РАЗРАБОТКА МЕТОДА ШУМООЧИСТКИ РЕЧЕВЫХ СИГНАЛОВ ДЛЯ УЛУЧШЕНИЯ КАЧЕСТВА БИОМЕТРИЧЕСКОЙ ГОЛОСОВОЙ ВЕРИФИКАЦИИ

А.А. Лепендин, Р.С. Насретдинов
Алтайский государственный университет, Барнаул
email: andrey.lependin@gmail.com

Аннотация. Системы верификации дикторов в последнее время получили широкое применение в большом спектре информационных систем. Этот способ подтверждения личности является крайне удобным, поскольку для регистрации речевых образцов необходим лишь микрофон, имеющийся по умолчанию в большинстве электронных устройств. Однако качество работы подобных систем существенно снижается, когда речевой образец был записан в зашумленных условиях. В данной работе предложена новая модель шумоочистки на основе рекуррентных нейронных сетей, которая была апробирована для задачи верификации дикторов. Разработанный подход продемонстрировал на наборе данных DNS Challenge 2020 лучшее качество очистки от шума в сравнении с альтернативными. Он позволил существенно уменьшить уровень ошибок модельной системы верификации дикторов на тестовом наборе данных VoxCeleb1.

Ключевые слова: биометрическая верификация, улучшение качества речи, подавление шума, глубокое обучения, рекуррентная нейронная сеть.

Введение. В настоящее время растет популярность голосовых биометрических систем для верификации дикторов. Подобные системы являются одними из наиболее доступных, что объясняется несколькими причинами. Во-первых, для использования технологий определения личности человека по голосу не требуется дорогостоящего оборудования. Во-вторых, существует большое количество исследований в области обработки аудио с использованием методов машинного обучения и искусственного интеллекта, которые демонстрируют наилучшее качество работы в этих задачах.

Реальные экспериментально наблюдаемые ошибки при голосовой верификации личности могут оказываться существенно выше оценок этих ошибок на модельных данных и малых тестовых выборках. Одной из причин низкого качества является присутствие в речевых сигналах посторонних фоновых шумов. Очевидным способом решения данной проблемы является добавление зашумленных образцов в данные при обучении моделей верификации, но это не всегда приводит к существенному улучшению качества работы на новых зашумленных голосовых сигналах. Альтернативой является предварительная обработка сигналов с использованием отдельных моделей шумоочистки.

В данной работе предложен новый подход к очистке зашумленного одноканального речевого сигнала на основе рекуррентной нейросетевой модели, предназначенный для предварительной обработки речи в системе голосовой верификации.

Постановка задачи. Верификация дикторов – это задача, в которой система определяет, произносился ли входной речевой образец заявленным человеком [1, с. 5]. На этапе тестирования такой системы из речевых образцов пользователей извлекается пара векторов признаков: один извлекается из регистрируемого речевого образца, а второй обычно усредняется по набору образцов из некоторой базы для конкретного проверяемого пользователя. Далее между этими векторами вычисляется мера схожести, на основе значения которой и происходит верификация пользователя. При зашумлении речевых сигналов качество верификации может существенно снижаться. Возможным решением данной проблемы является

предварительное использование некоторого метода шумоочистки к обрабатываемым речевым образцам (рисунок 1).

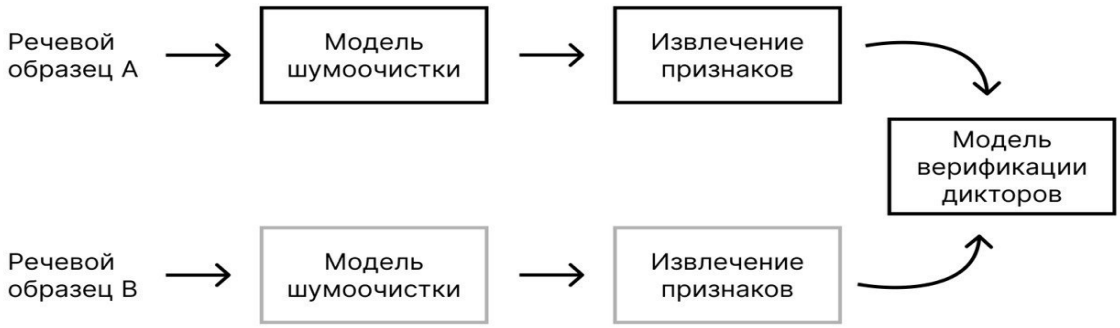

Рисунок 1. Система верификации дикторов с использованием модели шумоочистки.

