УДК 004.032.26, 004.048 DOI: 10.14529/cmse220204

МЕТОД ПОДАВЛЕНИЯ АКУСТИЧЕСКОГО ЭХА НА ОСНОВЕ РЕКУРРЕНТНОЙ НЕЙРОННОЙ СЕТИ И АЛГОРИТМА КЛАСТЕРИЗАЦИИ

© 2022 Д.М. Шаход, О.Л. Ибряева

Южно-Уральский государственный университет
(454080 Челябинск, пр. им. В.И. Ленина, д. 76)
E-mail: ghiathlovealaa@gmail.com, ibriaevaol@susu.ru

Поступила в редакцию: 01.04.2022

В статье решается задача подавления акустического эха 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 дополнительно с двунаправленной рекуррентной нейронной сетью BLSTM. Для оценки использования алгоритмов кластеризации EM, Mean-Shift, k-Means, модели были обучены и протестированы на базе данных TIMIT. Для каждой модели были вычислены метрики ERLE, PESQ, STOI, характеризующие ее качество. Использование алгоритмов кластеризации EM, Mean-Shift оказалось неэффективным по сравнению с алгоритмом BLSTM при соотношении сигнал/эхо 10 дБ. При соотношении сигнал/эхо 6 дБ BLSTM+Mean-Shift привел к незначительному улучшению метрики PESQ по сравнению с алгоритмом BLSTM. Результаты экспериментов показали эффективность предложенной модели BLSTM при использовании сети с алгоритмом K-Means, по сравнению с использованием чистой BLSTM для подавления эха 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%.

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

ОБРАЗЕЦ ЦИТИРОВАНИЯ

Шаход Д.М., Ибряева О.Л. Метод подавления акустического эха на основе рекуррентной нейронной сети и алгоритма кластеризации // Вестник ЮУрГУ. Серия: Вычислительная математика и информатика. 2022. Т. 11, № 2. С. 43–58. DOI: 10.14529/cmse220204.

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 статье 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. В работе рассмотрены три метода кластеризации (EM, Mean-Shift, k-Means) и проведено сравнение алгоритмов на сигналах базы данных TIMIT на основе об-

щепринятых метрик в обработке речи: ERLE, STOI, PESQ. Показано, что дополнительное использование кластеризации k-Means улучшает работу модели BLSTM.

Статья организована следующим образом. Раздел 1 содержит краткий обзор работ по тематике исследования. В разделе 2 представлен метод BLSTM с кластеризацией. В разделе 3 приведены результаты вычислительных экспериментов по оценке эффективности предложенных методов. Заключение резюмирует полученные результаты и описывает направления будущих исследований.

1. Обзор связанных работ

Традиционно задача акустического эхоподавления решается за счет адаптации акустической импульсной характеристики между громкоговорителем и микрофоном с использованием фильтра с конечной импульсной характеристикой [8]. Одним из наиболее широко используемых адаптивных алгоритмов является NLMS (Normalized Least Mean Square) [9], имеющий хорошую надежность при низкой сложности. Однако он, как и все адаптивные алгоритмы [8], имеет недостаток возможной расходомости из-за корреляции между полезным сигналом и подавляемым эхом. Эта корреляция имеет место, например, во время так называемого «двойного разговора» и большинство алгоритмов строится на попытке сторонними средствами распознать двойную речь в конкретный момент времени и приостановить обучение адаптивного фильтра [10]. С помощью такого метода можно предотвратить расхождение фильтра, однако обучение значительно замедляется.

Сигнал, поступающий на микрофон, содержит не только эхо и речь на ближнем конце, но и фоновый шум, с которым система акустического эхоподавления сама по себе не способна справиться. Для подавления фонового шума и остаточных эхо-сигналов, которые существуют на выходе системы обычно используют пост-фильтры. Наример, в работе [11] авторы объединили адаптивный алгоритм с методом подавления шума на основе кратковременного спектрального затухания и получили высокую степень удаления эха при наличии фонового шума.

Другим вариантом устранения остаточного эха является использование мощных моделей глубокого обучения [12], которые способны моделировать сложные нелинейные искажения, вносимые усилителями мощности и громкоговорителями. Такие модели являются все более популярной альтернативой нелинейному моделированию систем акустического эхоподавления. Так, авторы работы [13] смоделировали нелинейную систему как модель Хаммерштейна и использовали полносвязную нейронную сеть с двумя слоями, за которой следует адаптивный фильтр для идентификации параметров модели. В недавней работе [2] использована глубокая нейронная сеть для оценки усиления остаточного шума как на дальнем конце сигнала, так и на выходе системы [14], чтобы удалить нелинейные компоненты эхо-сигнала.

Глубокое обучение показало большой потенциал для разделения речи [15, 16]. Способность рекуррентных нейронных сетей RNN моделировать изменяющиеся во времени функции может играть важную роль в решении проблем акустического эхоподавления. LSTM (Long short-Term memory) [17] — вариант RNN, разработанный для решения проблемы исчезновения градиента, присущей традиционным RNN, уже показал хорошую производительность в задачах распознавания и улучшения речи в шумных условиях [18, 19]. Работа [20] также отмечает лучшее качество модели, использующей LSTM по сравнению с глубокой полносвязной нейронной сетью.

