DOI: 10.15514/ISPRAS-2022-34(4)-10

Метод улучшения качества речи с использованием модифицированного кодирующего-декодирующего пирамидального трансформера

А.А. Лепендин, ORCID: 0000-0001-5097-5023 <andrey.lependin@gmail.com>
Р.С. Насретдинов, ORCID: 0000-0003-3368-523X <uniform97@gmail.com>
И.Д. Ильяшенко, ORCID: 0000-0001-5119-3832 <ilya-ilyash@yandex.ru>
Алтайский государственный университет,
656049, Россия, г. Барнаул, пр. Ленина, д. 61

Аннотация. Развитие новых технологий голосового общения привело к необходимости совершенствования методов улучшения качества речи. Современные пользователи информационных систем предъявляют высокие требования как к разборчивости голосового сигнала, так и к его субъективно воспринимаемому качеству. Данная работа посвящена развитию нового подхода к решению актуальной задачи улучшения качества речи. Для этого было предложено использовать модифицированную нейронную сеть пирамидального трансформера, использующую двухкомпонентную структуру «кодер-декодер». Кодирующая компонента сети осуществляла сжатие спектра голосового сигнала в пирамидальную серию внутренних представлений. Декодирующая компонента, используя преобразования самовнимания, восстанавливала маску комплексного отношения очищенного и искаженного сигналов на основе вычисленных кодером внутренних представлений. Были рассмотрены две возможные функции потерь для обучения предложенной нейросетевой модели. Показано, что использование частотного кодирования, подмешиваемого к входным данным, позволило улучшить качество работы предложенного подхода. Реализованная на языке Python и библиотеке глубокого обучения PyTorch нейронная сеть обучалась и тестировалась на наборе данных DNS Challenge 2021. Она продемонстрировала высокое качество работы по сравнению с другими современными методами улучшения качества речи. В работе был проведен качественный анализ процесса обучения реализованной нейронной сети, который показал, что предлагаемая нейросетевая модель постепенно переходила от простого маскирования шума на ранних эпохах обучения к восстановлению пропущенных формантных компонент голоса говорящего на более поздних эпохах. Это приводило к высоким значениям численных метрик качества работы предложенного подхода и высокому субъективному качеству речи.

Ключевые слова: улучшение качества речи; очистка от шума; маскирование шума; глубокая нейронная сеть; глубокое обучение; архитектура кодер-декодер; пирамидальный трансформер; самовнимание.

Для цитирования: Лепендин А.А., Насретдинов Р.С., Ильяшенко И.Д. Метод улучшения качества речи с использованием модифицированного кодирующего-декодирующего пирамидального трансформера. Труды ИСП РАН, том 34, вып. 4, 2022 г., стр. 135-152.. DOI: 10.15514/ISPRAS-2022-34(4)-10

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

Speech Enhancement Method Based on Modified Encoder-Decoder Pyramid Transformer

A.A. Lependin, ORCID: 0000-0001-5097-5023 <andrey.lependin@gmail.com>
R.S. Nasretdinov, ORCID: 0000-0003-3368-523X <uniform97@gmail.com>
I.D. Ilyashenko, ORCID: 0000-0001-5119-3832 <ilya-ilyash@yandex.ru>
Altai State University,
61, Lenin st., Barnaul, 656049, Russia

Abstract. The development of new technologies for voice communication has led to the need of improvement of speech enhancement methods. Modern users of information systems place high demands on both the intelligibility of the voice signal and its perceptual quality. In this work we propose a new approach to solving the problem of speech enhancement. For this, a modified pyramidal transformer neural network with an encoder-decoder structure was developed. The encoder compressed the spectrum of the voice signal into a pyramidal series of internal embeddings. The decoder with self-attention transformations reconstructed the mask of the complex ratio of the cleaned and noisy signals based on the embeddings calculated by the encoder. Two possible loss functions were considered for training the proposed neural network model. It was shown that the use of frequency encoding mixed with the input data has improved the performance of the proposed approach. The neural network was trained and tested on the DNS Challenge 2021 dataset. It showed high performance compared to modern speech enhancement methods. We provide a qualitative analysis of the training process of the implemented neural network. It showed that the network gradually moved from simple noise masking in the early training epochs to restoring the missing formant components of the speaker's voice in later epochs. This led to high performance metrics and subjective quality of enhanced speech.

Keywords: speech enhancement; noise reduction; noise masking; deep neural network; deep learning; encoder-decoder architecture; pyramid transformer; self-attention.

For citation: Lependin A.A., Nasretdinov R.S., Ilyashenko I.D. Speech Enhancement Method Based on Modified Encoder-Decoder Pyramid Transformer. Trudy ISP RAN/Proc. ISP RAS, vol. 34, issue 4, 2022, pp. 135-152 (in Russian). DOI: 10.15514/ISPRAS-2022-34(4)-10

Acknowledgements. This work was supported by the grant from the Russian Science Foundation, project no. 22-21-00199, https://rscf.ru/en/project/22-21-00199/.

1. Введение

В настоящее время технологии цифровой обработки и передачи речи получили широкое распространение. Растет частота использования голосовых помощников в мобильных устройствах и интеллектуальных колонках. Множество мобильных устройств, планшетов и персональных компьютеров используется для аудио- и видео-звонков, телеконференций. Естественным образом повышаются требования пользователей к качеству передаваемого речевого сигнала. В связи с этим в последние годы увеличилось число практико-ориентированных исследовательских работ и проектов, ориентированных на разработку новых методов повышения качества речи (speech enhancement) [1]. К числу требований, предъявляемым к новым разработкам следует отнести следующее.

- Эффективное подавление фоновых шумов в аудиосигнале, соответствующее сценариям повседневного использования. Современные пользователи часто работают в открытых общественных пространствах, на улице, в шумных помещениях, транспорте. Из-за этого сигнал содержит фрагменты речи окружающих данного пользователя людей, крики детей, природные шумы, звуки уличного транспорта. Большинство подобных шумов является нестационарными, их мощность может быть сравнима и даже превосходить мощность полезного речевого сигнала. Их спектральные характеристики также имеют широкий «разбег»: широкополосные сигналы уличного трафика, узкополосные фоновые шумы или сигналы со спектром сложной формы в случае речевых или речеподобных шумов;
- Улучшение качества должно не только повышать разборчивости речи. Это минимально

необходимое, но не достаточное условие. Особую важность представляет высокое субъективное качество улучшенной речи. Оно предполагает помимо разборчивости сохранение характерных интонационных и тембральных особенностей голоса.

В появлении новых, более эффективных методов заинтересованы крупные IT-компании, так как это становится еще одним «продающим» фактором для их программных продуктов. В частности, это подтверждается активностью компании Microsoft, уже на протяжении нескольких лет проводящей открытые соревнований в области алгоритмов улучшения качества речи – Deep Noise Suppression (DNS) Challenge [2, 3]. Крупнейшие научные конференции Interspeech и ICASSP, посвященные методам обработки речи, предоставляют свои площадки для обсуждения новых подходов к развитию методов улучшения качества речи.

Основной вектор развития новых подходов к улучшению качества речи связан в настоящее время с развитием современных методов глубокого обучения. Большинство работ последних лет предполагают использование глубоких нейросетевых моделей. Появляются представительные открытые наборы голосовых записей, образцов шумовых сигналов реального мира, модельных и измеренных импульсных характеристик помещений. Все это способствует быстрому прогрессу как в разработке новых нейросетевых моделей для обработки искаженной речи, так и в появлении новых специализированных методов их обучения.

В данной работе предложен новый подход к решению задачи улучшения качества зашумленной речи, основанный на применении глубокой нейронной сети типа трансформер.

