УДК 004.032.26 DOI 10.25730/VSU.0536.19.021

Метод шумоподавления в речевых сигналах с помощью нейронной сети

С. Г. Лялин
аспирант кафедры электронных вычислительных машин, преподаватель кафедры систем автоматизации управления, Вятский государственный университет.
Россия, г. Киров. E-mail: lyalinstas@gmail.com

Аннотация. Задача идентификации диктора по голосу имеет широкую сферу применения. В частности, это может быть обеспечение безопасного входа на веб-сайт. В этом случае идентифицирующая сторона не может контролировать звукозаписывающее оборудование и шумовую обстановку пользователя, поэтому речевые сигналы могут оказаться зашумленными и не пригодными для принятия достоверного решения. Появляется необходимость в предварительной фильтрации входных данных от помех. Подвергаемые фильтрации цифровые данные должны сохранить максимум значимой информации, используемой для вычисления голосовых признаков. Одним из вариантов создания цифровых фильтров сегодня является применение искусственных нейронных сетей.

В данной статье описан метод фильтрации голосовых записей от шума с помощью искусственной нейронной сети. Разработанный метод протестирован на базе голосовых записей, записанных в различных реальных условиях, и показал свою эффективность. Он может применяться в любой сфере, где требуется фильтрация помех в речевых сигналах.

Ключевые слова: речевой сигнал, подавление шума, искусственная нейронная сеть, машинное обучение, задача предсказания.

Введение. Наличие помех в речевых данных может существенно осложнять задачу идентификации диктора по голосу, так как голосовые признаки по этим данным могут быть вычислены неверно. В результате точность идентификации будет низкой и система будет не пригодна к использованию. В то же время в реальных системах далеко не всегда есть возможность контролировать звукозаписывающее оборудование и шумовую обстановку, в которой осуществляется запись речевых данных, поэтому единственным вариантом решения проблемы остается фильтрация сигнала после записи. Кроме того, в реальных сигналах АЧХ шума распределена неравномерно, и от диктора к диктору она может сильно отличаться. Поэтому задача фильтрации сигнала является нетривиальной.

Цель исследования. Целью данного исследования является разработка универсального метода автоматической фильтрации речевых данных от шума, учитывающего особенности шумовой обстановки для каждой голосовой записи.

Задачи исследования. В процессе исследования необходимо разработать метод фильтрации голосовых записей от помех, подготовить базу голосовых данных для настройки и тестирования алгоритма, провести соответствующие эксперименты.

Методы исследования. Исходными данными в задаче являются голосовые данные. Они представляют собой одноканальные wav-файлы с частотой дискретизации 16 кГц и длительностью порядка 3-5 секунд. Файлы содержат речевые данные на английском языке реальных пользователей различной национальности, пола и возраста. Данные были получены с сайта voxforge.org. Эти голосовые записи сформированы непосредственно пользователями с учетом собственного записывающего оборудования, поэтому на многих файлах присутствует шум. Из множества записей было отобрано 20 файлов для настройки и тестирования метода.

Идея метода состоит в следующем. Как правило, диктор начинает говорить не сразу в момент включения записи и выключает запись не сразу по окончании фразы. Поэтому в начале и конце голосовой записи присутствуют участки, на которых отсутствует речь, но ясно различим шум. Эти участки могут использоваться для построения модели шума. Шум, извлеченный из этих участков, может быть наложен поверх исходной голосовой записи. Тогда можно считать, что исходная запись – это некий эталон, а запись с наложенным поверх шумом – зашумленный сигнал, который требуется научиться фильтровать, то есть приводить его к эталону. Таким образом, полученные данные могут являться обучающими для некоторого алгоритма машинного обучения, который по зашумленному сигналу учится восстанавливать эталонный. Результатом будет являться обученная модель, применив которую к исходному (эталонному) сигналу, можно получить очищенный от шумов сигнал. Обученная модель в данном случае будет выступать неким нелинейным цифровым фильтром.

Рассмотрим далее приведенный выше алгоритм более подробно.