2. Предложенная модель

2.1. Постановка задачи

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

$$y(n) = d(n) + s(n) + v(n). \tag{1}$$

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

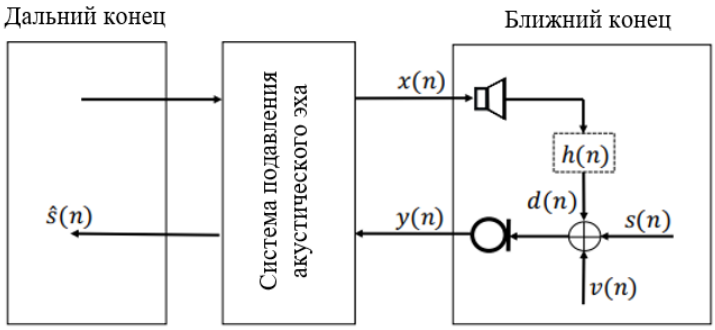

Рис. 1. Аддитивная модель акустического эха

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).

Наша задача — выделить из смеси $y(n)$ полезный сигнал $s(n)$, убрав нежелательную помеху $d(n)$.

На рис. 2 представлена схема предлагаемой нами модели для решения данной задачи. 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 $s(n)$.

2.2. Входные данные для модели

Исходные данные представляют собой аудофайлы из базы данных TIMIT (описанной в параграфе 3.1). 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 кратковременное преобразование Фурье 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.

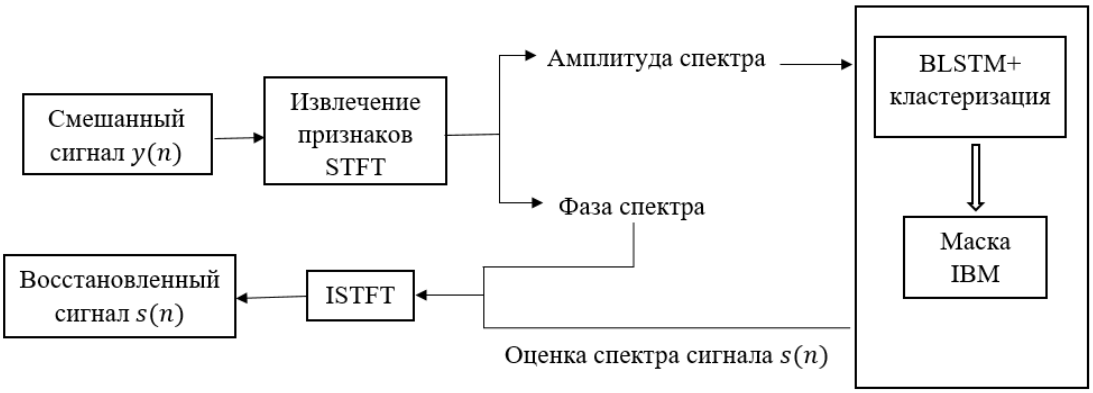

Рис. 2. Схема предложенной модели выделения сигнала $s(n)$

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.

2.3. 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 [21]:

$$IBM = \begin{cases} 1, & \text{if } \frac{S_T(t,f)}{S_I(t,f)} > 1 \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

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 [22], 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

$$Y = S_T + S_I \tag{3}$$

и, с помощью маски IBM, можно восстановить спектрограмму полезного сигнала $s(n)$ [21]:

$$S_T = IBM \odot Y. \tag{4}$$

Здесь оператор $\odot$ представляет собой поэлементное умножение.

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

2.4. Описание модели BLSTM+clustering

Модель двунаправленной рекуррентной нейронной сети BLSTM имеет два слоя, Каждый слой 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 размером 100×129 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 (они, как и результаты для других моделей, представлены в параграфе 3.4) и нами было решено использовать дополнительно глубокую кластеризацию.

На рис. 3 приведена структура нейронной сети BLSTM+clustering. 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 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. В качестве алгоритмов кластеризации в работе использовались известные алгоритмы K-Means, Mean-shift, EM (Expectation-maximization).

2.5. Метрики качества

Для оценки качества модели в работе используются три метрики.

1. Уровень ослабления эхо-сигнала ERLE (Echo Return Loss Enhancement).

ERLE измеряет, насколько хорошо акустическое эхо удаляется из смешанного сигнала и определяется по формуле [23]:

$$ERLE = \lim_{n\to\infty} 10\log_{10}\left(\frac{E[y^2(n)]}{E[e^2(n)]}\right), \tag{5}$$

где E — среднее значение. Чем выше этот показатель, тем лучше, он говорит о меньшем остаточном эхе e.

2. Перцепционная оценка качества речи PESQ (Perceptual Evaluation of Speech Quality).

PESQ — общепринятая метрика для оценки качества речи, которая сравнивает исходный сигнал $s(n)$ с полученной для него оценкой $\hat{s}(n)$. Процедура вычисления PESQ является довольно сложной, ее описание приводится, в частности, в работе [24].

В большинстве практических случаев величина PESQ принимает значения в диапазоне от 0.5 до 4.5. Более высокий балл указывает на лучшее качество.

3. Кратковременная объективная разборчивость речи STOI (Short-Time Objective Intelligibility).

STOI практически всегда идет рядом с PESQ и отвечает за «понимаемость» речи [25]. Обычно значения STOI находятся в диапазоне $[0, 1]$ и, можно считать, означают процент правильности работы модели.