2. Задача улучшения качества речи

Улучшение качества речи является одним из важных элементов современных телекоммуникационных систем. В зависимости от условий применения методы улучшения качества могут быть разделены на несколько типов. Выделяют адаптивные и неадаптивные подходы в зависимости от того, подстраиваются ли параметры используемых при улучшении речи методов под не стационарно меняющиеся полезный и шумовой компонент сигнала. В зависимости от числа микрофонов и их расположения выделяют одноканальное и многоканальное улучшение речи. Существующие многоканальные методы демонстрируют лучшее качество, однако в реальных сценариях использования число микрофонов не превосходит двух, расположены они относительно близко друг к другу, так что фактически их можно рассматривать как один канал. В работе решалась задача адаптивного одноканального улучшения качества речи.

2.1 Постановка задачи

При разработке предложенного метода предполагалось, что искажение, вносимое в одноканальный речевой сигнал, можно разделить на две части [1]. Первая, аддитивная, отвечала за присутствующие в записи фоновые шумы. Вторая, мультипликативная, определялась наличием реверберации. Искаженный голосовой сигнал представлялся следующей моделью:

$$Y(t) = S(t) * H(t) + N(t) = X(t) + N(t), \quad (1)$$

где $S(t)$ – чистый сигнал без искажений, $H(t)$ – импульсная характеристика помещения, применяемая для моделирования реверберации, $X(t)$ – сигнал с реверберацией, $N(t)$ – аддитивный шум, * – операция свертки. После разбиения сигнала на перекрывающиеся временные окна и быстрого преобразования Фурье, получившееся представление сигнала можно записать следующим образом:

$$Y(f,t) = S(f,t) \cdot H(f,t) + N(f,t) = X(f,t) + N(f,t), \quad (2)$$

где $Y(f,t)$, $S(f,t)$, $H(f,t)$, $X(f,t)$ – комплексное представление сигналов в частотно-временной области, индекс $f = 0, \dots, F-1$ отвечает за номер соответствующей частотной полосы, $t = 0, \dots, T-1$ – за номер временного окна сигнала.

В данной работе решалась исключительно задача восстановления сигнала X, который мог представлять собой как реверберированную версию чистого сигнала S, так и сам чистый сигнал S. Набор данных включал в себя как сигналы без реверберации, так и с ней. Все оценки качества работы также проводились двух версий сигнала x.

2.2 Основные методы улучшения качества речи

Одноканальное улучшение речи согласно [1] может производиться следующими классическими методами: спектральным вычитанием, статистическим моделированием, алгоритмами выделения подпространств и методами маскирования. Каждая из этих групп обладает своими достоинствами и недостатками, связанными как с ограничениями на типы подавляемого шума, так и с требуемой вычислительной мощностью. Основной объединяющей чертой этих методов в классической постановке является опора на явно или не явно используемые статистические модели аддитивного шума, что ограничивает их применение. Еще одним препятствием для использования подобных подходов является низкое субъективное качество результатов их работы.

В силу указанных ограничений в последние годы широкое распространение стали получать методы улучшения качества, основанные на глубоких нейронных сетях. Они в той или иной степени стали развитием перечисленных выше классических методов. Появились нейросетевые аналоги методов статистического моделирования [4], адаптивной фильтрации [5]. Однако наибольшее число работ [6-13] было связано с развитием методов маскирования шума. Этот подход в классической постановке [1] связан с разделением всех элементов частотно-временного представления зашумленного сигнала (2) на два вида – относящиеся к речи и относящиеся к шуму. Метод шумоочистки должен вычислить так называемую идеальную бинарную маску (ideal binary mask) вида [1]:

$$\mathrm{IBM}(f,t) = \begin{cases} 0, & Y(f,t) \text{ содержит аддитивный шум,} \\ 1, & Y(f,t) \text{ содержит полезный сигнал.} \end{cases} \quad (3)$$

Подобный подход получил дальнейшее развитие в виде нескольких модификаций методов маскирования. В работе [11] было предложено вычислять идеальные маски вещественного отношения амплитуд сигнала и шума (IRM), позволявшие учитывать относительный вклад шумовой добавки в отдельных участках спектра голосового сигнала и, более того, к возможности частичного восстановления полезной компоненты из шумового фона. Далее, метод идеальной маски комплексного отношения (CRM) [12] позволил учесть фазовую составляющую сигнала, что повысило не только разборчивость речи, но и субъективное качество речи.

Для вычисления масок в работах [6, 12, 13] применялся подход на основе кодирующей-декодирующей архитектуры. Кодирующая компонента нейронной сети сжимала представление сигнала, по этому представлению декодером вычислялась требуемая маска. В [13] для кодирования и декодирования применялись комбинации сверточных и рекурсивных слоев. Слои свертки позволяли эффективно сжимать частотно-временное представление обрабатываемого сигнала, а рекурсивные слои осуществляли моделирование временных зависимостей в нем. Недостатком такого подхода являлось следующее. Рецептивное поле преобразования свертки является хорошо локализованным, поэтому при вычислении сжатого представления спектра сигнала могли теряться связи между различными участками спектра сигнала. Рекурсивные слои, в свою очередь, обладают эффектом «забывания» внутренних

состояний [14]. В целом это приводило к неточному восстановлению голосовых формант и потере качества.

Новым подходом, позволяющим учитывать нелокальные связи между удаленными по частотной или временной оси участками голосового сигнала, явились архитектуры типа трансформер (transformer neural network), использующие механизм самовнимания (self-attention) [15]. Подобные нейронные сети нашли применение во всех основных задачах глубокого обучения – от обработки естественных языков [15] до автоматической сегментации изображений [16, 17], постепенно вытесняя предыдущие поколения глубоких нейронных сетей, основанные на преобразованиях свертки и рекуррентных слоях. В задачах обработки речи преобразование самовнимания применялось для автоматического распознавания голоса [18], оценки эмоциональной окраски [19] и синтеза речи [20]. В задаче улучшения качества речи данный механизм использовался как замена части рекуррентных и сверточных преобразований [9]. Принципиально новым в данной работе является использование механизма самовнимания на всех этапах обработки искаженного голосового сигнала.

3. Предлагаемый метод

Предложенная нейросетевая модель была основана на модификации сети пирамидального трансформера PVT (Pyramid Vision Transformer) [16]. На вход сети подавалась спектрограмма обрабатываемого сигнала $\mathrm{Y(f,t)}$. Сеть имела архитектуру «кодер-декодер» и позволяла предсказывать комплексную маску отношения очищенного и зашумленного сигнала $\mathrm{M(f,t)}$. С помощью вычисленной маски восстанавливалось комплексное частотно-временное представление очищенного сигнала $\widehat{\mathrm{X}}\mathrm{(f,t)}$:

$$\widehat{\mathrm{X}} = (\mathrm{Y_r} \odot \mathrm{M_r} - \mathrm{Y_i} \odot \mathrm{M_i}) + \mathrm{j}(\mathrm{Y_r} \odot \mathrm{M_i} + \mathrm{Y_i} \odot \mathrm{M_r}), \quad (4)$$

где индексами r и i обозначены вещественная и мнимые части комплексных значений, j – комплексная единица, $\odot$ - операция поэлементного умножения.

Таким образом, для зашумленного сигнала Y, процесс улучшения качества записывался следующим образом:

$$\widehat{\mathrm{X}} = \mathrm{ISTFT}\left(\mathrm{STFT(Y)} \odot \mathrm{M}\left(\mathrm{STFT(Y)}\right)\right), \quad (5)$$

где STFT(.) и ISTFT(.) представляли собой прямое и обратное оконное преобразование Фурье, а M(.) – представляло собой преобразование, вычисляемое предлагаемой нейронной сетью.

