8. Yuan Z. Bearing fault diagnosis using a speed-adaptive network based on vibro-speed data fusion and majority voting / Z. Yuan, Z. Ma, X. Li [et al.] — DOI: 10.1088/1361-6501/AC46EE. — Text : electronic // Meas. Sci. Technol. IOP Publishing. — 2022. — Vol. 33, № 5. — P. 055112.
9. Chen Z. Rolling Bearing Fault Diagnosis Using Time-Frequency Analysis and Deep Transfer Convolutional Neural Network / Z. Chen, J. Cen, J. Xiong. — DOI: 10.1109/ACCESS.2020.3016888. — Text : electronic // IEEE Access. Institute of Electrical and Electronics Engineers Inc. — 2020. — Vol. 8. — P. 150248–150261.
10. Prognostics Center of Excellence. — Data Repository. — URL: https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/#bearing (accessed: 21.01.2022). — Text : electronic.
11. Zhang R. Fault Diagnosis from Raw Sensor Data Using Deep Neural Networks Considering Temporal Coherence / R. Zhang, Z. Peng, L. Wu [et al.]. — DOI: 10.3390/s17030549. — Text : electronic // Sensors. — 2017. — Vol. 17, № 3. — P. 549.
12. Xu Y. A hybrid deep-learning model for fault diagnosis of rolling bearings / Y. Xu, Z. Li, S. Wang [et al.] — DOI: 10.1016/j.measurement.2020.108502. — Text : electronic // Measurement. Elsevier. — 2021. — Vol. 169. — P. 108502.
13. Huang M. Fault diagnosis of rolling bearing based on empirical mode decomposition and convolutional recurrent neural network / M. Huang, T. Huang, Y. Zhao [et al.]. — DOI: 10.1088/1757-899X/1043/4/042015. — Text : electronic // IOP Conf. Ser. Mater. Sci. Eng. IOP Publishing. — 2021. — Vol. 1043, № 4. — P. 042015.

Д. М. Шаход, О. Л. Ибряева

Южно-Уральский государственный университет, г. Челябинск

УДК 004.032.26, 004.048

НЕЙРОСЕТЕВОЙ АЛГОРИТМ ПОДАВЛЕНИЯ ЭХА В УСЛОВИЯХ ДВОЙНОГО РАЗГОВОРА

Аннотация. *В работе решается задача подавления акустического эха в условиях двойного разговора на основе нейронной сети, оценивающей идеальную двоичную маску IBM из признаков, извлеченных из смеси сигналов ближнего и дальнего конца. Новизна предложенного метода заключается*

в использовании алгоритма кластеризации (EM, Mean-Shift, k-Means) дополнительно с двунаправленной рекуррентной нейронной сетью BLSTM.

Ключевые слова: *идеальная двоичная маска, сигнал ближнего конца, сигнал дальнего конца, двунаправленная рекуррентная нейронная сеть, кластеризация.*

Введение. Алгоритмы восстановления речевого сигнала, искаженного аддитивным некоррелированным шумом, в случае, когда доступен только зашумленный сигнал, широко применяются в различных областях цифровой обработки речевых сигналов, таких как распознавание речи, распознавание говорящего, детектирование речевой активности, улучшение качества и разборчивости речевых сигналов и др. [1]. С развитием эффективных методов машинного обучения широкое распространение стали получать алгоритмы подавления шума на основе глубоких нейронных сетей [2-4]. Одними из наиболее используемых методов шумоподавления являются методы, основанные на оценке частотно-временных масок [5]. Например, в работах [6, 7] в роли целевого выхода нейросетевой модели выступает идеальная двоичная маска (ideal binary mask, IBM). В настоящей работе разработан алгоритм на основе двунаправленной рекуррентной сети (Bidirectional Long Short-Term Memory, BLSTM) выходом которой является маска IBM. Ключевой особенностью нашего алгоритма является использование кластеризации на выходе нейронной сети.

Постановка задачи. Рассматриваемая модель приведена на рис. 1. Сигнал микрофона $\boldsymbol{y(n)}$ состоит из сигнала на ближнем конце $\boldsymbol{s(n)}$, эха $\boldsymbol{d(n)}$ и фонового шума $\boldsymbol{v(n)}$:

$$y(n) = d(n) + s(n) + v(n) \qquad (1)$$

Для простоты в этой работе будем считать $v(n) = 0$.