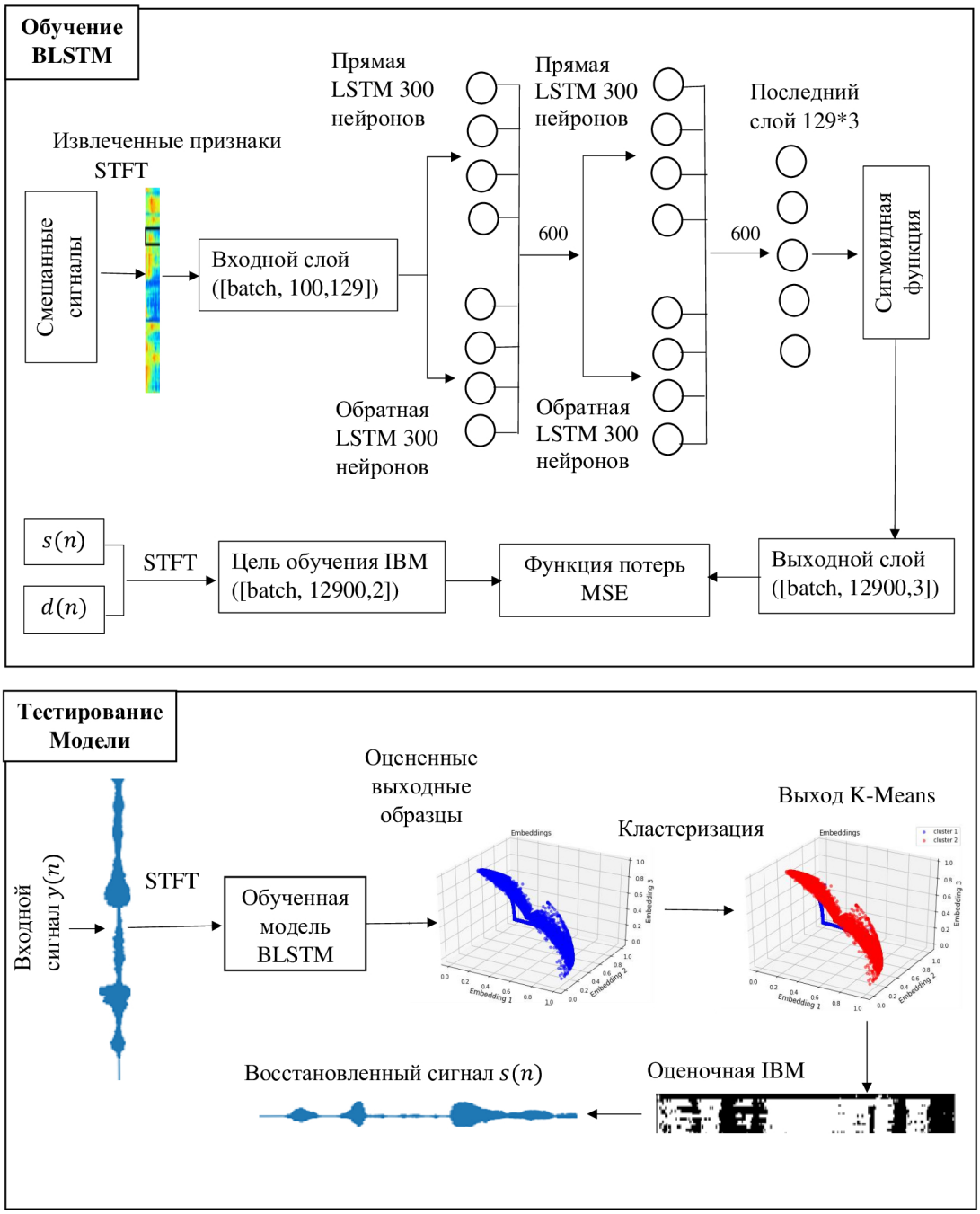

Рис. 3. Структура предлагаемой модели

3. Результаты экспериментов

3.1. Набор данных

Набор данных TIMIT — один из известных наборов данных для акустико-фонетических исследований, а также для разработки и оценки систем автоматического распознавания речи. TIMIT содержит записи 630 носителей восьми основных диалектов американского английского языка, каждый из которых читает десять предложений с фонетически богатым звучанием. Набор данных записан с частотой 16 KHz в Texas Instruments, Inc, Массачусетским технологическим институтом [21].

3.2. Оценка эффективности модели BLSTM

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 экспериментах 13606, 1855 и 3093, соответственно.

RIR генерируется при времени реверберации $T60 = 0.5$с (время, необходимое для уменьшения RIR на 60 dB) с использованием метода источника изображения ISM (Image Source Method) [26]. 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. Источник помехи озвучивает содержимое wav-файла (из TIMIT), начиная с 1.3 секунды.

Отношение сигнал/эхо SER вычисляется по формуле:

$$SER = 10 \lg \left(\frac{E[s^2(n)]}{E[d^2(n)]} \right). \tag{6}$$

Далее в работе будут приведены результаты работы моделей для случая $SER = 6$ дБ.

На рис. 4 представлены графики изменения метрики Accuracy (доли правильных ответов) для тренировочной и валидационной выборок в процессе 100 эпох обучения. Видно, что метрика достигла почти постоянного порога, который однако составляет немногим более 50% для валидационной выборки, что совершенно неудовлетворительно.