3.1 Сеть пирамидального визуального трансформера PVT

Сеть PVT являлась альтернативой кодирующей компоненты сверточных сетей типа U-Net [21], позволяющей решать задачи классификации изображений, детектирования объектов и семантической сегментации. Она обеспечивала последовательное вычисление нескольких промежуточных представлений входного изображения. При этом использовалась прогрессивная стратегия сжатия представлений [16]. Они образовывали «пирамиду» представлений, содержащую все более общие признаки входного изображения.

Главной отличительной особенностью сети PVT по сравнению с U-Net-подобными сверточными сетями было использование для понижения размерности признакового представления преобразования нового типа. В этом преобразовании сначала проводилось разбиение входного представления F размера $\mathrm{H \times W}$ с числом каналов C на $\mathrm{HW/P^2}$ непересекающихся участков (патчей) за счет применения свертки с совпадающими по величине размером фильтра ($\mathrm{P \times P}$) и величиной сдвига (stride=P). Тем самым вычислялись закодированные представления патчей (patch embeddings) размера $(\mathrm{HW/P^2}) \times \mathrm{C'}$, где $\mathrm{C'}$

обозначено новое число каналов. После к этим представлениям для учета взаимного расположения патчей друг относительно друга добавлялись обучаемые вектора позиционного кодирования. Далее применялся трансформер [15, 16], содержащий L последовательных преобразований самовнимания (self-attention). Выход трансформера представлялся в виде нового сжатого признакового представления $\mathrm{F'}$ размера $\frac{\mathrm{H}}{\mathrm{P}} \times \frac{\mathrm{W}}{\mathrm{P}} \times \mathrm{C'}$.

Сеть PVT содержала 4 последовательных блока преобразований, описанных выше. На вход этой цепочки преобразований подавалось цветное трехканальное изображение. В результате вычислялось пирамидальное представление изображения $\{\mathrm{F_1}, \mathrm{F_2}, \mathrm{F_3}, \mathrm{F_4}\}$. Именно элементы этого представления, все или по отдельности, использовались для решения перечисленных выше задач классификации, детектирования и сегментации.

3.2 Модификация сети PVT для задачи улучшения качества речи

Концептуально задачи шумоочистки и сегментации изображений похожи, так как размеры входа и выхода нейронной сети должны совпадать. Содержательное отличие, которое следует учитывать, заключалось в том, что в изображении обе размерности (по ширине и по высоте изображения) равноправны, в то время как в частотно-временном представлении они имеют разный смысл (номер частотной полосы и номер временного окна). Учитывая указанные соображения, в архитектуру сети PVT были внесены некоторые изменения.

3.2.1 Частотное кодирование

Входом модифицированной сети являлась комплексная спектрограмма $\mathrm{Y(f,t)}$, размера $F \times T \times 2$, где два канала отвечали за вещественную и мнимую компоненты. Перед вычислением пирамидального представления она конкатенировалась с матрицей частотного позиционного кодирования. Тем самым ко входному сигналу подмешивалась информация о том, какой частотной полосе соответствуют тот или иной участок спектрограммы. В качестве гипотезы предполагалось, что это могло помочь нейросетевой модели эффективно сопоставлять различные области спектра и поддерживать непрерывность голосовых формант при восстановлении речи.

Использовались K-мерные вектора кодирования частоты вида [6]:

$$\mathrm{p(f,k)} = \left[\cos\left(\frac{\pi \mathrm{f}}{\mathrm{F}}\right), \cos\left(\frac{2\pi \mathrm{f}}{\mathrm{F}}\right), \ldots, \cos\left(\frac{2^{K-1}\pi \mathrm{f}}{\mathrm{F}}\right)\right], \quad (6)$$

где F – это максимальная частота и $\mathrm{k} = 0, \ldots, K-1$ – индекс вектора кодирования. Они не зависели от времени и не являлись обучаемыми параметрам сети. Пример частотного кодирования размера K=10 для F=256 частотных полос представлен на рис. 1.

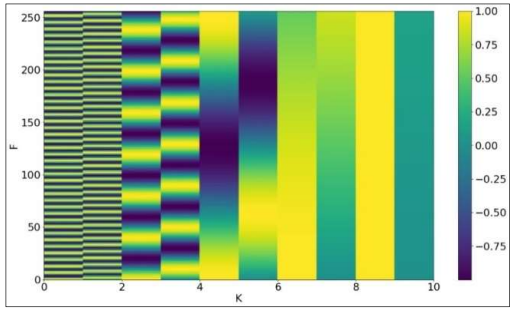

Рис. 1. Пример частотного кодирования с векторами размером K=10 для F=256 частотных полос
Fig. 1. An example of positional encoding with vectors of size K=10 for F=256 frequency bands

3.2.2 Разбиение на патчи и декодирование комплексной маски

На рис. 2 представлена схема предложенной нейронной сети с архитектурой «кодер-декодер». После применения входной свертки к спектрограммам с частотным кодированием они подавались на цепочку преобразований сжатия, обозначенных как B_1-B_4, являющихся по сути нейронной сетью PVT (выделены на рис. 2 зеленым). Были проведены несколько модификаций этих преобразований.

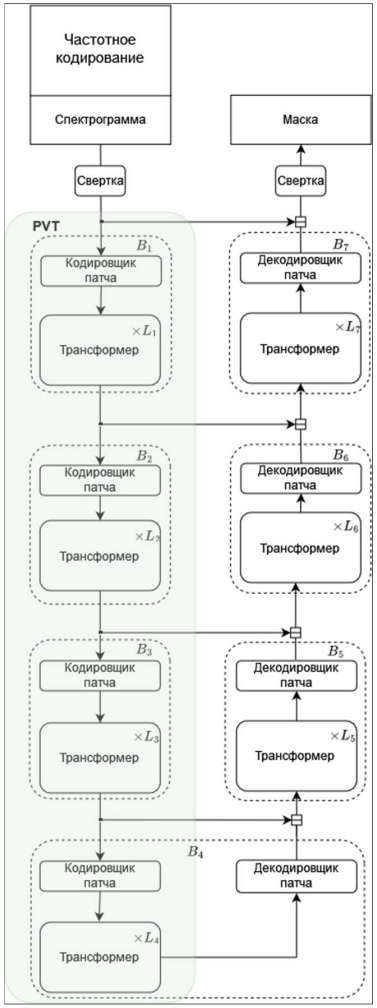

Рис. 2. Схема модифицированной нейронной сети для улучшения качества речи; зеленым выделена часть сети аналогичная пирамидальному кодировщику сети PVT.
Fig. 2. Scheme of the modified neural network for speech enhancement; the part of the network similar to the PVT network pyramidal encoder is highlighted in green.

1. Разбиение проводилось на прямоугольные патчи различных размеров. Требовалось учитывать больше информации о соседних частотных полосах сигнала и одновременно обеспечить достаточное сжатие входного частотно-временного представления для

ограничения числа параметров модели (размеры используемых патчей приведены ниже в табл. 1).

2. Последнее из преобразований сети PVT (B_4) было модифицировано. К нему был добавлен декодировщик патчей, который представлял собой транспонированную свертку, с последующей прибавкой обучаемого вектора позиционного кодирования. Следует отметить, что все вектора позиционного кодирования как кодировщике, так и в декодировщике были не зависимы друг от друга при обучении сети.

3. В отличие от задачи сегментации, был заменен используемый декодер. В работе [22] для развертывания пирамидального представления применялся декодер FPN, использующий последовательные сверточные слои. В данной работе был разработан оригинальный декодер, основанный на трансформер-преобразованиях, аналогичных применяемым в PVT. Для увеличения размера последовательных результатов в блоках B_5-B_7 использовалась упомянутая выше связка транспонированной свертки и позиционного кодирования. При этом на вход блоков B_5-B_7 подавался не только выход предыдущего блока, но его конкатенация с выходом соответствующего блока пирамидального преобразования PVT. Этот способ передачи дополнительной информации о представлениях сигнала разного масштаба был концептуально похож на применяемому в других кодирующих-декодирующих сетях.