Изначально голосовая запись подвергается процедуре стандартизации. Эта процедура выравнивает амплитуду сигнала, центрируя среднее значение амплитуды и устанавливая среднеквадратическое отклонение, равное единице (1).

$$x = \frac{x - \mu}{\sigma} \tag{1}$$

Необходимость процедуры стандартизации амплитуды вызвана тем, что исходные данные в голосовых записях не нормированы (могут быть очень малы или очень велики, различны для всех пользователей). По этой причине задавать универсальные пороговые значения невозможно, без стандартизации они будут разными для всех пользователей. Стандартизация решает эту проблему. На рисунке 1 приведен исходный речевой сигнал и тот же сигнал после процедуры стандартизации.

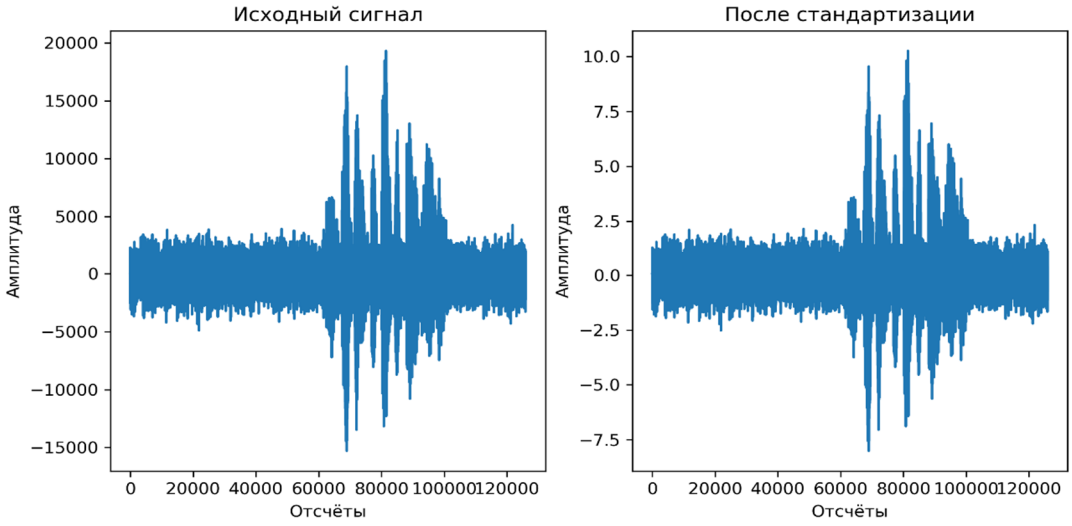


Рис. 1. Сигнал до и после процедуры стандартизации

Следующий этап алгоритма состоит в поиске участков шума в начале и конце файла. Для этого строится огибающая сигнала, по которой затем путем сравнения с некоторым пороговым значением выбираются участки, соответствующие шуму. Для вычисления огибающей выполняется свертка модуля отсчетов сигнала с прямоугольным окном длительностью 30 мс. Данная длительность получена эмпирически, путем наблюдения за файлами из выборки. Если задать значение меньше, огибающая будет подвержена локальным всплескам мощности сигнала. Если задать значение больше, будут сглажены важные фрагменты сигнала. На рисунке 2 слева показан сигнал, вычисленная для него огибающая и выбранное пороговое значение. На том же рисунке справа показан увеличенный фрагмент графика. Значения, меньшие порога, из начала и конца файла принимаются за шум. На рисунке 2 эти участки выделены прямоугольниками.