Предлагаемый метод шумоочистки. Для представления входного образца зашумленной речи $X(t)$ использовалась модель аддитивного шума [2, с. 93]:

$$\widehat{X}(t,\ f)\ = \widehat{S}(t,\ f)\ +\ \widehat{N}(t,\ f), \qquad (1)$$

где t – дискретное время (от 0 до $T-1$), f – дискретная частота (от 0 до $F-1$), $\widehat{S}(t,\ f)$, $\widehat{X}(t,\ f)$ и $\widehat{N}(t,\ f)$ – частотно-временные представления чистого $S(t)$, зашумленного $X(t)$ сигналов и аддитивной шумовой добавки $N(t)$ соответственно. В данной работе не учитывались мультипликативные искажения сигнала, такие как реверберация.

Предлагаемая в работе нейросетевая модель для шумоочистки в качестве входных данных принимала последовательные действительные магнитуды преобразования Фурье входного сигнала, представленные в виде вектора $\vec{X} = [|X(t,0)|,\ |X(t,1)|, \dots,\ |X(t,F-1)|]^T$ длины F. Они вычислялись по всему диапазону дискретных частот для каждого временного окна с индексом t. Для обработки векторов $\vec{X}$ использовалась схема, представленная на рисунке 2.

Первый этап обработки включал в себя преобразование сигнала на всем диапазоне частот. Основной целью данного этапа было извлечение широкополосных спектральных паттернов в сигнале. Для этого входной вектор $\vec{X}$ подавался на вход нейросетевой модели G_{full}, которая представляла собой двуслойную LSTM-сеть [3, с. 242] с размером скрытого слоя 512 элементов:

$$G_{full}(\vec{X})\ = \mathrm{ReLU}\left(\mathrm{FC}\left(\mathrm{LSTM}(\vec{X})\right)\right), \qquad (2)$$

где $\mathrm{LSTM}(x)$ обозначена двуслойная LSTM-сеть, FC – полносвязный нейронный слой [3, с. 105], а ReLU – функция активации типа rectified linear unit [3, с. 107].

Во время второго этапа обработки извлекалась информация об узкополосных деталях спектра речевого сигнала. Его целью являлись как учет стационарных составляющих полезного сигнала, так и более точное выделение шумовой составляющей. Входные данные для второго этапа представляли собой результат конкатенации разбитых на перекрывающиеся полосы частот входных векторов $\vec{X}$ и $G_{full}(\vec{X})$ (на рисунке 2 операции разбиения обозначены как P_X и P_Y). Ширина частотных полос составляла $2n+1$ отсчетов с центральными частотами $f = 0, \dots, F-1$, где n равнялось 15 для исходного вектора данных $\vec{X}$ и 1 для вектора $G_{full}(\vec{X})$. Для граничных полос, индексы отсчетов в которых находились вне диапазона $0 \le f \le F-1$, значения частот продолжались циклически.

Более конкретно, результатом преобразования P_X к вектору отсчетов $\vec{X}$ частотно-временного представления сигнала являлась матрица:

$$P_X(\vec{X}) = \begin{bmatrix} \vec{X}(F-15) & \vec{X}(F-14) & \dots & \vec{X}(15) \\ \vec{X}(F-14) & \vec{X}(F-13) & \dots & \vec{X}(16) \\ \vdots & \vdots & \ddots & \vdots \\ \vec{X}(F-16) & \vec{X}(F-15) & \dots & \vec{X}(14) \end{bmatrix} \quad (3)$$

размера $F \times 31$, а для векторов $\vec{Y} = G_{full}(\vec{X})$ результатом применения P_Y была матрица размера $F \times 3$ вида:

$$P_Y(\vec{Y}) = \begin{bmatrix} \vec{Y}(F-1) & \vec{Y}(0) & \vec{Y}(1) \\ \vec{Y}(0) & \vec{Y}(1) & \vec{Y}(2) \\ \vdots & \vdots & \vdots \\ \vec{Y}(F-2) & \vec{Y}(F-1) & \vec{Y}(0) \end{bmatrix} \quad (4)$$