3.3 Функции потерь

Обучение нейронных сетей для улучшения качества речи потребовало и подбора наиболее подходящей функции потерь. Входными данными для них являлись образцовый сигнал $\mathrm{X}(\mathrm{t})$ и улучшенный сигнал $\widehat{\mathrm{X}}(t)$, вычисленный согласно (4). Использовались отрицания нескольких существующих метрик оценки качества очистки от шума, так как минимизация значений этих метрик при обучении нейронной сети приводила бы к высокому достигнутому качеству работы. В численных экспериментах были апробированы следующие варианты функций потерь:

– отрицание отношения сигнал-шум (SNR):

$$\mathcal{L}(\widehat{\mathrm{X}},\mathrm{X}) = -\mathrm{SNR}(\widehat{\mathrm{X}},\mathrm{X}) = -10\log_{10}\left(\frac{\|\mathrm{X}\|^2}{\|\widehat{\mathrm{X}}-\mathrm{X}\|^2}\right); \quad (7)$$

– отрицание масштабно-инвариантного отношения сигнал-возмущение (SI-SDR) [23]:

$$\mathcal{L}(\widehat{\mathrm{X}},\mathrm{X}) = -\text{SI-SDR}(\widehat{\mathrm{X}},\mathrm{X}) = -10\log_{10}\left(\frac{\|\alpha\mathrm{X}\|^2}{\|\widehat{\mathrm{X}}-\alpha\mathrm{X}\|^2}\right), \quad (8)$$

где $\alpha = \frac{\widehat{\mathrm{X}}^T X}{\|\mathrm{X}\|}$ – коэффициент масштабирования, корректирующий амплитуду чистого сигнала, для того чтобы откорректировать завышенную оценку отношения сигнал-шум в случае не ортогональных векторов X и $\widehat{\mathrm{X}} - \mathrm{X}$.

4. Набор данных

Для апробации предложенного подхода и предварительной оценки его эффективности использовался обучающий набор, построенный по методике соревнований DNS Challenge 2021 года [2] с образцами с частотой дискретизации 16 кГц. Образцы из соревнований DNS Challenge 2022 года [3] не использовались в данной работе по следующим причинам. Во-первых, в DNS Challenge 2022 были представлены существенно большие по размерам образцы с частотой дискретизации 48 кГц, что при доступных авторам вычислительных мощностях не позволяло эффективно проводить вычислительные эксперименты. Во-вторых, с точки зрения практического применения частотный диапазон 20 Гц-8 кГц,

соответствующий частоте дискретизации 16 кГц, достаточен для представления человеческой речи в ситуации обычного диалога. За пределы данной полосы частот голос может выходить только в отдельных специфических ситуациях пения или крика.

При построении обучающей выборки согласно [2] использовались три набора данных.

1) Образцы неискаженной речи из нескольких открытых наборов данных (LibriVox, M-AILABS и др.). Следует отметить, что часть англоязычных образцов была отсеяна из-за низкого качества записи. Для отбора образцов применялась методика ITU-T P.808 [24] субъективной оценки качества. Отсеивались образцы со средней субъективной оценкой MOS ниже 4,3 по 5-бальной шкале.
2) Образцы шума из подмножества, полученного на основе двух наборов данных Audioset и Freesound. Подмножество состояло из 60000 образцов, относящихся к 150 классам шумов. Каждый класс шумовых сигналов содержал не менее 500 образцов для балансировки выборки.
3) Набор, состоящий из 3076 измеренных и более 110 тысяч синтезированных импульсных характеристик помещений, выбранных из наборов данных openSLR26 и openSLR28.

Синтез зашумленных примеров обучающей выборки осуществлялся следующим образом. Случайно выбиралась неискаженная запись голоса и набор из нескольких случайно выбранных образцов шума. Шумовые записи масштабировались по амплитуде так, чтобы SNR суммы чистой и шумовой компонент соответствовал выбранному случайно значению, которые случайно выбирались из равномерного распределения на отрезке [0 дБ, 40 дБ].

Тестовая выборка состояла из набора образцов с соотношением сигнал-шум распределённом на отрезке от 0 дБ до 20 дБ. Эти образцы являлись частью набора данных DNS Challenge 2021 [2] и были синтезированы независимо организаторами этих соревнований. Все приведенные ниже оценки качества работы, в том числе сравнение с существующими альтернативными решениями, вычислялись на данном фиксированном наборе образцов.

5. Обучение модели

5.1 Особенности реализации

Предложенная нейронная сеть была реализована на языке программирования Python с использованием фреймворка глубокого обучения PyTorch. Обучение проводилась на рабочей станции с двумя видеокартах Nvidia GeForce 1080Ti. Общее время обучения лучшей модели составило 45 дней. Модель обучалась методом оптимизации Adam со скоростью обучения 1e-4 для всех проведённых экспериментов.

5.2 Гиперпараметры модифицированной нейронной сети

Длительность каждого зашумленного образца составляла 2 с, частота дискретизации равнялась 16 кГц. Каждый образец разбивался на временные окна 32 мс с перекрытием 16 мс. Для вычисления частотно-временного представления использовалось быстрое преобразование Фурье с окном Хэмминга. Размер входного тензора составлял $256 \times 128 \times 2$ (F=256, T=128). Для частотного кодирования использовался вектор длины K=10. Размер ядра начальной и конечной сверток составлял 3×5 со сдвигом 1.

В табл. 1 приведены значения всех гиперпараметров основной части нейронной сети. Использовались следующие обозначения (аналогичные обозначениям сети PVT [16]):

- P_i – размер патча в блоке кодирования B_i;
- $\widehat{P}_i$ – размер патча в блоке декодирования B_i;
- C_i – количество каналов на выходе блока B_i;
- L_i – количество слоев трансформера слоев в блоке B_i;

- R_i – коэффициент сжатия в слоях трансформера блока B_i;
- H_i – количество голов в слоях трансформера в блоке B_i;
- E_i – коэффициент расширения в трансформере в блоке B_i.

Табл. 1. Гиперпараметры кодирующей и декодирующей компонент модифицированной сети PVT
Table. 1. Hyperparameters of the encoder and the decoder of the modified PVT network

Стадия	Номер блока	Размер выхода	Тип слоя (преобразование)	Гиперпараметры
Кодирование	1	$\frac{F}{8} \times \frac{T}{2}$	Кодировщик патча	$P_1 = (8, 2);\ C_1 = 64$
			Трансформер	$E_1 = 8, H_1 = 1; R_1 = 8, L_1 = 2$
	2	$\frac{F}{16} \times \frac{T}{4}$	Кодировщик патча	$P_2 = (2, 2);\ C_2 = 128$
			Трансформер	$E_2 = 8, H_2 = 2; R_2 = 4, L_2 = 2$
	3	$\frac{F}{32} \times \frac{T}{4}$	Кодировщик патча	$P_3 = (2, 1);\ C_3 = 320$
			Трансформер	$E_3 = 4, H_3 = 5; R_3 = 2, L_3 = 2$
	4	$\frac{F}{32} \times \frac{T}{4}$	Кодировщик патча	$P_4 = (1, 1);\ C_4 = 512$
			Трансформер	$E_4 = 4, H_4 = 8; R_4 = 1, L_4 = 2$
			Декодировщик патча	$\hat{P}_4 = (1, 1);\ C_4 = 320$
Декодирование	5	$\frac{F}{16} \times \frac{T}{4}$	Трансформер	$E_5 = 4, H_5 = 4; R_5 = 2, L_5 = 2$
			Декодировщик патча	$\hat{P}_5 = (2, 1);\ C_5 = 128$
	6	$\frac{F}{8} \times \frac{T}{2}$	Трансформер	$E_6 = 8, H_6 = 2; R_6 = 4, L_6 = 2$
			Декодировщик патча	$\hat{P}_6 = (2, 2);\ C_6 = 64$
	7	$F \times T$	Трансформер	$E_7 = 8, H_7 = 1; R_7 = 8, L_7 = 2$
			Декодировщик патча	$\hat{P}_7 = (8, 2);\ C_7 = 16$

6. Результаты и обсуждение

6.1 Используемые метрики оценки качества

В большинстве современных работ [6-13] для оценки качества методов улучшения речи не применялось отношение сигнал-шум, так как оно не отражает ни реальную разборчивость, ни воспринимаемое слушателем субъективное качество восстановленной речи. Поэтому в данной работе использовались следующие методы оценки качества.