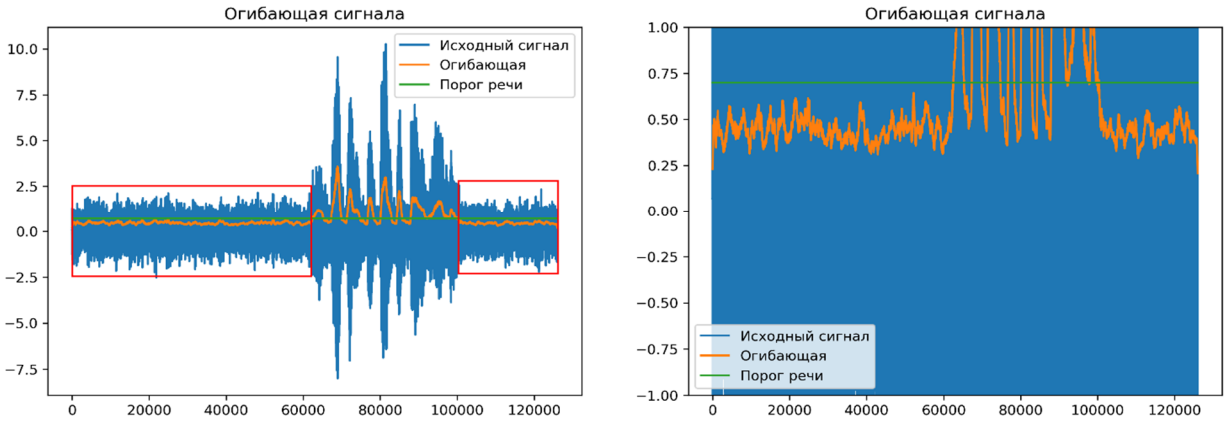


Рис. 2. Огибающая сигнала и порог отсечения шума

В качестве порога отсечения шума выбрано значение 0.7. Значение также выбрано эмпирически после исследования всех файлов выборки.

На следующем шаге алгоритма участок шума увеличивается по длительности до размера исходного сигнала, чтобы его можно было наложить поверх исходного сигнала, который будет выступать эталоном при обучении.

Далее выполняется фильтрация шума ФНЧ с частотой среза 4 кГц. Используется фильтр Баттерворта порядка 8. Данный фильтр нужен для того, чтобы исключить из фильтрации нейронной сетью неинформативную область высоких частот. Применение фильтра дает следующий эффект. Область высоких частот (от 4 кГц) не содержит информации о речи, частотная полоса которой обычно не превышает 3 кГц [1], поэтому при итоговой фильтрации сигнала нейронной сетью эту область можно игнорировать. Когда к исходному (эталонному) сигналу добавляется шум, пропущенный через ФНЧ, сигнал в области высоких частот у обоих сигналов будет одинаковым. Тогда нейронная сеть при обучении будет обращать внимание только на значимую область низких частот, не изменяя высокочастотную область. Это не имеет принципиального значения для технического анализа голосовых записей, но дает меньше искажений при восприятии отфильтрованного сигнала «на слух».

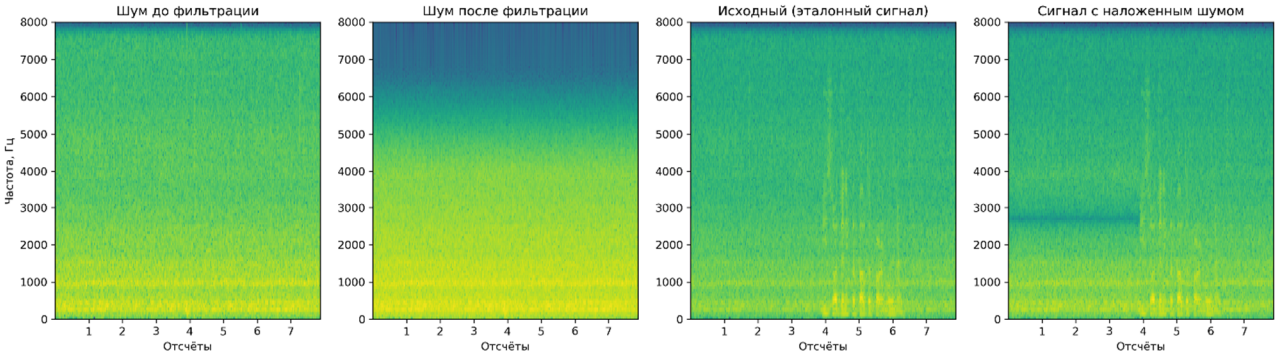
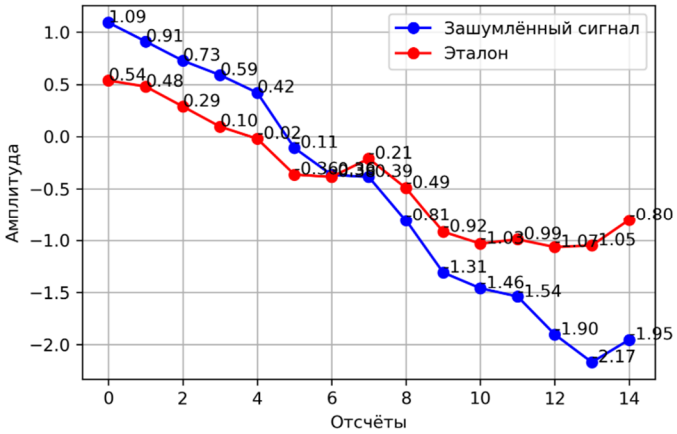


Рис. 3. Шум, отфильтрованный шум, эталонный сигнал и сигнал с наложенным шумом

На рисунке 3 показан шум, извлеченный из исходной записи, затем шум, пропущенный через ФНЧ с полосой 4 кГц, затем исходный (эталонный) сигнал и тот же сигнал, но с добавленным шумом.