Эхо-сигнал $d(n)$ формируется за счет отражения сигнала с дальнего конца $x(n)$ от стенок комнаты и моделируется путем свертки $x(n)$ с импульсной характеристикой $h(n)$ помещения RIR (Room Impulse Responses).

Задача данного исследования — выделить из смеси $y(n)$ полезный сигнал $s(n)$, убрав нежелательную помеху $d(n)$.

Из сигнала микрофона $y(n)$ с помощью кратковременного преобразования Фурье STFT (Short Time Fourier Transform) извлекаются признаки, которые служат входными данными для двунаправленной рекуррентной нейронной сети BLSTM. Выходом нейронной сети является бинарная маска Ideal Binary Mask (IBM), которая часто используется в качестве цели в задаче разделения речи от помех. Используя IBM маску, можно оценить спектр сигнала ближнего конца и, с помощью обратного преобразования Фурье ISTFT, восстановить $s(n)$.

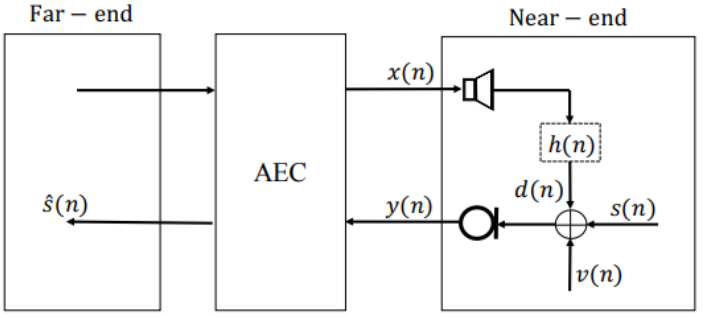

Рис. 1. Аддитивная модель акустического эха

Входные данные для модели

Исходные данные представляют собой аудофайлы из базы данных TIMIT. Из этих данных мы (случайным образом) выбирали сигналы ближнего конца $s(n)$, дальнего $x(n)$ и формировали соответствующие сигналы микрофона $y(n)$. К этим сигналам, передискретизированным с частоты 16 кГц до 8 кГц (с целью уменьшения времени обработки данных) было применено STFT с окном Ханнинга шириной 256 точек, что соответствует временной длине в 32 миллисекунды и 129 элементам разрешения по частоте. Для увеличения обучающей выборки была проведена аугментация данных, заключающаяся в перекрытии временных сигналов на 50%.

Полученные спектрограммы Y сигналов микрофона $y(n)$ были разбиты на блоки 100×129 для того, чтобы у нейронной сети всегда был вход фиксированного размера. Итоговый объем обучающей

выборки составил 13606 образцов, объем тестовых данных — 3093 образца, объем валидациооного набора данных — 1855 образцов. Таким образом, набор данных разбит на тренировочный, тестовый, валидационный датасеты в соотношении примерно 75% — 15% — 10%. Также были найдены спектрограммы цели $s(n)$ и помехи $d(n)$, которые использовались, чтобы найти маску IBM, являющуюся выходом модели для данного входа Y.

Выход модели

Выходом нейронной сети (целью обучения, target) является идеальная двоичная маска IBM — одна из наиболее часто используемых масок в задачах распознования речи. IBM определяется как [8]:

$$IBM(t,f) = \begin{cases} 1 \; if \; \dfrac{S_T(t,f)}{S_I(t,f)} > 1 \\ 0 \; otherwise \end{cases} \quad (2)$$

Здесь $S_T = S_T(t,f)$ и $S_I = S_I(t,f)$ — спектрограммы цели $s(n)$ и помехи $d(n)$, соответственно. Некоторые исследователи, например [9], считают, что при использовании IBM восстановленная речь звучит не естественно, однако разборчивость речи при этом очень хорошая.

Если мы обозначим за Y спектрограмму сигнала микрофона, то

$$Y = S_T + S_I \quad (3)$$

и, с помощью маски IBM, можно восстановить спектрограмму полезного сигнала $s(n)$ [8]:

$$S_T = IBM \odot Y \quad (4)$$

Здесь оператор $\odot$ представляет собой поэлементное умножение.

По известной спектрограмме для $s(n)$ можно далее восстановить сам сигнал с помощью обратного преобразования Фурье.