1) Perceptual Evaluation of Speech Quality (PESQ) [24] – алгоритм автоматической оценки качества речевых записей, моделирующий субъективную оценку MOS по непрерывной шкале от -0,5 (наихудшее качество) до 4,5 (наилучшее качество). Выделялись два основных вида PESQ-метрики: узкополосная NB-PESQ для сигналов с частотой дискретизации 8 кГц, и широкополосная WB-PESQ для сигналов с частотой дискретизации 16 кГц;
2) Short-Time Objective Intelligibility Measure (STOI) [25] – алгоритм автоматической оценки разборчивости речи в диапазоне от 0 (наихудшее качество) до 1 (наилучшее качество). Оценка разборчивости в нем производилась с использованием сравнения спектральных характеристик чистого и очищенного голосовых сигналов;

3) Scale-Invariant Signal Distortion Rate (SI-SDR) [23] – устойчивая к различиям в громкости речевых сигналов модификация соотношения сигнал-шум.

Все эти методы в качестве входных данных требовали пары речевых сигналов, состоящие из искаженных и соответствующих очищенных образцов.

6.2 Выбор функции потерь

Для выбора лучшей функции потерь (раздел 3.3) использовалась специально синтезированная подвыборка набора данных DNS Challenge 2021 [2]. Она состояла из 125 часов чистой речи, добавление шума проводилось аналогично полной выборке данных. Количество эпох обучения равнялось 12. Результаты представлены в табл. 2. Видно, что самый стабильный результат на всех метриках показала модель, которая обучалась с функцией потерь -SI-SDR. Использование SNR в качестве функции потерь существенно снижает качество работы как для чистых образцов, так и для образцов с реверберацией, причем для последних относительный разрыв больше, что, вероятно, связано с наличием в образцовых «чистых» сигналах существенных искажений.

Табл. 2. Сравнение качества работы для двух функций потерь
Table. 2. Performance comparison for two loss functions

Функция потерь	С реверберацией				Без реверберации			
	STOI	SI-SDR	NB-PESQ	WB-PESQ	STOI	SI-SDR	NB-PESQ	WB-PESQ
-SI-SDR	0,9295	13,43	2,764	2,028	0,8892	13,1	3,014	2,333
-SNR	0,9188	10,73	2,554	1,887	0,873	10,43	2,777	2,171

6.3 Использование частотного кодирования

Экспериментально была проверена необходимость использования частого кодирования (раздел 3.2.2). Для проверки был использован полный набор DNS Challenge 2021. В модели в качестве функции потерь было применено отрицание SI-SDR. Количество эпох обучения равнялось 53. Как видно из табл. 3, использование частотного кодирования существенно улучшает метрики качества.

Табл. 3. Сравнение качества работы при включении/выключении частотного кодирования
Table. 3 Performance comparison of models with and without frequency encoding.

Использование частотного кодирования	С реверберацией				Без реверберации			
	STOI	SI-SDR	NB-PESQ	WB-PESQ	STOI	SI-SDR	NB-PESQ	WB-PESQ
Нет	0,9305	17,05	3,489	2,992	0,9618	18,52	3,252	2,73
Да	0,9362	17,51	3,561	3,11	0,9658	19,06	3,347	2,811

6.4 Сравнение с существующими решениями

Лучшей нейросетевой моделью в серии экспериментов оказалась версия модифицированной нейронной сети с включенным частотным кодированием. Использовалась функция потерь – SI-SDR. Обучение лучшей модели осуществлялось в течение 121 эпохи на полном наборе данных DNS Challenge 2021.

Предложенная модель тестировалась и сравнивалась на тестовой выборке синтетического набора данных соревнования DNS Challenge 2021 (табл. 4). В первой строке табл. 4 приведены значения метрик для необработанных зашумленных образцов тестовой выборки. Ниже приведены оценки метрик качества сетей PoCoNet [6], FullSubNet [7], предыдущей модели TS-LSTM [26], разработанной авторами данной работы, и предложенной модели. Модель PoCoNet использовала смешанную архитектуру кодировщик-декодировщик,

сочетающую сверточные слои с блоками самовнимания. Сети FullSubNet и TS-LSTM использовали рекуррентные слои. Показано, что предложенная модель превосходит существующие аналоги по всем метрикам.

Табл. 4. Сравнение качества работы предложенной модели с альтернативными решениями
Table. 4 Performance comparison of the proposed model with alternative methods.

Модель	С реверберацией				Без реверберации			
	STOI	SI-SDR	NB-PESQ	WB-PESQ	STOI	SI-SDR	NB-PESQ	WB-PESQ
Без обработки	0,8662	9,030	2,753	16822	0,9152	9,07	2,454	1,582
PoCoNet	-	-	-	2,832	-	-	-	2,748
FullSubNet	0,9262	15,750	3,473	2,969	0,9611	17,29	3,305	2,777
TS-LSTM	0,8600	9,376	3,050	2,220	0,9605	17,58	3,338	2,832
Предложенная модель	**0,9402**	**17,67**	**3,592**	**3,176**	**0,9692**	**19,53**	**3,398**	**2,899**

6.5 Особенности процесса обучения предложенной нейронной сети

При разработке новых нейросетевых методов интерес представляет интерпретация того, как идет обучение модели, на каких этапах этого процесса происходит качественное изменение результатов работы нейронной сети. Это имеет как научную, так и практическую значимость, так как может позволить лучше понять каким образом строить процесс обучения нейронных сетей для улучшения речи и, в более широком смысле, может помочь при создании новых подходов к обработке речевых сигналов.

В табл. 5 вынесены усредненные метрики качества работы обучаемой нейросети на тестовой выборке на 4, 21, 53 и 121 эпохах. Видно, что количественные оценки качества меняются все медленнее по мере того, как идет процесс обучения, и, как показывают приведенные ниже примеры, это не связано напрямую с качественным изменением поведения обучаемой сети.

Табл. 5. Сравнение качества работы предложенной модели на промежуточных эпохах обучения
Table. 5 Performance comparison of the proposed model at intermediate training epochs

Номер эпохи	С реверберацией				Без реверберации			
	STOI	SI-SDR	NB-PESQ	WB-PESQ	STOI	SI-SDR	NB-PESQ	WB-PESQ
4	0,913	15,48	3,294	2,687	0,9492	16,65	3,062	2,399
21	0,9281	15,48	3,464	2,964	0,9661	18,9	3,334	2,806
53	0,9362	17,51	3,561	3,11	0,9658	19,06	3,347	2,811
121	0,9402	17,67	3,592	3,176	0,9692	19,53	3,398	2,899

На рис. 3 представлены спектрограммы для речевого сигнала с шумом из класса «пение птиц». Участки спектра непосредственно со звуками птиц выделены белыми прямоугольниками. Также в паузах видны добавки широкополосного шума (в диапазоне частот до 4,5-5 кГц). Уже после 4 эпохи обучения (рис. 3в) они были почти полностью вычищены сетью, однако искажения от пения птиц остались. После 21 эпохи (рис. 3г) эти искажения практически полностью исчезли, а на 54 эпохе (рис. 3д) сеть научилась достраивать на этом месте форманты полезного речевого сигнала, формируя их из шума.