На следующем этапе формируется обучающая выборка для алгоритма машинного обучения. В данном случае решается задача предсказания амплитуды отсчета по амплитудам p предыдущих отсчетов. Матрица входов будет иметь размер $m \times p$, где m – количество примеров, p – количество отсчетов из зашумленного сигнала, по которым будет предсказываться отсчет в отфильтрованном сигнале. Вектор выходов имеет размерность $m \times 1$ и содержит значения амплитуды отсчетов из эталонного сигнала. Рисунок 4 поясняет процесс формирования выборки.



1.09	0.91	0.73	0.10
0.91	0.73	0.59	-0.02
0.73	0.59	0.42	-0.36
	• • •		
-1.54	-1.90	-2.17	-0.80

Зашумлённые данные Эталон

Рис. 4. Формирование выборки для алгоритма машинного обучения

На рисунке 4 приведен пример формирования выборки для $p = 3$. 3 первых отсчета зашумленного сигнала являются вектором признаков первого обучающего примера. Четвертый отсчет эталонного сигнала является целевым значением первого обучающего примера. Затем выполняется сдвиг на один отсчет и формируется следующая строка таблицы, и так до тех пор, пока не будет

достигнут конец файла. В результате получается матрица признаков, содержащая данные из зашумленного сигнала, и вектор ответов, содержащий данные из эталонного (исходного) сигнала.

Затем все примеры выборки случайным образом перемешиваются и делятся в отношении 70% на 30%. Первая часть используется в качестве обучающих примеров, вторая – в качестве контрольных при кросс-валидации.

После того как исходные данные подготовлены, выполняется построение модели. Существует большое количество методов машинного обучения, решающих задачу регрессии. В данной работе выбрана искусственная нейронная сеть (многослойный персептрон), так как данная модель является обобщением более простых моделей (линейной и логистической регрессий) и способна моделировать сложные нелинейные зависимости. На рисунке 5 приведена архитектура нейронной сети.

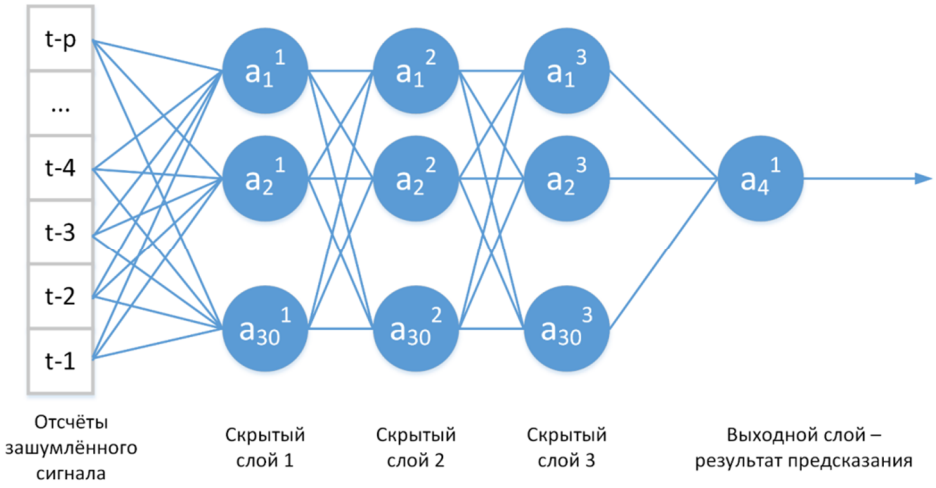


Рис. 5. Архитектура нейронной сети

Архитектура состоит из 5 слоев. Первый слой – входной. Сюда поступают значения из выборки – зашумленные значения, которые сеть должна научиться фильтровать. Затем следуют 3 скрытых слоя по 30 нейронов в каждом с функцией активации LeakyReLU, которая является усовершенствованным вариантом функции активации ReLU. Для ReLU характерны простота реализации, эффективность и отсутствие проблемы затухания градиентов при обучении сети. LeakyReLU считается улучшенной версией ReLU и дает большую точность на большинстве тестов [2]. Последний слой – выходной – содержит один нейрон с линейной функцией активации, так как сеть должна предсказывать непрерывное значение целевой переменной – амплитуду очищенного от шума сигнала.

