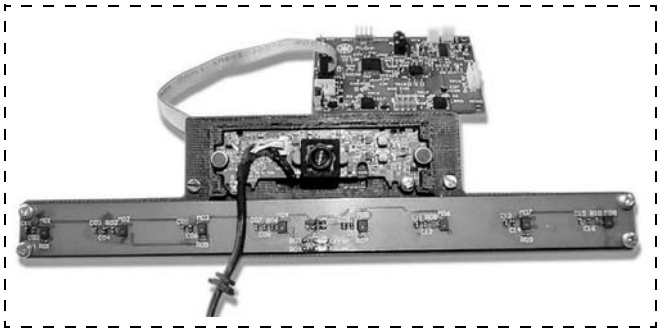

Рис. 3. Версия массива микрофонов со встроенной видеокамерой

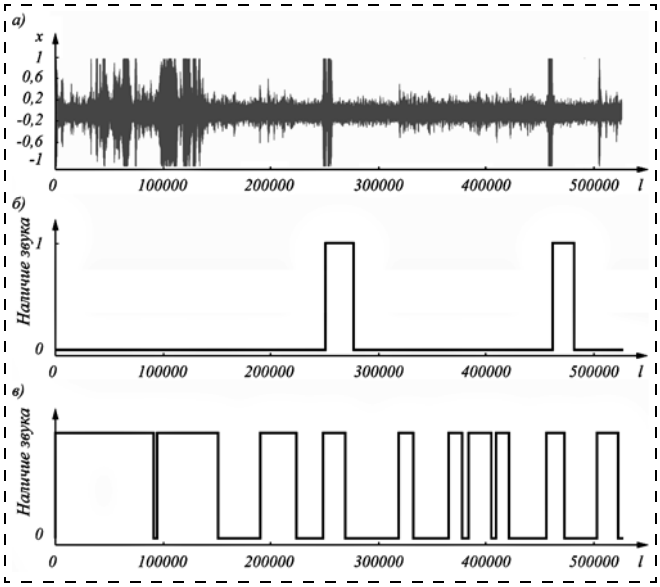

Рис. 4. Результаты сравнения разработанной системы с детектором голоса, использующим только звуковые данные (горизонтальная ось — номер измерения):
а — звуковой сигнал с первого микрофона; *б* — результат детектирования голосовой активности с помощью разработанного метода; *в* — результат детектирования голосовой активности с помощью детектора голоса, использующего только голосовые данные

нов с PDM-интерфейсом [15] был спроектирован в SolidWorks и изготовлен прототип устройства для аудиовизуального детектирования голосовой активности, представленный на рис. 3. Прототип состоит из трех печатных плат:

- восьмиканального линейного массива микрофонов с расстоянием между микрофонами 3 см (нижняя плата на рис. 3);
- платы захвата звука с восьмиканального массива микрофонов (вверху на рис. 3);
- платы видеокамеры с объективом (в середине рис. 3).

Прототип по USB-шине подключается к компьютеру с видеокартой с поддержкой технологии CUDA, на котором проводятся все вычисления.

Результаты тестирования

Было проведено сравнение качества работы разработанной системы в сильно зашумленных условиях и детектора голосовой активности из проекта с открытым исходным кодом WebRTC, который также использовал восьмиканальный линейный массив микрофонов, но не использовал данные от видеокамеры. Алгоритм детектирования голосовой активности из WebRTC работал следующим образом:

- формирование диаграммы направленности;
- постфильтрация;
- шумоподавление;
- детектирование наличия голоса.

Сравнение проводили при наличии постоянного фонового источника шума, а также при наличии голосового источника, который находился вне поля зрения видеокамеры и рассматривался как нецелевой. На верхнем графике рис. 4 этот источник звучит первую четверть времени проведения испытаний. Целевой источник появляется дважды и верно детектируется обоими детекторами наличия голоса. Однако детектор наличия голоса, использующий только голосовые данные, имеет очень большое число ложных срабатываний из-за высокой зашумленности обстановки — 88 % времени его активации является ложным в проведенном эксперименте.

Заключение

Разработанный метод детектирования голосовой активности показал высокое качество разметки голосового сигнала, которое при соблюдении условия нахождения говорящего человека в поле видимости видеокамеры превышает результаты, показываемые детекторами голоса, использующими только звуковую информацию. Разработанный детектор пригоден для реальных промышленных применений в голосовых пользовательских интерфейсах и робототехнике. В дальнейшем для повышения точности детектирования голоса необходимо проработать калмановскую фильтрацию, использующую данные с массива микрофонов и видеокамеры. Для снижения вычислительной сложности алгоритма и соответственно его системных требований необходимо упростить архитектуру нейронной сети, используемой для обработки видео.