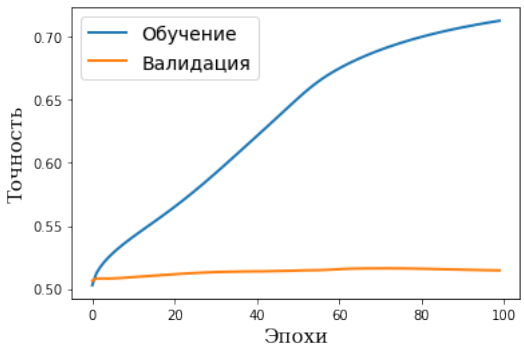

Рис. 4. Оценка эффективности модели BLSTM

На рис. 5 показан пример спектрограммы сигнала микрофона, оценочной маски IBM, являющейся целевым выходом модели BLSTM, а также настоящей маски (Real IBM).

На рис. 6 показаны сигнал входа для модели BLSTM $y(n)$ (Mixture signal), сигнал ближнего конца $s(n)$ (Near-end signal), ресинтезированный целевой сигнал $\hat{s}(n)$ (Target signal),

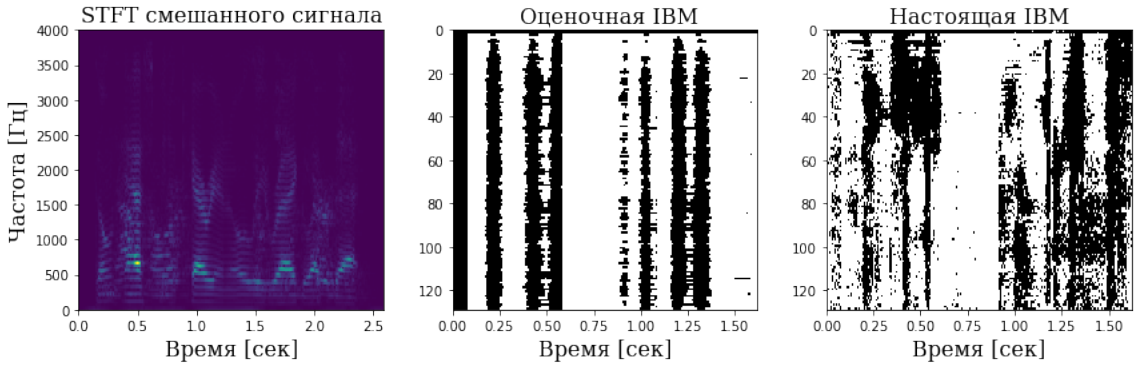

Рис. 5. Спектрограмма сигнала микрофона, оценочная маска IBM, полученная методом BLSTM и настоящая маска IBM

и эхо (Echo) $d(n)$. Можно видеть, что сигналы $s(n)$ и $\hat{s}(n)$ очень похожи. Значение метрик PESQ и STOI составило 1.03 и 0.846, соответственно.

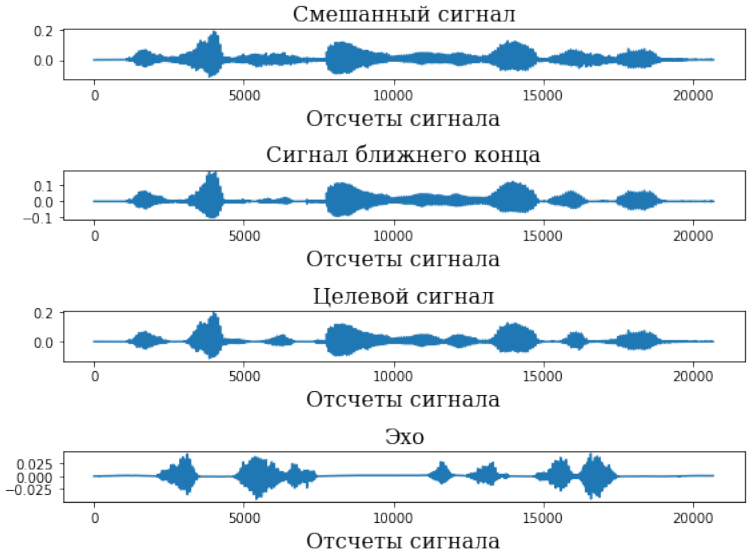

Рис. 6. Входной и выходной сигналы модели BLSTM

3.3. Оценка эффективности модели BLSTM+K-Means

Перейдем к построению предложенной модели после добавления кластеризации на выходе BLSTM. На рис. 7 представлены графики изменения метрики Accuracy для тренировочной и валидационной выборок в процессе 100 эпох обучения. Отметим, что в отличие от аналогичного рис. 4 для простой модели BLSTM, сейчас достигнута точность около 78% на валидационной выборке.

На рис. 8 показаны точки в трехмерном пространстве, являющиеся выходом обучающей модели, а также показано разбиение этих точек на два класса, полученное с помощью алгоритма кластеризации K-Means.

На рис. 9 показан пример спектрограммы сигнала микрофона, оценочной маски IBM, являющейся целевым выходом модели BLSTM+k-Means, а также настоящей маски (Real IBM).