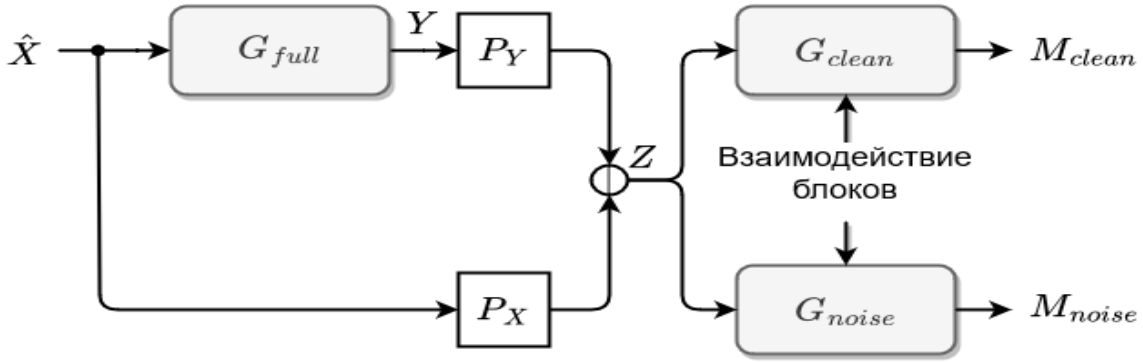

Рисунок 2. Предлагаемая архитектура нейронной сети для подавления шума.

Представления $P_X(\vec{X}), P_Y(\vec{Y})$ конкатенировались вдоль первой размерности в результирующую матрицу $Z = P_X(\vec{X}) \mid P_Y(\vec{Y})$ размера $F \times (31 + 3) = F \times 34$. Полученные значения Z подавались одновременно на две LSTM-сети G_{clean}и G_{noise} с независимо подбираемыми при обучении параметрами. Эти сети предсказывали маски M_{clean} и для M_{noise}, при помощи которых вычислялись полезная и шумовая составляющие входного сигнала.

Сети G_{clean}и G_{noise} представляли собой двуслойные LSTM-сети с размерностью скрытого слоя 384 и взаимодействием между слоями друг друга. Выходы каждого внутреннего слоя первой LSTM-сети смешивались с выходами во второй сети и наоборот. Обмен информацией между сетями G_{clean}и G_{noise} позволял подавлять нежелательные шумы в сигнале и в то же время более эффективно восстанавливать возникающие разрывы в формантах речевого спектра.

Смешивание выходов слоев LSTM осуществлялось следующим образом:

$$\bar{S}^i_{clean} = S^i_{clean} + S^i_{noise} \odot \sigma(BN(conv(S^i_{clean} \mid S^i_{noise}), \quad (5)$$

$$\bar{S}^i_{noise} = S^i_{noise} + S^i_{clean} \odot \sigma(BN(conv(S^i_{noise} \mid S^i_{clean}), \quad (6)$$

где S^i_{clean}, S^i_{noise} – выходы i-го (i = 1, 2) слоя LSTM-сети G_{clean} или G_{noise} соответственно, σ – сигмоидальная функция, $\odot$ - операция поэлементного умножения, BN – операция бэтч-нормализации [3, с. 153], а $\bar{S}^i_{clean}$ и $\bar{S}^i_{noise}$ – значения выходов после взаимодействия.

Сети G_{clean}и G_{noise} вычисляли мультипликативные маски для восстановления чистого сигнала и шума следующим образом:

$$M_{clean} = FC\left(LSTM_{inter}(Z, S^1_{noise}, S^2_{noise})\right), \quad (7)$$

$$M_{noise} = FC\left(LSTM_{inter}(Z, S^1_{clean}, S^2_{clean})\right), \quad (8)$$

где $LSTM_{inter}$ обозначена сеть LSTM с представленными выше блоками взаимодействия.