Список литературы

1. **Ramнrez J., Górriz J. M., Segura J. C.** Voice activity detection. Fundamentals and speech recognition system robustness // Robust Speech Recognition and Understanding. Vienna: I-TECH Education and Publishing. 2007. P. 1—22.
2. **Woo K., Yang T., Park K., Lee C.** Robust voice activity detection algorithm for estimating noise spectrum // Electronics Letters. 2000. Vol. 36, N. 2. P. 180—181.
3. **Mousazadeh S., Cohen I.** Voice activity detection in presence of transient noise using spectral clustering // IEEE Trans. Audio, Speech, Language Process. 2013. Vol. 21, N. 6. P. 1261—1271.
4. **Obuchi Y.** Framewise speech-nonspeech classification by neural networks for voice activity detection with statistical noise suppression // IEEE International Conference on Acoustics, Speech and Signal Processing. Shanghai, 2016, March. P. 5715—5719.
5. **Montazzolli S., Jung C. R., Gelb D.** Audiovisual voice activity detection using off-the-shelf cameras // IEEE International Conference on Image Processing. Quebec, 2015, September. P. 3886—3890.
6. **Ying D., Yan Y., Dang J., Soong F. K.** Voice activity detection based on an unsupervised learning framework // IEEE Trans. Audio, Speech, Language Process. 2011. Vol. 19, N. 8. P. 2624—2633.
7. **Popović B., Pakoci E., Pekar D.** Advanced Voice Activity Detection on Mobile Phones by Using Microphone Array and Phoneme-Specific Gaussian Mixture Models // SISY. Subotica, 2016, August. P. 45—50.

8. **Grondin F., Michaud F.** Noise Mask for TDOA Sound Source Localization of Speech on Mobile Robots in Noisy Environments // IEEE International Conference Robotics and Automation. Stockholm, 2016, May.

9. **Tashev I., Mirsamadi S.** DNN-based Causal Voice Activity Detector // Information Theory and Applications Workshop. San Diego, 2016, February.

10. **Julier S., Uhlmann J.** A new extension of the Kalman filter to nonlinear systems // 11th International Symposium on Aerospace/ Defense Sensing, Simulation and Controls. Vol. Multi- Sensor Fusion, Tracking and Resource Management II. Orlando, 1997.

11. **King D. E.** Max-Margin Object Detection // Cornell University Library. 31.12.15. URL: https://arxiv.org/pdf/1502.00046.pdf (дата обращения: 18.08.2017).

12. **Kazemi V., Sullivan J.** One Millisecond Face Alignment with an Ensemble of Regression Trees // IEEE Conference on Computer Vision and Pattern Recognition. Columbus, 2014, June.

13. **Bradski G., Kaehler A.** Learning OpenCV. Computer Vision with the OpenCV Library. Sebastopol: O'Reilly Media, 2008. P. 580.

14. **Tashev I.** Sound Capture and Processing. Practical Approaches. The City of New York: John Wiley & Sons, 2009. P. 365.

15. **Суворов Д. А., Жуков Р. А.** Устройство синхронного сбора данных с массива MEMS микрофонов с PDM интерфейсом. Патент России № 172596. 2017. Бюл. № 20.

Audiovisual Voice Activity Detector Based on Deep Convolutional Neural Network and Generalized Cross-Correlation

D. A. Suvorov, dmitry.suvorov@skolkovotech.ru, **R. A. Zhukov,** roman.zhukov@skolkovotech.ru,
D. O. Tsetserukov, d.tsetserukou@skoltech.ru,
Skolkovo Institute of Science and Technology, Moscow, 143026, Russian Federation,
S. L. Zenkevich, zenkev@bmstu.ru,
Bauman Moscow State Technical University, Moscow, 105005, Russian Federation

Corresponding author: **Suvorov Dmitry A.,** Ph. D., student,
e-mail: dmitry.suvorov@skolkovotech.ru

Accepted on August 20, 2017

This paper presents a voice activity detector (VAD) which uses the data from the compact linear microphone array and a video camera, so developed VAD is robust to external noise conditions. It is able to ignore non-speech sound sources and speaking persons located out of the area of the interest. A deep convolutional neural network processes images from the video camera for searching face and lips of the speaking person. It was trained using the Max-Margin Object Detection loss. Pixel coordinates of found lips are converting to directions to lips in camera coordinate system using optical camera model. The sound from the microphone array is processing using the weighted GCC-PHAT algorithm and Kalman filtering. VAD searches for speaking lips on the video. It becomes activated only if the video camera finds lips and the microphone array confirms that there is a sound source in this direction. A prototype of the system based the linear microphone array with 30 mm spacing between microphones and the video camera was developed, manufactured using a 3D printer and tested in the laboratory conditions. The accuracy of the system was compared with the open source VAD from the WebRTC project (developed by Google) which uses only audio features extracted from the same microphone array. Developed VAD showed a high sustainability to external noise. It ignored the noise from not-target directions during 100 % of the testing time. And the VAD from the WebRTC had 88 % of false positive activations.