Кроме того, при обучении сети использовался подход EarlyStopping, который препятствует переобучению модели. Суть его заключается в том, что обучение останавливается в тот момент, когда на кривой обучения ошибка на кросс-валидационной (контрольной) выборке, достигнув плато, начинает увеличиваться, в то время как ошибка на обучающей выборке стабильно снижается. Еще одним достоинством такого подхода является ускорение процесса обучения. Типичный график кривой обучения приведен на рисунке 6.

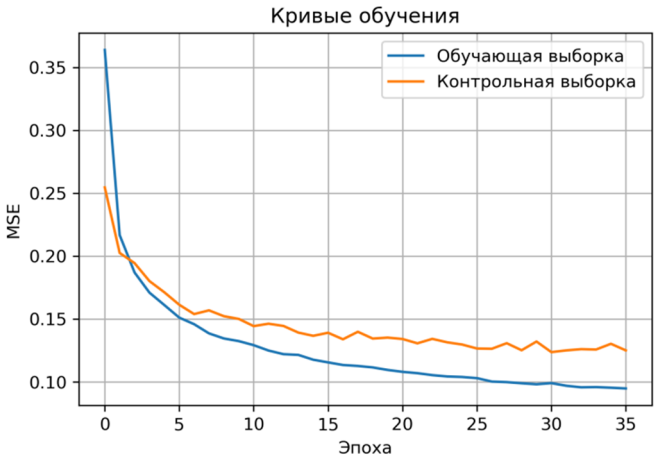


Рис. 6. Кривые обучения

Как было показано выше, предсказание отфильтрованного отсчета выполняется по предыдущим отсчетам. При этом чем выше значение p, тем большая предыстория учитывается при построении прогноза. Для поиска оптимального значения параметра p были рассмотрены значения из ряда (10, 30, 100, 300, 1000). Для каждого значения на каждом из файлов выборки была посчитана среднеквадратическая ошибка (MSE) между фактическими значениями эталонного сигнала и теми значениями, которые предсказала нейронная сеть. Очевидно, что чем меньше MSE, тем лучше модель предсказывает значения, иными словами, тем лучше она отфильтровывает шум. Данные поиска оптимального значения p, усредненные по всем файлам выборки, приведены в таблице 1 и на рисунке 7.

Таблица 1

MSE для зашумленного и отфильтрованного сигнала при разных p

Значение	10	30	100	300	1000
MSE	1.928935	1.937442	1.872025	1.918838	**1.502585**

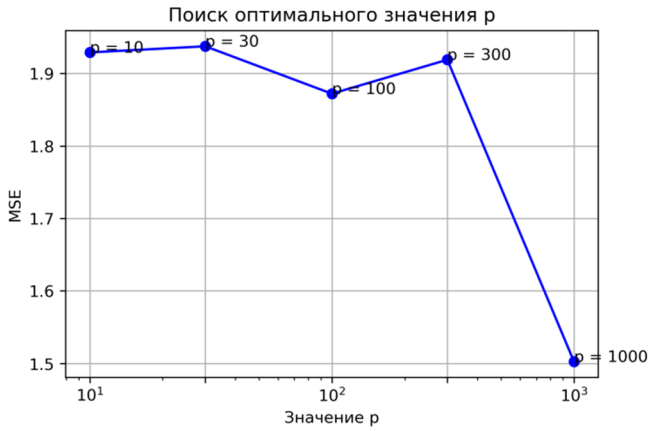


Рис. 7. MSE для зашумленного и отфильтрованного сигнала при разных p

Результаты исследований, их обсуждение. Таким образом, точнее всего прогноз получается при значении $p = 1000$. Именно с этим значением далее проводилось тестирование алгоритма для различных файлов.