Обучение предложенной модели шумоочистки. Процесс обучения предложенной модели нейронной сети проводился на наборе триплетов: чистый сигнал $S(t)$, шум $N(t)$ и их сумма (входной зашумленный сигнал $X(t)$). Для каждого из этих триплетов заранее рассчитывалась маска комплексного идеального отношения (cIRM) [4, с. 487] для восстановления чистой речи:

$$\widetilde{M}_{clean} = (X_r S_r + X_i S_i)/(X_r^2 + S_i^2) + j(X_r S_i - X_i S_r)/(X_r^2 + X_i^2), \quad (9)$$

где индексы r и i обозначают действительную и мнимую составляющие преобразования Фурье сигнала $\hat{S}(t,f)$ и шума $\hat{N}(t,f)$, j – обозначает мнимую единицу. Аналогично рассчитывалась идеальная маска для выделения шума из сигнала $X(t)$. Для улучшения сходимости модели использовались сжатые представления идеальных масок cIRM [5, с. 90]:

$$\left(cIRM(\widetilde{M})\right)_x = K \cdot \left(1 - \exp\left(-C \cdot \widetilde{M}_x\right)\right) / \left(1 + \exp\left(-C \cdot \widetilde{M}_x\right)\right), \quad (10)$$

где x означает r или i - действительные или мнимые компоненты идеальной маски. Параметры сжатия, аналогично работе [5, с. 91], принимали значения $K = 10$ и $C = 0{,}1$.

Функция потерь состояла из двух компонентов, каждый из которых представлял собой среднеквадратичную ошибку (MSE) предсказания маски комплексного отношения. Первый член – это отклонения по маскам зашумленного сигнала M_{noise} со сжатыми идеальными масками, а второй – по маскам чистого сигнала M_{clean}:

$$\mathcal{L} = MSE\left(M_{clean}, cIRM\left(\widetilde{M}_{clean}\right)\right) + MSE\left(M_{noise}, cIRM\left(\widetilde{M}_{noise}\right)\right). \quad (11)$$

Используемые наборы данных. Для обучения модели шумоочистки использовался набор данных из конкурса по шумоочистке Microsoft DNS-Challenge 2020 [6, с. 2494]. Он состоял из примеров чистых речевых сигналов (около 500 часов речи от 2150 дикторов) и шумовых фрагментов (150 типов шумов по 10 секунд каждый). Для получения зашумленных образцов чистые сигналы и шум складывались с некоторым коэффициентом, подбираемым для различных уровней отношения сигнал / шум (SNR). SNR оценивался на фрагментах сигнала, в которых активны и речь, и шум. Это делалось для того, чтобы избежать превышения громкости при импульсных типах шума, таких как закрытие дверей, грохот, лай собаки и т. д. Длительность образцов зашумленных речевых сигналов составляла 30 с. В данной работе было использовано 60000 подобных образцов.

Для обучения и тестирования модели верификации дикторов использовался набор данных VoxCeleb1 [7, с. 2618], который содержал более 100 000 голосовых образцов 1251 дикторов, взятых из видеозаписей из хостинга YouTube. Видеозаписи снимались в большом количестве различных акустических сред: открытый стадион, студийные интервью, выступления на конференциях, любительские ролики, снятые на портативные устройства. Количество образцов, записанных мужчинами и женщинами, составляло 55 и 45 процентов соответственно. Спикеры выбирались из широкого спектра национальностей, видов акцента, профессий и возрастов. В записях присутствовали реальные шумы, которые состояли из фоновой речи, смеха, акустических характеристик помещения.

Метрики оценки качества. Для оценки качества модели шумоочистки использовались следующие метрики:

1. Оценка качества речи PESQ [8, с. 750], имитирующая «ручную» экспертную оценку по смещенной пятибалльной шкале от -0,5 (наименьшее качество) до 4,5 (наивысшее качество). Использовались два типа показателей PESQ: узкополосный (NB-PESQ), предназначенный для сигналов с частотой дискретизации 8 кГц, и широкополосный (WB-PESQ) для сигналов с частотой дискретизации 16 кГц;

2. STOI [9, с. 4215] – показатель разборчивости речи, измеряемый в процентах. Разборчивость оценивалась по опорному и искаженному передискретизированных сигналах с частотой 10 кГц на основе сравнения их спектральных характеристик;

3. Масштабно-инвариантное отношение сигнал/искажение SI-SDR [10, с. 627], являвшееся несмещенной модификацией обычного отношения сигнал-шум.