Keywords: *voice activity detector, microphone array, convolutional networks*

Acknowledgments

The research was partially implemented due to the grant from the Fund for the Promotion of Innovation (project No. 102GRNTIS5/26071).

For citation:

Suvorov D. A., Zhukov R. A., Tsetserukov D. O., Zenkevich S. L. Audiovisual Voice Activity Detector Based on Deep Convolutional Neural Network and Generalized Cross-Correlation, *Mekhatronika, Avtomatizatsiya, Upravlenie*, 2018, vol. 19, no. 1, pp. 53—57

DOI: 10.17587/mau.19.53-57

References

1. **Ramнrez J., Górriz J. M., Segura J. C.** Voice activity detection. Fundamentals and speech recognition system robustness, *Robust Speech Recognition and Understanding*, Vienna, I-TECH Education and Publishing, 2007, pp. 1—22.

2. **Woo K., Yang T., Park K., Lee C.** Robust voice activity detection algorithm for estimating noise spectrum, *Electronics Letters*, 2000, vol. 36, no. 2, pp. 180—181.

3. **Mousazadeh S., Cohen I.** Voice activity detection in presence of transient noise using spectral clustering, *IEEE Trans. Audio, Speech, Language Process*, 2013, vol. 21, no. 6, pp. 1261—1271.

4. **Obuchi Y.** Framewise speech-nonspeech classification by neural networks for voice activity detection with statistical noise suppression, *IEEE International Conference on Acoustics, Speech and Signal Processing*, Shanghai, March 2016, pp. 5715—5719.

5. **Montazzolli S., Jung C. R., Gelb D.** Audiovisual voice activity detection using off-the-shelf cameras, *IEEE International Conference on Image Processing,* Quebec, September 2015, pp. 3886—3890.

6. **Ying D., Yan Y., Dang J., Soong F. K.** Voice activity detection based on an unsupervised learning framework, *IEEE Trans. Audio, Speech, Language Process*, 2011, vol. 19, no. 8, pp. 2624—2633.

7. **Popović B., Pakoci E., Pekar D.** Advanced Voice Activity Detection on Mobile Phones by Using Microphone Array and Phoneme-Specific Gaussian Mixture Models, *SISY*, Subotica, August 2016, pp. 45—50.

8. **Grondin F., Michaud F.** Noise Mask for TDOA Sound Source Localization of Speech on Mobile Robots in Noisy Environments, *IEEE International Conference Robotics and Automation,* Stockholm, May 2016.

9. **Tashev I., Mirsamadi S.** DNN-based Causal Voice Activity Detector, *Information Theory and Applications Workshop*, San Diego, February 2016.

10. **Julier S., Uhlmann J.** A new extension of the Kalman filter to nonlinear systems, *11th International Symposium on Aerospace/Defense Sensing, Simulation and Controls. Vol. Multi- Sensor Fusion, Tracking and Resource Management II*, Orlando, 1997.

11. **King D. E.** Max-Margin Object Detection, *Cornell University Library*. 31.12.15, available at: https://arxiv.org/pdf/1502.00046.pdf (date of access: 18.08.2017).

12. **Kazemi V., Sullivan J.** One Millisecond Face Alignment with an Ensemble of Regression Trees, I*EEE Conference on Computer Vision and Pattern Recognition*, Columbus, June 2014.

13. **Bradski G., Kaehler A.** Learning OpenCV. Computer Vision with the OpenCV Library, Sebastopol, O'Reilly Media, 2008, 580 p.

14. **Tashev I.** Sound Capture and Processing. Practical Approaches, The City of New York, John Wiley & Sons, 2009, 365 p.

15. **Suvorov D. A., Zhukov R. A.** *Ustrojstvo sinhronnogo sbora dannyh s massiva MEMS mikrofonov s PDM interfejsom* (Device for synchronous data capturing from the array of MEMS microphones with PDM interface), Russian patent № 172596, 2017, Bul. № 20 (in Russian).