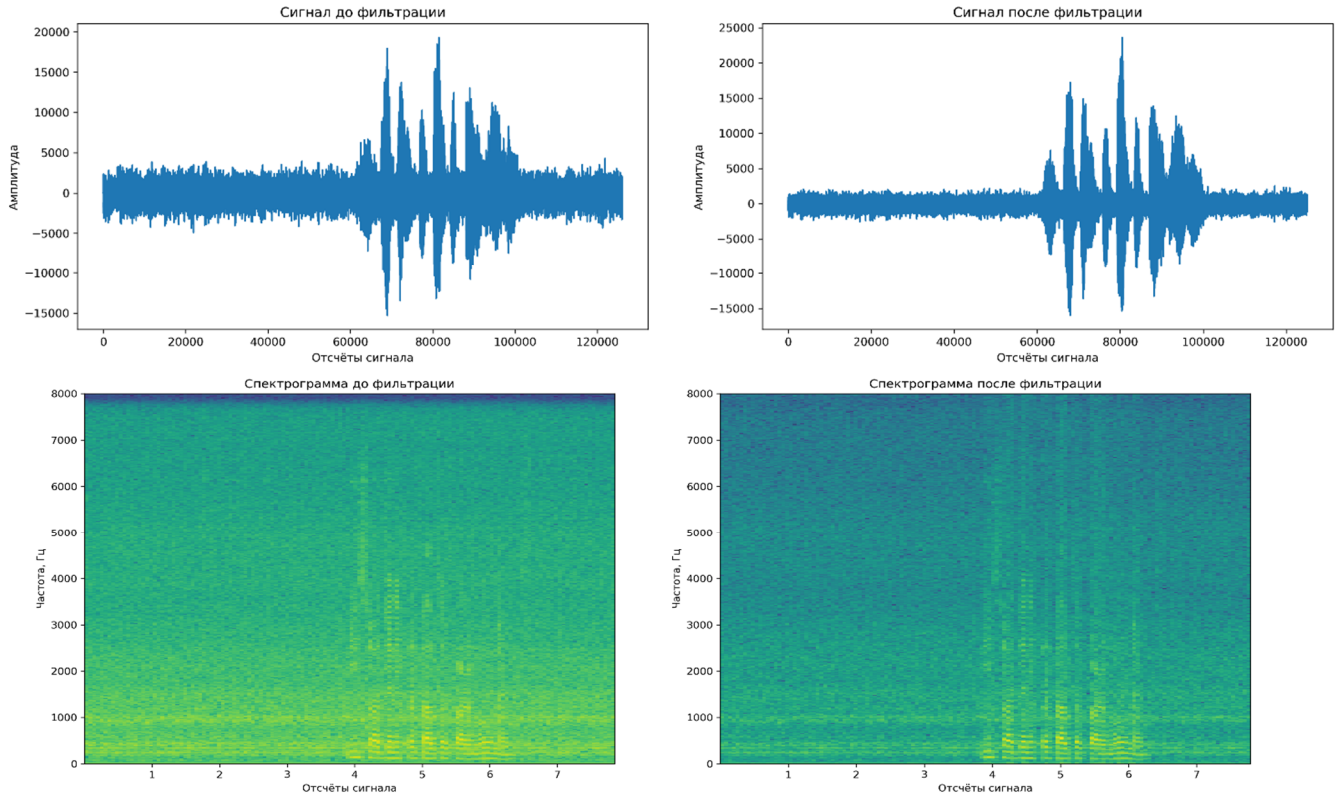


Рис. 8. Результат фильтрации сигнала

На рисунке 8 отчетливо видно, что уровень шума уменьшился, при этом значимая информация сохранилась.

Кроме того, одним из вариантов тестирования сигнала является проверка на искусственно сгенерированном эталонном сигнале. В качестве такого искусственного сигнала выступает синусоида с частотой 1 кГц. К этому сигналу сначала добавляется шум, а затем выполняется фильтрация зашумленного сигнала обученной моделью. Результат фильтрации показан на рисунке 9. Видно, что, несмотря на высокий уровень шума, исходная синусоида была хорошо восстановлена. На спектрограмме отфильтрованного сигнала, кроме того, видны дополнительные частоты с шагом 1 кГц, которых не было в исходном эталонном сигнале. Природа этого явления пока неясна. Тем не менее можно сделать вывод, что модель работоспособна и справляется с задачей фильтрации речевого сигнала.