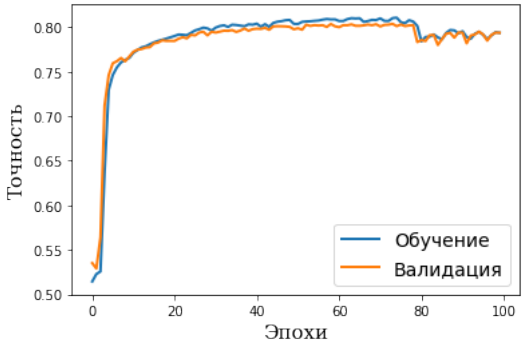

Рис. 7. Оценка эффективности модели BLSTM+K-Means

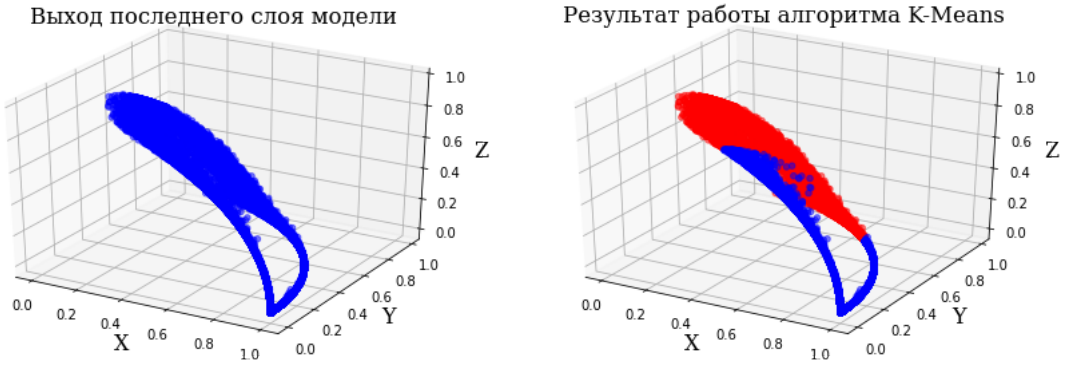

Рис. 8. Выход модели и результат k-Means кластеризации

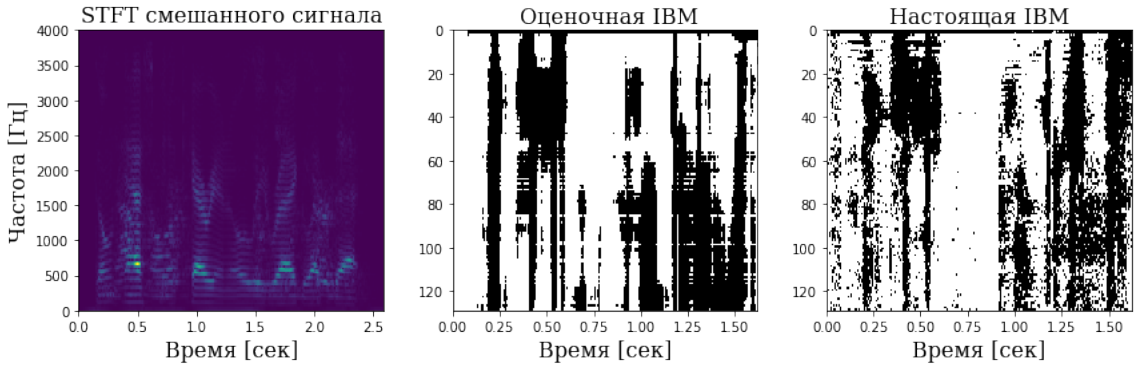

Рис. 9. Спектрограмма сигнала микрофона, оценочная маска IBM, полученная методом BLSTM+K-Means и настоящая маска IBM

На рис. 10 показаны сигнал входа BLSTM+K-Means, сигнал ближнего конца, ресинтезированный целевой сигнал и эхо. Значение метрик PESQ и STOI, характеризующих качество восстановления сигнала, составило 2.1 и 0.911, соответственно, и превысило их значения в случае простой модели BLSTM.

Аналогично были проанализированы алгоритмы BLSTM+EM, BLSTM+MeanShift, основанные на использовании методов кластеризации EM и MeanShift. Полученные значения метрик приведены в табл. 1.

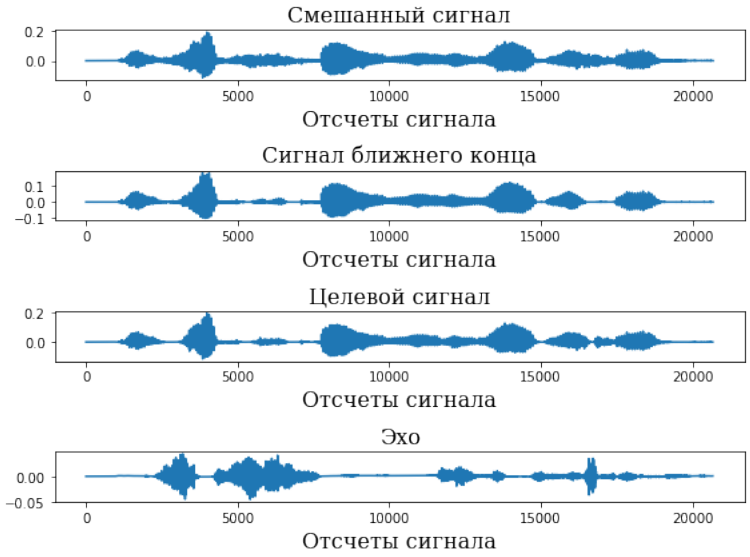

Рис. 10. Входной и выходной сигналы модели BLSTM+K-Means

3.4. Сравнение эффективностей моделей

Таблица 1 содержит значения метрик ERLE, PESQ и STOI, полученных четырьмя рассматриваемыми методами при $SER = 6$ дБ и $SER = 10$ дБ.

Таблица 1. Сравнение эффективности моделей

Метод	ERLE	PESQ	STOI
SER = 6 дБ			
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**
SER = 10 дБ			
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**

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.