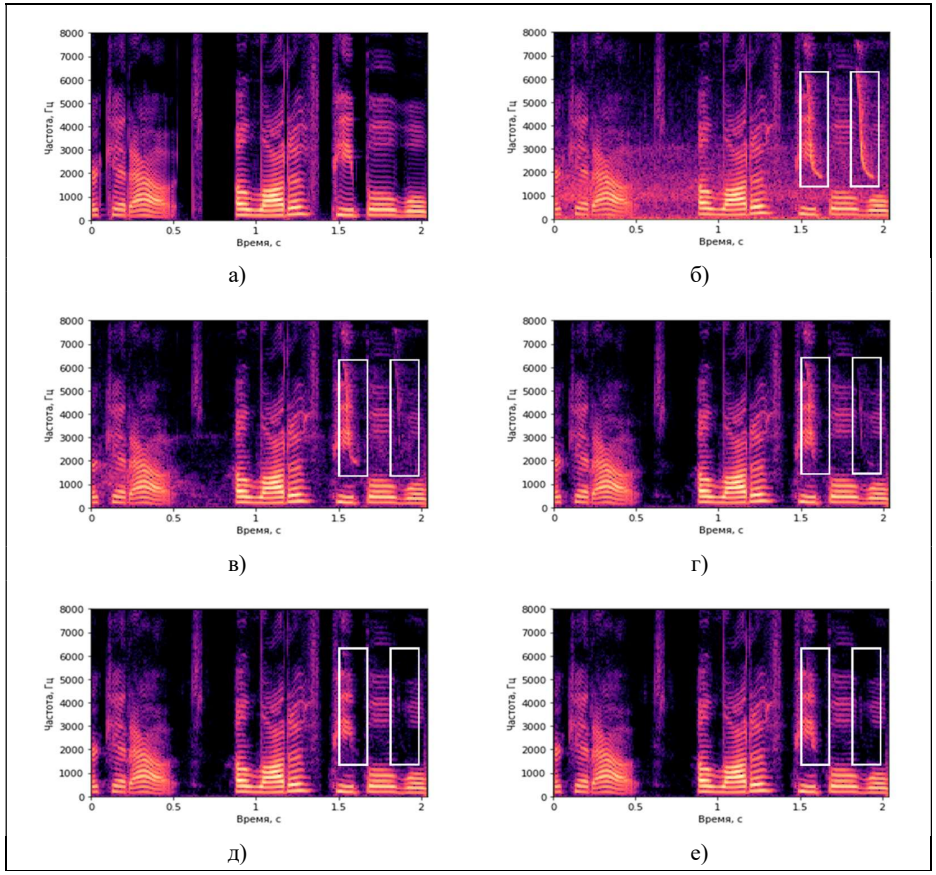

*Рис. 3. Пример спектрограмм чистого речевого сигнала **без реверберации** (а), с широкополосным шумом из класса **«пение птиц»** (б) и результатов работы модели после 4 (в), 21 (г), 53 (д) и 121 (е) эпох обучения*

*Fig. 3. Example of spectrograms of a clean speech signal **without reverberation** (a), with wideband noise from the **"birdsong"** class (b) and the results of the model after 4 (c), 21 (d), 53 (e) and 121 (f) epochs learning*

В случае того же исходного чистого сигнала, но с накладываемой реверберацией (рис. 4), видно, что сеть убирала искажения как от широкополосного шума, так и от птичьего пения уже с 4 эпохи (рис. 4в).

Дальнейшее обучение сети (рис. 4г-е) не существенно повлияло на вид спектра восстанавливаемого речевого сигнала. Это легко объясняется тем, что спектр реверберированной речи имеет меньше мелких деталей, формантные полосы более размыты.

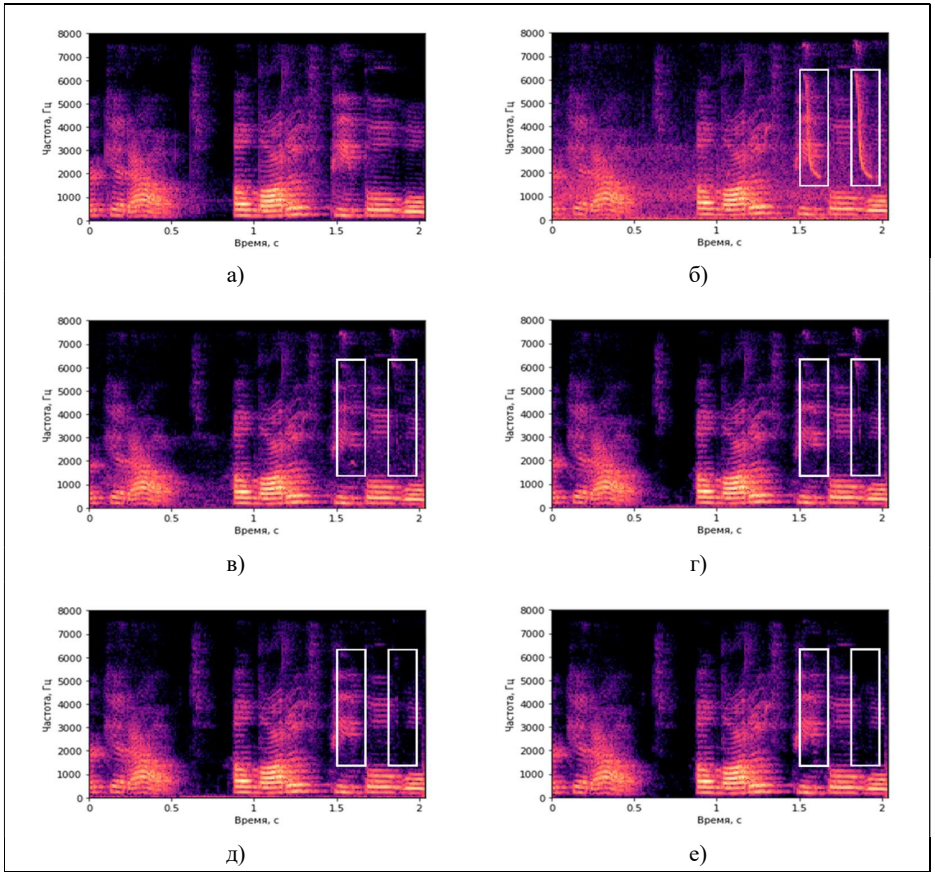

*Рис. 4. Пример спектрограмм чистого речевого сигнала **с реверберацией** (а), с широкополосным шумом из класса **«пение птиц»** (б) и результатов работы модели после 4 (в), 21 (г), 53 (д) и 121 (е) эпох обучения*

*Fig. 4. Example of spectrograms of a clean speech signal **with reverberation** (a), with wideband noise from the **"birdsong"** class (b) and the results of the model after 4 (c), 21 (d), 53 (e) and 121 (f) epochs learning*

На рис. 5 представлены спектрограммы сигнала без реверберации с шумом класса «музыка» (звучащий на фоне музыкальный инструмент типа синтезатора). Шумовое искажение сосредоточено в узкой полосе частот, выделенной на рис. 5 белым прямоугольником. Начиная с 4 эпохи обучения (рис. 5в) сеть устранила значительную часть искажения, а в течение последующих эпох (рис. 5г-е), училась восстанавливать недостающие части спектральных компонент. Интересно отметить, что в области, выделенной зеленым прямоугольником, сеть сначала (рис. 5в) «воспринимала» часть шума на 1,5 кГц за форманту речевого сигнала и пыталась продолжить её, однако после следующих эпох обучения она устранила этот артефакт.