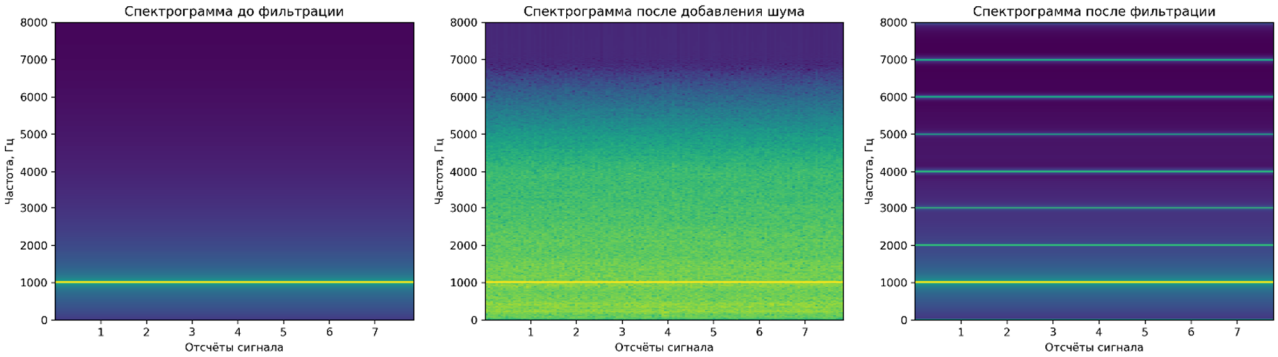


Рис. 9. Проверка модели на искусственно сгенерированной синусоиде с частотой 1 кГц

Выводы. В данной работе был описан метод фильтрации голосовых сигналов с помощью искусственной нейронной сети. Метод является универсальным и полностью автоматическим. Метод был протестирован на базе из 20 сигналов, записанных в разных условиях с использованием разного оборудования, и показал свою эффективность – уровень шума снижается, при этом вся значимая информация сохраняется.

К недостаткам метода можно отнести появление некоторых артефактов, которые отсутствовали в исходном сигнале (см. рис. 9). Кроме того, процесс обучения модели выполняется довольно долго – на процессоре Intel Core i3 6 поколения – порядка 20 секунд на 3-5 секундный файл. В качестве мер по повышению производительности метода можно предложить обучать модель не на всех данных, а на некотором их подмножестве, а также упростить архитектуру сети. Еще одним потенциальным улучшением метода может стать итеративная фильтрация – запуск фильтрации на одном файле несколько раз подряд для более существенного подавления шума.

В целом, разработанный метод является универсальным и подходит для использования в широком спектре систем, например, при очистке от шума записей телефонных разговоров.

Список литературы

1. Частота голоса. URL: https://ru.wikipedia.org/wiki/Частота_голоса (дата обращения: 10.02.2019).
2. Это нужно знать: Ключевые рекомендации по глубокому обучению (Часть 2). URL: http://datareview.info/article/eto-nuzhno-znat-klyuchevyie-rekomendatsii-po-glubokomu-obucheniyu-chast-2 (дата обращения: 10.02.2019).

Noise reduction method in speech signals using neural network

S. G. Lyalin
post-graduate student of the Department of electronic computing machines,
lecturer of the Department of control automation systems, Vyatka State University.
Russia, Kirov. E-mail: lyalinstas@gmail.com

Abstract. The task of speaker identification by voice has a wide scope of application. In particular, it can be used to provide a secure login to the website. In this case, the identifying party cannot control the recording equipment and the noise environment of the user, so the voice signals may be noisy and not suitable for making a reliable decision. There is a need to pre-filter the input data from interference. Subjected to the filtering of digital data need to maintain maximum meaningful information used to evaluate voice characteristics. One of the options for creating digital filters today is the use of artificial neural networks.

This article describes a method of filtering voice recordings from noise using an artificial neural network. The developed method is tested on the basis of voice recordings recorded in various real conditions, and has shown its effectiveness. It can be used in any field where noise filtering in speech signals is required.

Keywords: speech signal, noise suppression, artificial neural network, machine learning, prediction problem.

References

1. *Chastota golosa* – The frequency of the voice. Available at: https://ru.wikipedia.org/wiki/Частота_голоса (date accessed: 10.02.2019).

2. *Eto nuzhno znat': Klyuchevye rekomendacii po glubokomu obucheniyu (Chast' 2)* – This should be known: Key recommendations for deep learning (Part 2). Available at: http://datareview.info/article/eto-nuzhno-znat-klyuchevyie-rekomendatsii-po-glubokomu-obucheniyu-chast-2 (date accessed: 10.02.2019).