Можно видеть, что и в случае $SER = 10$ дБ 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 статье 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 дальнейшем предложенная модель BLSTM+k-Means будет использована для задачи подавления акустического эха при наличии шума и нелинейных искажений.

Литература

1. Benesty J., Jensen J., Christensen M., Chen J. Speech Enhancement: A Signal Subspace Perspective. Elsevier Academic Press, 2014. 129 p. DOI: 10.1016/C2013-0-16082-5.
2. Lee C.M., Shin J.W., Kim N.S. DNN-based residual echo suppression // Interspeech 2015, Dresden, Germany, September 6–10, 2015. ISCA, 2015. P. 1775–1779. DOI: 10.21437/Interspeech.2015-412.
3. Zhang H., Wang D. Deep learning for acoustic echo cancellation in noisy and double-talk scenarios // Interspeech 2018, Hyderabad, India, September 2–6, 2018. ISCA, 2018. P. 3239–3243. DOI: 10.21437/Interspeech.2018-1484.
4. Zhang H., Tan K., Wang D. Deep learning for joint acoustic echo and noise cancellation with nonlinear distortions // Interspeech 2019, Graz, Austria, September 15–19, 2019. ISCA, 2019. P. 4255–4259. DOI: 10.21437/Interspeech.2019-2651.
5. Wang D. On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis // Speech Separation by Humans and Machines / ed. by P. Divenyi. Springer, Boston, MA, 2005. P. 181–197. DOI: 10.1007/0-387-22794-6_12.
6. Li N., Loizou P.C. Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction // J. Acoust. Soc. Am. 2008. Vol. 123, no. 3. P. 1673–1682. DOI: 10.1121/1.2832617.
7. Brungart D.S., Chang P.S., Simpson B.D., Wang D. Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation // J. Acoust. Soc. Am. 2006. Vol. 120, no. 6. P. 4007–4018. DOI: 10.1121/1.2363929.
8. Benesty J., Gänsler T., Morgan D.R., *et al.* Advances in network and acoustic echo cancellation. Springer, Berlin, Heidelberg, 2001. 222 p. DOI: 10.1007/978-3-662-04437-7.
9. Enzner G., Buchner H., Favrot A., Kuech F. Chapter 30 - Acoustic Echo Control // Academic Press Library in Signal Processing: Volume 4 / ed. by J. Trussell, A. Srivastava, A.K. Roy-Chowdhury, *et al.* Elsevier, 2014. P. 807–877. DOI: 10.1016/B978-0-12-396501-1.00030-3.
10. Hamidia M., Amrouche A. A new robust double-talk detector based on the Stockwell transform for acoustic echo cancellation // Digital Signal Processing. 2017. Vol. 60. P. 99–112. DOI: 10.1016/j.dsp.2016.09.001.

11. Ykhlef F., Ykhlef H. A post-filter for acoustic echo cancellation in frequency domain // 2014 Second World Conference on Complex Systems (WCCS), Agadir, Morocco, Nov. 10–12, 2014. IEEE, 2014. P. 446–450. DOI: 10.1109/ICoCS.2014.7060938.

12. Kuech F., Kellermann W. Nonlinear residual echo suppression using a power filter model of the acoustic echo path // 2007 International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, Honolulu, HI, USA, April 15–20, 2007. IEEE, 2007. P. I-73–I-76. DOI: 10.1109/ICASSP.2007.366619.

13. Malek J., Koldovský Z. Hammerstein model-based nonlinear echo cancelation using a cascade of neural network and adaptive linear filter // 2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC), Xi'an, China, Sept. 13–16, 2016. IEEE, 2016. P. 1–5. DOI: 10.1109/IWAENC.2016.7602906.

14. Yang F., Wu M., Yang J. Stereophonic acoustic echo suppression based on wiener filter in the short-time fourier transform domain // EEE Signal Processing Letters. 2012. Vol. 19, no. 4. P. 227–230. DOI: 10.1109/LSP.2012.2187446.

15. Wang D., Chen J. Supervised speech separation based on deep learning: an overview // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2018. Vol. 26, no. 10. P. 1702–1726. DOI: 10.1109/TASLP.2018.2842159.

16. Wang Y., Narayanan A., Wang D. On training targets for supervised speech separation // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2014. Vol. 22, no. 12. P. 1849–1858. DOI: 10.1109/TASLP.2014.2352935.

17. Hochreiter S., Schmidhuber J. Long Short-Term Memory // Neural Computation. 1997. Vol. 9, no. 8. P. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

18. Erdogan H., Hershey J.R., Watanabe S., Roux J.L. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks // 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, April 19–24, 2015. IEEE, 2015. P. 708–712. DOI: 10.1109/ICASSP.2015.7178061.

19. Weninger F., Erdogan H., Watanabe S., *et al.* Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR // Latent Variable Analysis and Signal Separation. Vol. 9237 / ed. by E. Vincent, A. Yeredor, Z. Koldovský, P. Tichavský. Cham: Springer International Publishing, 2015. P. 91–99. Lecture Notes in Computer Science. DOI: 10.1007/978-3-319-22482-4_11.

20. Chen J., Wang D. Long short-term memory for speaker generalization in supervised speech separation // The Journal of the Acoustical Society of America. 2017. Vol. 141, no. 6. P. 4705–4714. DOI: 10.1121/1.4986931.