Для оценки качества работы системы верификации использовались стандартные для этой задачи метрики EER [11, с. 589] и минимальная функция стоимости обнаружения minDCF [11, с. 593]. EER – это коэффициент, равный вероятности отказа целевого диктора P_{fa} и вероятности пропуска самозванца P_{miss} в случае их равенства для выбранного порога принятия решения. Значение minDCF представляло собой минимум взвешенной суммы вероятностей P_{fa} и P_{miss} вида $DCF = 0{,}1P_{miss} + 0{,}01P_{fa}$, вычисляемый по всем пороговым значениям.

Результаты и обсуждение. Для реализации предложенного подхода использовался язык программирования Python 3.7, нейронные сети реализовывались с использованием библиотеки PyTorch 1.7. Обучение модели осуществлялось методом градиентного спуска с адаптивным вычислением моментов [3, с. 172] со стандартными параметрами. Остальные гиперпараметры представлены в таблице 1. Количество эпох обучения (полного прогона всех обучающих данных) составляло 44.

Таблица 1. Гиперпараметры нейросетевой модели для шумоочистки речевого сигнала.

Наименование гиперпараметра	Значение
Частота дискретизации, кГц	16
Число частотных полос F	512
Размер окна, отсчетов	512
Размер сдвига, отсчетов	256
Размер скрытого слоя LSTM в G_full	512
Размер скрытого слоя FFN в G_full	512
Размер скрытого слоя LSTM в G_clean и G_noise	384
Размер скрытого слоя FFN в G_clean и G_noise	384

В таблице 2 показаны результаты сравнения с современными моделями шумоочистки на тестовом подмножестве набора данных DNS Challenge 2020 (чем выше значение метрик PESQ, STOI и SI-SDR, тем лучше).

Таблица 2. Сравнение качества предложенной модели шумоочистки с аналогами.

Метод шумоочистки	Число параметров, $\times 10^6$	Задержка, мс	WB-PESQ	NB-PESQ	STOI, %	SI-SDR
Зашумленный сигнал	-	-	1,582	2,454	91,52	9,071
NSNet [11, с. 80]	5,1	32	2,145	2,873	94,47	15,613
FullSubNet [12, с. 2]	5,6	32	2,777	3,305	96,11	17,29
Предложенный метод	7,5	32	2,832	3,338	96,05	17,58

В верхней строке таблицы 2, помеченной как «Зашумленный сигнал», показаны значения метрик до применения моделей шумоочистки. Из таблицы видно,

что предлагаемая модель превосходит альтернативные подходы при сопоставимом количестве обучаемых параметров. При этом число параметров модели и задержка при обработке речевого сигнала была сопоставимой с другими методами.

Результаты использования модели шумоочистки в задаче верификации дикторов приведены в таблице 3. В качестве модели для выделения высокоуровневых признаков из речевого сигнала использовалась ResNetSE [13, с. 7134]. В ней использовались «squeeze and excitation» блоки, разработанные для усиления репрезентативной мощности модели, позволяя ей динамически выполнять поканальную перекалибровку выходных данных. Данная модель зарекомендовала себя в большом количестве исследовательских работ, демонстрируя высокую производительность на множестве наборов данных и задач.

К речевым образцам при оценке качества верификации добавлялись шумы из набора данных DNS-Challenge в диапазоне отношений сигнал-шум от -5 дБ до 20 дБ. Как можно увидеть из таблицы 3, использование предложенного метода значительно повышало качество верификации практически на всех уровнях зашумления.

Таблица 3. Качество верификации дикторов с использованием шумоочистки.

Уровень шума, дБ	До шумоочистки		После шумоочистки	
	EER, %	minDCF	EER, %	minDCF
-5	15,82	0,67	10,57	0,53
0	9,10	0,50	6,64	0,39
5	5,69	0,28	4,44	0,30
10	3,84	0,27	3,21	0,24
15	2,97	0,21	2,74	0,20
20	2,55	0,19	2,58	0,19

Выводы. В данной работе предложена и реализована новая нейросетевая архитектура для очистки речевых сигналов от фонового шума, основанная на рекуррентной нейронной сети с двумя выделенными блоками оценивания чистого голосового сигнала и шума. Предложенная архитектура имеет относительно небольшое количество параметров, малую задержку при оценке маски для восстановления сигнала, что позволяет ей работать в режиме реального времени. Качество восстановления очищенной речи превзошло существующие аналоги. Данная модель вносит слабые искажения в фазовые характеристики сигнала и подходит для широкого класса практических задач. Было проведено ее тестирование в задаче предварительной обработки зашумленных речевых образцов при верификации дикторов на тестовом наборе зашумленных голосовых сигналов. Использование модели позволило значительно улучшить качество работы системы верификации.