Описание модели BLSTM+clustering

Модель двунаправленной рекуррентной нейронной сети BLSTM содержит две однонаправленные рекуррентные сети LSTM с 300

нейронами, одна из которых обрабатывает сигнал в прямом направлении, а другая — в обратном. Выходной полносвязный слой имеет сигмоидную функцию активации, и диапазон значений в [0, 1], который легко (с установкой порогового значения 0.5) трансформируется в дискретный выход из нулей и единиц, соответствующий IBM маске.

Для обучения сети был выбран оптимизатор adam, в качестве функции потерь использовалась среднеквадратичная ошибка MSE (Mean Square Error). Скорость обучения была равна 0.01. Количество эпох обучения было равно 100.

Результаты описанной модели чистой BLSTM оказались неудовлетворительными, поэтому нами было решено использовать дополнительно глубокую кластеризацию. В этом случае мы увеличили размеры матрицы весов и смещения в три раза на последнем слое нейронной сети и теперь ее выход представляет собой матрицу размера (12900, 3) (в отличие от матрицы (12900, 1) для модели BLSTM). Далее применяется алгоритм кластеризации K-Means к 12900 точкам в трехмерном пространстве. Разделяя данные на два класса, получаем вектор из нулей и единиц, который затем преобразуем в матрицу, соответствующую маске IBM.

Результаты. Для обучения, валидации и тестирования из исходного набора данных с 630 носителями были случайным образом выбраны записи 462, 68 и 100 человек, соответственно. Поскольку каждый человек читает 10 предложений, у нас имеется 4620 аудиофайлов для обучения, 680 для валидации и 1000 для тестирования. Случайно выбранная пара из этого набора представляет собой, как правило, аудиофайлы с речью разных людей, которые мы берем в качестве сигналов ближнего $s(n)$ и дальнего конца $x(n)$. Из сигнала $x(n)$ путем свертки с импульсной характеристикой $h(n)$ помещения RIR формируется эхо-сигнал $d(n)$. Смешивая $d(n)$ с $s(n)$, получаем сигнал микрофона. Таким образом, у нас имеется 2310, 340 и 500 пар аудиофайлов для обучения, валидации и тестирования. Поскольку длительность аудиофайлов различна, то из каждого из них мы получаем различное число спектрограмм (в среднем, около 3, но с учетом аугментации и перекрытия в 50%, около 6).

Окончательно, объем обучающей, валидационной и тестовой выборок составил в наших эксперимепнтах 13606, 1855 и 3093, соответственно. RIR генерируется при времени реверберации T60 = 0.5c (время, необходимое для уменьшения RIR на 60 dB) с использованием метода ISM [10]. Размер комнаты для моделирования составляет (9, 7.5, 3.5) m, микрофон находится в позиции (6.3, 4.87, 1.2) m внутри комнаты, источник помехи в (2.5, 3.73, 1.76) m.

Таблица 1 содержит значения метрик ERLE, PESQ и STOI, полученных четырьмя рассматриваемыми методами при (signal-to-echo ratio) SER = 6 дБ.

Таблица 1

Сравнение моделей при SER = 6 дБ

Метод	*ERLE*	*PESQ*	*STOI*
BLSTM	6.8	1.03	0.846
BLSTM+EM	3.5	0.91	0.714
BLSTM+Mean-Shift	-2.6	1.17	0.808
BLSTM+K-Means	**8.1**	**2.1**	**0.911**

Отметим, что для вычисления метрик использовались данные, которые не участвовали в процессе обучения. Как можно видеть, использование алгоритма k-Means улучшило все показатели, в то время как другие алгоритмы кластеризации почти всегда ухудшают работы модели BLSTM.

В табл. 2 отражены метрики эффективности моделей в случае SER = 10 дБ.

Таблица 2

Сравнение моделей при SER = 10 дБ

Метод	*ERLE*	*PESQ*	*STOI*
BLSTM	8.7	2.23	0.865
BLSTM+EM	5.3	1.59	0.770
BLSTM+Mean-Shift	-1.8	2.14	0.846
BLSTM+K-Means	**11.2**	**2.65**	**0.924**

Можно видеть, что и в этом случае добавление k-Means к BLSTM показывает улучшение значений ERLE примерно на 2.5 дБ, PESQ на 0.42 и STOI на 0.059. Таким образом, метрика STOI, характеризующая разборчивость речи, улучшилась на 7%, а метрика PESQ, характеризующая качество восстановления речи, на 18.8%. Использование алгоритмов Mean-Shift и EM не улучшило производительность модели BLSTM.