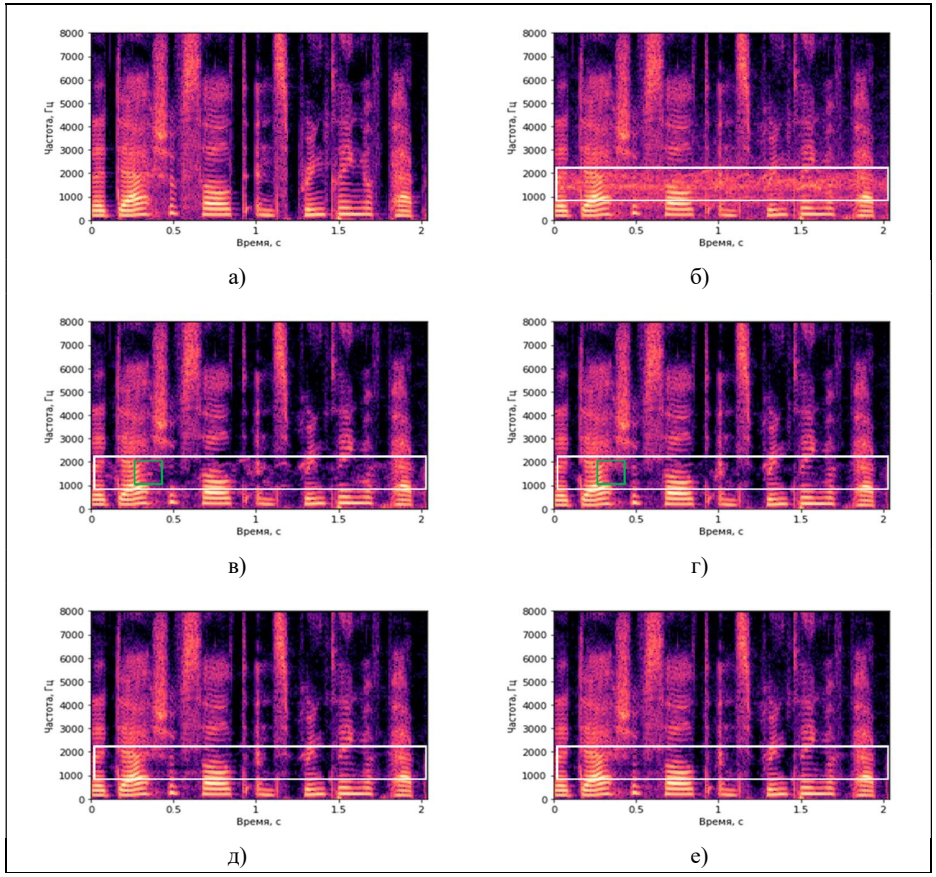

*Рис. 5. Пример спектрограмм чистого речевого сигнала **без реверберации** (а), с узкополосным шумом из класса **«музыка»** (б) и результатов работы модели после 4 (в), 21 (г), 53 (д) и 121 (е) эпох обучения*
*Fig. 5. Example of spectrograms of a clean speech signal **without reverberation** (a), with narrowband noise from the **"music"** class (b) and the results of the model after 4 (c), 21 (d), 53 (e) and 121 (f) training epochs*

На рис. 6 представлены спектрограммы для сигнала с искажением из класса «шум моря». В данном сигнале присутствовало широкополосное зашумление и звуки морских птиц, выделенные белым прямоугольником. Сигнал в процессе обучения последовательно улучшался без резких изменений до 53 эпохи (рис. 6д). В отличие от предыдущих примеров, более существенные изменения в работе обучаемой сети, пришлись как раз на поздний период между 53 (рис. 6д) и 121 (рис. 6е) эпохами, что совершенно не нашло отражения в средних количественных оценках качества из табл. 5.

Таким образом, в процессе обучения нейронная сеть довольно быстро (за несколько первых эпох) училась выделению и удалению шумовых искажений, протяженных как в частотной (шириной порядка килогерц), так и во временной (от десятых долей секунды и больше) областях. При этом восстановление деталей полезного сигнала было не точным (рис. 5в). Дальнейшее обучение привело к тому, что сеть стала не только удалять шумы, но и правильно достраивать отдельные форманты восстанавливаемой речи. Чем выше мощность или

сложнее структура аддитивного шума, тем больше эпох обучения требовалось для эффективной работы.

Слабое относительное изменение усредненных метрик (табл. 5) на поздних эпохах без приведенного выше анализа спектрограмм могло интерпретироваться как достижение сходимости параметров сети, что привело бы к ранней остановке процесса обучения, и, как следствие, недостаточному качеству восстановления речевых сигналов со «сложной» шумовой добавкой (рис. 6). Поэтому возможной рекомендацией при обучении нейросетевых методов шумоочистки является необходимость дополнительного выборочного контроля работы обучаемой нейронной сети на отдельных образцах с качественно разными классами шумов.

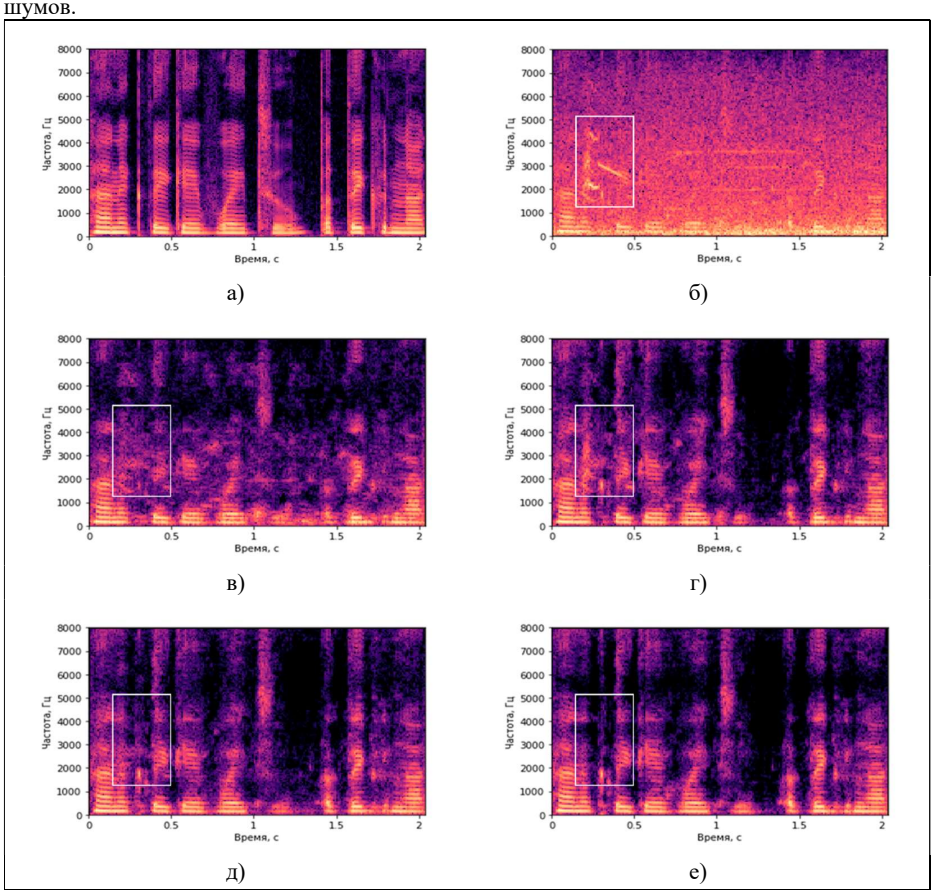

*Рис. 6. Пример спектрограмм чистого речевого сигнала **без реверберации** (а), с широкополосным шумом из класса **«шум моря»** (б) и результатов работы модели после 4 (в), 21 (г), 53 (д) и 121 (е) эпох обучения*
*Fig. 6. Example of spectrograms of a clean speech signal **without reverberation** (a), with wideband noise from the " sound of the sea" class (b) and the results of the model after 4 (c), 21 (d), 53 (e) and 121 (f) training epochs*

7. Заключение

В данной работе разработан новый тип кодирующей-декодирующей нейронной сети для решения задачи улучшения качества речи. Он основан на использовании модифицированной сети пирамидального трансформера и оригинальном методе декодирования иерархического представления обрабатываемого аудиосигнала.