Исследование выполнено за счет гранта Российского научного фонда № 22–21–00199, https://rscf.ru/project/22-21-00199/.

Библиографический список.

1. Beigi H. Fundamentals of speaker recognition. – NY, Dordrecht, Heidelberg, London: Springer, 2011 – 942 с.

2. Loizou P.C. Speech Enhancement: Theory and Practice. – M.: Boca Raton. FL. USA: CRC Press, 2007. – 716 с.

3. Николенко С.И., Кадурин А.А., Архангельская Е.О. Глубокое обучение. – М., СПб. : Питер, 2018 – 480 с.

4. Williamson D.S., Wang Y., Wang D. Complex ratio masking for monaural speech separation // IEEE/ACM Transactions on Audio, Speech, and Language Processing. – 2016. –No 3 (24). – P. 483-492.
5. Nasretdinov R.S., Ilyashenko I.D., Lependin A.A. Two-stage method of speech denoising by long short-term memory neural network // 11th International Conference on High-Performance Computing Systems and Technologies in Scientific Research, Automation of Control and Production, HPCST 2021, Barnaul 21-22 May 2021. CCIS, Vol. 1526. – Springer, 2022. – P. 86-97.
6. Reddy, C.K.A., Gopal, V., Cutler, R., Beyrami, E., Cheng, R., Dubey, H., Matusevych, S., Aichner, R., Aazami, A., Braun, S., Rana, P., Srinivasan, S., Gehrke, J. The INTERSPEECH 2020 Deep Noise Suppression Challenge: datasets, subjective testing framework, and challenge results // Proc. Interspeech 2020. – 2020. – P. 2492-2496.
7. Nagrani A., Chung J.S., Zisserman A. VoxCeleb: a large-scale speaker identification dataset // Proc. Interspeech 2017 – 2017. – pp. 2616-2620.
8. Rix A.W., Beerends J.G., Hollier M.P. Hekstra A.P. Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs // 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. – 2001. – P. 749-752.
9. Taal C.H., Hendriks R.C., Heusdens R., Jensen J., A short-time objective intelligibility measure for time-frequency weighted noisy speech // 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. – 2010. – P. 4214-4217.
10. Roux J.L., Wisdom S., Erdogan H., Hershey J.R. SDR – half-baked or well done? // ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2019. – P. 626-630.
11. Braun S., Tashev I. Data augmentation and loss normalization for deep noise suppression // 22nd International Conference on Speech and Computer (SPECOM). LNAI 12335. – Springer, 2020. – P. 79–86.
12. Hao X. FullSubNet: A Full-Band and Sub-Band Fusion Model for Real-Time Single-Channel Speech Enhancement / X. Hao, X. Su, R. Horaud, X. Li // IEEE International Conference on Acoustics, Speech, and Signal Processing. - 2021. - P. 1-5.
13. Hu J., Shen L., Sun G. Squeeze-and-excitation networks // Proceedings of the IEEE conference on computer vision and pattern recognition. – 2018. – P. 7132-7141.

DEVELOPMENT OF A SPEECH SIGNALS NOISE CLEANING METHOD TO IMPROVE THE QUALITY OF BIOMETRIC VOICE VERIFICATION

A.A. Lependin, R.S. Nasretdinov
Altai state university, Barnaul
email: andrey.lependin@gmail.com

Annotation. Speaker verification systems have recently been widely used in a wide range of information systems. This method of identity verification is extremely convenient, since only a microphone, which is available by default in most electronic devices, is needed to register speech samples. However, the performance of such systems is significantly reduced when the speech sample was recorded in noisy environment. In this paper, a new speech enhancement model based on recurrent neural networks was proposed, which was tested for the problem of speaker verification. On the DNS Challenge 2020 data set, the developed approach demonstrated the best quality of noise removal in comparison with alternative approaches. It made it possible to significantly reduce the level of errors in the model system for verifying speakers on the VoxCeleb1 test data set.

Keywords: biometric verification, speech enhancement, noise cancelling, deep learning, recurrent neural network.