Заключение. В работе предложена модель восстановления зашумленного сигнала на основе двунаправленной рекуррентной нейронной сети BLSTM с IBM маской на выходе. Сеть обучалась и тестировалась на наборе данных TIMIT и показала недостаточную эффективность. Далее модель была модифицирована добавлением дополнительного этапа кластеризации данных. Были рассмотрены три метода кластеризации: k-Means, Mean-Shift, EM. Использование метода k-Means привело к существенному улучшению показателей ERLE, PESQ, STOI, в отличие от методов Mean-Shift, EM. В сценариях с двойным разговором, при соотношении сигнал/эхо 10 дБ метрика STOI, характеризующая разборчивость речи, улучшилась на 7%, а метрика PESQ, характеризующая качество восстановления речи, на 18.8%.

СПИСОК ЛИТЕРАТУРЫ

1. Benesty J. Speech Enhancement: A Signal Subspace Perspective / J. Benesty, J. Jensen, M. Christensen, J. Chen. — Text : direct // Elsevier Academic Press. — 2014. — 129 p.
2. Lee C. M. DNN-based residual echo suppression / C. M. Lee, J. W. Shin, N. S. Kim. — Text : direct // Interspeech. — 2015. — Dresden, 2015. — P. 1775-1779.
3. Zhang H. Deep learning for acoustic echo cancellation in noisy and double-talk scenarios / H. Zhang, D. Wang. — Text : direct // Interspeech 2018. — Hyderabad, 2018. — P. 3239-3243.
4. Zhang H. Deep learning for joint acoustic echo and noise cancellation with nonlinear distortions / H. Zhang, K. Tan, D. Wang. — Text : direct // Interspeech 2019. — Graz, 2019. — P. 4255-4259.
5. Wang D. On Ideal Binary Mask as the Computational Goal of Auditory Scene Analysis / D. Wang. — Text : direct // Speech Separation by Humans and Machines / ed. by P. Divenyi. — Springer, Boston, MA, 2005. — P. 181-197.

6. Li N. Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction / N. Li, P. C. Loizou. — Text : direct // J. Acoust. Soc. Am. — 2008. — Vol. 123, № 3. — P. 1673-1682.
7. Brungart D. S. Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation / D. S. Brungart, P. S. Chang, B. D. Simpson, D. Wang. — Text : direct // J. Acoust. Soc. Am. — 2006. — Vol. 120, № 6. — P. 4007-4018.
8. Zermini A. Deep Learning for Speech Separation: PhD thesis / A. Zermini. — Text : direct // University of Surrey, faculty of engineering, physical sciences, Centre for Vision. — Speech : Signal Processing (CVSSP), South East of England, UK, 2020.
9. Xia S. Using Optimal Ratio Mask as Training Target for Supervised Speech Separation / S. Xia, H. Li, X. Zhang. — Text : direct // 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). — Kuala Lumpur, 2017. — P. 163-166.
10. Allen J. B. Image method for efficiently simulating small-room acoustics / J. B. Allen, D. A. Berkley. — Text : direct // The Journal of the Acoustical Society of America. — 1998. — Vol. 65, № 4. — P. 943-950.

М. А. Устелемов, М. С. Воробьева
Тюменский государственный университет, г. Тюмень

УДК 004.94

ИССЛЕДОВАНИЕ ПОДХОДОВ К КЛАССИФИКАЦИИ НАЛИЧИЯ СРЕДСТВ ИНДИВИДУАЛЬНОЙ ЗАЩИТЫ ПО КАДРУ ЛИЦА НА ОСНОВЕ АНАЛИЗА ВИДЕОРЯДА МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

Аннотация. *В статье отражено исследование методов машинного обучения, которые могут служить компонентами автоматизированной системы выявления лиц, не использующих средства индивидуальной защиты.*

Ключевые слова: *fine-tuning, transfer learning, data augmentation, видеоряд, классификация, машинное зрение, трекинг объектов, набор данных, определение лица, средства индивидуальной защиты.*