Продемонстрирована важность добавления векторов частотного кодирования при формировании входных данных и проведен выбор наиболее подходящей функции потерь для наилучшего достижимого качества работы предложенного подхода.

Сравнение по метрикам качества показало, что разработанная нейронная сеть превосходит существующие современные решения. Анализ процесса обучения сети продемонстрировал постепенное совершенствование используемого сетью способа улучшения речевых сигналов – от маскирования шумовых областей в спектрограмме до восстановления искаженных формантных полос.

Разработанная нейросетевая модель может найти свое применение как непосредственно в прикладных приложениях, так и в качестве «учителя» при использовании метода дистилляции для обучения быстрых нейросетевых моделей-«учеников».

Список литературы / References

[1] Loizou P. Speech Enhancement. Theory and Practice, 2nd Edition. CRC Press, 2017. 711 p.

[2] Reddy C., Dubey H., Koishida K. Interspeech 2021 Deep Noise Suppression Challenge. arXiv preprint arXiv: 2101.01902, 2021, 5 p.

[3] Dubey H., Gopal V., Cutler R. et al. ICASSP 2022 Deep Noise Suppression Challenge. DNS C, 2022. arXiv preprint arXiv: 2202.13288, 2022. 5 p.

[4] Borgström B.J., Brandstein M.S., Dunn R.B. Improving Statistical Model-Based Speech Enhancement with Deep Neural Networks. In Proc. of the 16th International Workshop on Acoustic Signal Enhancement (IWAENC), 2018, pp. 471-475.

[5] Zheng C., Liu W. et al. Low-latency Monaural Speech Enhancement with Deep Filter-bank Equalizer. The Journal of the Acoustical Society of America, vol. 151, issue 5, 2021, article no. 3291.

[6] Umut I., Giri R. et al. PoCoNet: better speech enhancement with frequency-positional embeddings, semi-supervised conversational data, and biased loss. In Proc. of the INTERSPEECH 2020, 2020, pp. 2487-2491.

[7] Hao X., Su X. et al. FullSubNet: a full-band and sub-band fusion model for real-time single-channel speech enhancement. In Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2021, pp. 6633-6637.

[8] Xu, R., Wu, R. et al. Listening to sounds of silence for speech denoising. In Proc. of the Conference on Neural Information Processing Systems (NeurIPS 2020), 2020, pp. 9633-9648.

[9] Zheng C., Peng X., Zhang Y. Interactive Speech and Noise Modeling for Speech Enhancement. arXiv preprint arXiv: 2105.05537, 2020, 9 p.

[10] Luo, Y., Mesgarani, N. Conv-TasNet: surpassing ideal time-frequency magnitude masking for speech separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, issue 8, 2019, pp. 1256-1266.

[11] Liu Y., Zhang H. et al. Supervised speech enhancement with real spectrum approximation. In Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2019, pp. 5746-5750.

[12] Williamson D.S., Wang Y., Wang D. Complex ratio masking for monaural speech separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, issue 3, 2016, pp. 483-492.

[13] Tan K., Wang D.A. Convolutional recurrent neural network for real-time speech enhancement. 2018, 2018, pp. 3229-3233.

[14] Tallec C., Ollivier Y. Can recurrent neural networks warp time? In Proc. of the International Conference on Learning Representation, 2018, pp. 1-13.

[15] Vaswani A., Shazeer N. et al. Attention is all you need. In Proc. of the Conference on Neural Information Processing Systems (NIPS 2017), 2017, 11 p.

[16] Wang W., Xie E. et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, , pp. 548-558.

[17] Cao H., Wang Y. et al. Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv: 2105.05537, 2020. 14 c. DOI: 10.48550/arXiv.2105.05537

[18] Kun Wei, K., Guo P., Jiang N. Improving Transformer-based Conversational ASR by Inter-Sentential Attention Mechanism, arXiv preprint arXiv: 2207.00883, 2022, 5 p.

[19] Wang Y., Shen G. et al. Learning Mutual Correlation in Multimodal Transformer for Speech Emotion Recognition. In Proc. of the INTERSPEECH 2021, 2021, pp. 4518-4522.

[20] Wu, C., Xiu, Z. et al. Transformer-based acoustic modeling for streaming speech synthesis. c INTERSPEECH 2021, 2021, pp. 146-150.

[21] Ronneberger O., Fischer P., Brox T. U-Net: convolutional networks for biomedical image segmentation. Lecture Notes in Computer Science, vol. 9351, 2015, pp. 234-241.

[22] Lin T, Dollár P., Girshick R. Feature Pyramid Networks for Object Detection. In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 936-944.

[23] Roux J.L., Wisdom S. et al. SDR – Half-baked or well done? In Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2019, pp. 626-630.

[24] Naderi B., Cutler R. An opensource implementation of ITU-T recommendation P.808 with validation. In Proc. of the INTERSPEECH 2020, 2020, pp. 2862-2866.

[25] Taal C.H., Hendriks R.C. et al. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In Proc. of the IIEEE International Conference on Acoustics, Speech and Signal Processing, 2010, pp. 4214-4217.

[26] Nasretdinov R., Ilyashenko I., Lependin A. Two-Stage Method of Speech Denoising by Long Short-Term Memory Neural Network. Communications in Computer and Information Science, vol 1526, 2021, pp. 86-97.

Информация об авторах / Information about authors

Андрей Александрович ЛЕПЕНДИН – кандидат физико-математических наук, доцент кафедры информационной безопасности Института цифровых технологий, электроники и физики Алтайского государственного университета. Сфера научных интересов: цифровая обработка сигналов, машинное обучение, глубокие нейронные сети, голосовая биометрия, улучшение качества речи.

Andrey Aleksandrovich LEPENDIN – candidate of science in physics and mathematics, associate professor of the Department of information security of the Institute of Digital Technology, Electronics and Physics of Altai State University. Research interests: digital signal processing, machine learning, deep neural nets, voice biometry, speech enhancement.

Рауф Салаватович НАСРЕТДИНОВ – аспирант кафедры информационной безопасности Института цифровых технологий, электроники и физики Алтайского государственного университета. Сфера научных интересов: машинное обучение, глубокие нейронные сети, автоматическое распознавание речи, улучшение качества речи.

Rauf Salavatovich NASRETDINOV – post-graduate student of the Department of information security of the Institute of Digital Technology, Electronics and Physics of Altai State University. Research interests: machine learning, deep neural nets, automatic speech recognition, speech enhancement.

Илья Дмитриевич ИЛЬЯШЕНКО – аспирант кафедры информационной безопасности Института цифровых технологий, электроники и физики Алтайского государственного университета. Сфера научных интересов: машинное обучение, глубокие нейронные сети, биометрические методы верификации, улучшение качества речи.

Ilya Dmitrievich ILYASHENKO – post-graduate student of the Department of information security of the Institute of Digital Technology, Electronics and Physics of Altai State University. Research interests: machine learning, deep neural nets, biometric verification, speech enhancement.