21. Zermini A. Deep Learning for Speech Separation: PhD thesis / Zermini Alfredo. University of Surrey, faculty of engineering, physical sciences, Centre for Vision, Speech, Signal Processing (CVSSP), South East of England, UK, 2020. URL: https://openresearch.surrey.ac.uk/esploro/outputs/doctoral/99512310402346#file-0.

22. Xia S., Li H., Zhang X. Using Optimal Ratio Mask as Training Target for Supervised Speech Separation // 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, Malaysia, Dec. 12–15, 2017. IEEE, 2017. P. 163–166. DOI: 10.1109/APSIPA.2017.8282021.

23. Palmqvist M. Methods and algorithms for quality and performance evaluation of audio conferencing systems: PhD thesis / Palmqvist Maria. Umeå University, Faculty of Science, Technology, Department of Physics, Sweden, 2013. URL: http://umu.diva-portal.org/smash/get/diva2:630382/FULLTEXT01.pdf.

24. ITU-T Recommendation P. 862, Perceptual Evaluation of Speech Quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. 2001. URL: https://www.itu.int/rec/T-REC-P.862-200102-I/en.

25. Fu S.-W., Liao C.-F., Tsao Y. Learning with Learned Loss Function: Speech Enhancement with Quality-Net to Improve Perceptual Evaluation of Speech Quality // EEE Signal Processing Letters. 2020. Vol. 27. P. 26–30. DOI: 10.1109/LSP.2019.2953810.

26. Allen J.B., Berkley D.A. Image method for efficiently simulating small-room acoustics // The Journal of the Acoustical Society of America. 1998. Vol. 65, no. 4. P. 943–950. DOI: 10.1121/1.382599.

Шаход Джиах Михаил, магистрант, кафедра системного программирования, Южно-Уральский государственный университет (национальный исследовательский университет) (Челябинск, Российская Федерация)

Ибряева Ольга Леонидовна, к.ф.-м.н., доцент, кафедра системного программирования, Южно-Уральский государственный университет (национальный исследовательский университет) (Челябинск, Российская Федерация)

DOI: 10.14529/cmse220204

METHOD OF AN ACOUSTIC ECHO SUPPRESSION BASED ON RECURRENT NEURAL NETWORK AND CLUSTERING

© 2022 Gh.M. Shahoud, O.L. Ibryaeva

South Ural State University

(pr. Lenina 76, Chelyabinsk, 454080 Russia)

E-mail: ghiathlovealaa@gmail.com, ibriaevaol@susu.ru

Received: 01.04.2022

The article solves the problem of acoustic echo suppression based on a neural network that evaluates an ideal binary mask IBM using features extracted from a mixture of near-end and far-end signals. The novelty of the proposed method lies in the use of the clustering algorithm in addition to the bidirectional recurrent neural network BLSTM. To evaluate the use of the EM, Mean-Shift, k-Means clustering algorithms, the models have been trained and tested on the TIMIT database. For each model, the ERLE, PESQ, STOI metrics have been calculated to characterize its quality. The use of the EM and Mean-Shift clustering algorithms appeared to be inefficient compared to the BLSTM algorithm at a signal-to-echo ratio of 10 dB. With a signal-to-echo ratio of 6 dB, BLSTM+Mean-Shift resulted in a marginal improvement in the PESQ metric compared to the BLSTM algorithm. The results of the experiments show the effectiveness of the proposed BLSTM model when using a network with the K-Means algorithm, compared to using a pure BLSTM for echo cancellation in double-talk scenarios. With a signal-to-echo ratio of 10 dB, the STOI metric, which characterizes speech intelligibility, has improved by 7%, and the PESQ metric, which characterizes the quality of speech restoration, by 18.8%.

Keywords: ideal binary mask, near-end signal, far-end signal, bidirectional recurrent neural network, clustering, double-talk.

FOR CITATION

Shahoud Gh.M., Ibryaeva O.L. Method of an Acoustic Echo Suppression Based on Recurrent Neural Network and Clustering. Bulletin of the South Ural State University. Series: Computational Mathematics and Software Engineering. 2022. Vol. 11, no. 2. P. 43–58. (in Russian) DOI: 10.14529/cmse220204.

References

1. Benesty J., Jensen J., Christensen M., Chen J. Speech Enhancement: A Signal Subspace Perspective. Cambridge: Elsevier Academic Press, 2014. 129 p. DOI: 10.1016/C2013-0-16082-5.
2. Lee C.M., Shin J.W., Kim N.S. DNN-based residual echo suppression. Interspeech 2015, Dresden, Germany, September 6–10, 2015. ISCA, 2015. P. 1775–1779. DOI: 10.21437/Interspeech.2015-412.
3. Zhang H., Wang D. Deep learning for acoustic echo cancellation in noisy and double-talk scenarios. Interspeech 2018, Hyderabad, India, September 2–6, 2018. ISCA, 2018. P. 3239–3243. DOI: 10.21437/Interspeech.2018-1484.
4. Zhang H., Tan K., Wang D. Deep learning for joint acoustic echo and noise cancellation with nonlinear distortions. Interspeech 2019, Graz, Austria, September 15–19, 2019. ISCA, 2019. P. 4255–4259. DOI: 10.21437/Interspeech.2019-2651.

5. Wang D. On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis. Speech Separation by Humans and Machines / ed. by P. Divenyi. Springer, Boston, MA, 2005. P. 181–197. DOI: 10.1007/0-387-22794-6_12.

6. Li N., Loizou P.C. Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction. J. Acoust. Soc. Am. 2008. Vol. 123, no. 3. P. 1673–1682. DOI: 10.1121/1.2832617.

7. Brungart D.S., Chang P.S., Simpson B.D., Wang D. Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. J. Acoust. Soc. Am. 2006. Vol. 120, no. 6. P. 4007–4018. DOI: 10.1121/1.2363929.

8. Benesty J., Gänsler T., Morgan D.R., *et al.* Advances in network and acoustic echo cancellation. Springer, Berlin, Heidelberg, 2001. 222 p. DOI: 10.1007/978-3-662-04437-7.

9. Enzner G., Buchner H., Favrot A., Kuech F. Chapter 30 - Acoustic Echo Control. Academic Press Library in Signal Processing: Volume 4 / ed. by J. Trussell, A. Srivastava, A.K. Roy-Chowdhury, *et al.* Elsevier, 2014. P. 807–877. DOI: 10.1016/B978-0-12-396501-1.00030-3.

10. Hamidia M., Amrouche A. A new robust double-talk detector based on the Stockwell transform for acoustic echo cancellation. Digital Signal Processing. 2017. Vol. 60. P. 99–112. DOI: 10.1016/j.dsp.2016.09.001.

11. Ykhlef F., Ykhlef H. A post-filter for acoustic echo cancellation in frequency domain. 2014 Second World Conference on Complex Systems (WCCS), Agadir, Morocco, Nov. 10–12, 2014. IEEE, 2014. P. 446–450. DOI: 10.1109/ICoCS.2014.7060938.

12. Kuech F., Kellermann W. Nonlinear residual echo suppression using a power filter model of the acoustic echo path. 2007 International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, Honolulu, HI, USA, April 15–20, 2007. IEEE, 2007. P. I-73–I-76. DOI: 10.1109/ICASSP.2007.366619.

13. Malek J., Koldovský Z. Hammerstein model-based nonlinear echo cancelation using a cascade of neural network and adaptive linear filter. 2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC), Xi'an, China, Sept. 13–16, 2016. IEEE, 2016. P. 1–5. DOI: 10.1109/IWAENC.2016.7602906.

14. Yang F., Wu M., Yang J. Stereophonic acoustic echo suppression based on wiener filter in the short-time fourier transform domain. EEE Signal Processing Letters. 2012. Vol. 19, no. 4. P. 227–230. DOI: 10.1109/LSP.2012.2187446.

15. Wang D., Chen J. Supervised speech separation based on deep learning: an overview. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2018. Vol. 26, no. 10. P. 1702–1726. DOI: 10.1109/TASLP.2018.2842159.

16. Wang Y., Narayanan A., Wang D. On training targets for supervised speech separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2014. Vol. 22, no. 12. P. 1849–1858. DOI: 10.1109/TASLP.2014.2352935.

17. Hochreiter S., Schmidhuber J. Long Short-Term Memory. Neural Computation. 1997. Vol. 9, no. 8. P. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

18. Erdogan H., Hershey J.R., Watanabe S., Roux J.L. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, April 19–24, 2015. IEEE, 2015. P. 708–712. DOI: 10.1109/ICASSP.2015.7178061.

19. Weninger F., Erdogan H., Watanabe S., *et al.* Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR. Latent Variable Analysis and Signal Separation. Vol. 9237 / ed. by E. Vincent, A. Yeredor, Z. Koldovský, P. Tichavský. Cham: Springer, 2015. P. 91–99. Lecture Notes in Computer Science. DOI: 10.1007/978-3-319-22482-4_11.

20. Chen J., Wang D. Long short-term memory for speaker generalization in supervised speech separation. The Journal of the Acoustical Society of America. 2017. Vol. 141, no. 6. P. 4705–4714. DOI: 10.1121/1.4986931.

21. Zermini A. Deep Learning for Speech Separation: PhD thesis / Zermini Alfredo. University of Surrey, faculty of engineering, physical sciences, Centre for Vision, Speech, Signal Processing (CVSSP), South East of England, UK, 2020. URL: https://openresearch.surrey.ac.uk/esploro/outputs/doctoral/99512310402346#file-0.

22. Xia S., Li H., Zhang X. Using Optimal Ratio Mask as Training Target for Supervised Speech Separation. 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, Malaysia, Dec. 12–15, 2017. IEEE, 2017. P. 163–166. DOI: 10.1109/APSIPA.2017.8282021.

23. Palmqvist M. Methods and algorithms for quality and performance evaluation of audio conferencing systems: PhD thesis / Palmqvist Maria. Umeå University, Faculty of Science, Technology, Department of Physics, Sweden, 2013. URL: http://umu.diva-portal.org/smash/get/diva2:630382/FULLTEXT01.pdf.

24. ITU-T Recommendation P. 862, Perceptual Evaluation of Speech Quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. 2001. URL: https://www.itu.int/rec/T-REC-P.862-200102-I/en.

25. Fu S.-W., Liao C.-F., Tsao Y. Learning with Learned Loss Function: Speech Enhancement with Quality-Net to Improve Perceptual Evaluation of Speech Quality. EEE Signal Processing Letters. 2020. Vol. 27. P. 26–30. DOI: 10.1109/LSP.2019.2953810.

26. Allen J.B., Berkley D.A. Image method for efficiently simulating small-room acoustics. The Journal of the Acoustical Society of America. 1998. Vol. 65, no. 4. P. 943–950. DOI: 10.1121/1.382599.