УДК 621.317+681.849

В.И. СОЛОВЬЕВ, О.В. РЫБАЛЬСКИЙ, В.В. ЖУРАВЕЛЬ, Н.В. СЕМЕНОВА

ИССЛЕДОВАНИЯ МОДЕЛЕЙ РАСПОЗНАВАНИЯ ЗВУКОВ РЕЧИ НА ОСНОВЕ НЕЙРОННЫХ СЕТЕЙ ГЛУБОКОГО ОБУЧЕНИЯ ДЛЯ ЭКСПЕРТИЗЫ ЦИФРОВЫХ ФОНОГРАММ

Аннотация. Исследованы модели на основе нейронных сетей глубокого обучения, базирующиеся на общем подходе к паузам и сигналам речи как разным видам звуковой информации, зафиксированной в фонограмме, отличающимся по некоторым характеристикам. Такой подход позволяет формировать базы данных обучения с использованием общих для пауз и сигналов речи методов предварительной обработки информации, что обеспечивает более высокий уровень унификации методов обучения сетей, предназначенных для решения разных задач экспертизы.

Ключевые слова: аппаратура цифровой звукозаписи, база данных обучения, нейронная сеть глубокого обучения, цифровая обработка фонограмм, цифровая фонограмма, экспертиза.

В работе [1] проведен анализ современных исследований инструментария, предназначенного для экспертизы материалов и аппаратуры цифровой звукозаписи. Показана также возможность и целесообразность применения нейронных сетей глубокого обучения при создании полного комплекта инструментария для ее проведения. Однако построение такого инструментария без использования средств автоматизации подготовки больших массивов данных (Dataset) невозможно. Это обусловлено необходимостью разработки сотен тысяч эталонных образцов, требуемых как для обучения нейронной сети, так и построения кривых ошибок первого и второго рода. Точка пересечения этих кривых часто используется в качестве универсального критерия эффективности такого инструментария для любой экспертизы сложных технических объектов [2].

Известно, что в минимальный комплект инструментария для проведения экспертизы материалов и аппаратуры цифровой звукозаписи должны входить средства для идентификационных и диагностических исследований аппаратуры записи и фонограмм, а также сигналов речи дикторов, зафиксированных в фонограммах [3]. В процессе проведения такой экспертизы устанавливаются оригинальность изучаемых фонограмм, отсутствие в них следов монтажа, а также выполняется идентификация личности диктора по физическим параметрам сигналов его речи [4].

Весь комплекс этих задач можно решить с помощью программного обеспечения, построенного на основе нейронных сетей глубокого обучения [1, 5]. Успешность решения в значительной степени определяется формированием данных и их структурированием в нейронной сети.

Целью настоящей работы является исследование универсальных моделей на основе нейронных сетей, предназначенных для решения большинства задач экспертизы материалов и аппаратуры цифровой звукозаписи.

Многообразие задач, решаемых в процессе экспертизы, предполагает использование различных данных и методов обучения нейронных сетей. Однако это не всегда так, поскольку сигналы, обрабатываемые в таких сетях, имеют общую основу — фонограмму, записанную на аппаратуре цифровой звукозаписи. Рассмотрим методы формирования обучающей базы данных для разных задач экспертизы более подробно. Определим общие элементы решений, присущие разнородным задачам. Известно, что для выявления следов цифрового монтажа

требуется сегментация фонограмм на паузы, а для идентификации диктора по физическим параметрам сигналов речи — на отдельные звуки [6]. При этом идентификация диктора иногда предполагает использование в инструментарии элементов автоматической классификации и сегментации звуков речи [7]. Однако паузы между речевыми сигналами и сигналы речи сопровождаются собственными шумами аппаратуры записи, зафиксированными на фонограмме. Все сигналы расположены в звуковом диапазоне частот, а в паузах, кроме шумов, присутствуют звуковые сигналы окружающей среды.

В связи с этим рассмотрим модель формирования базы данных обучения, основанную на общем подходе к паузам и сигналам речи как разным видам звуковой информации, зафиксированной в фонограмме, отличающимся по некоторым характеристикам. Такой подход позволяет формировать базы данных обучения на общих методах предварительной обработки информации, содержащейся и в паузах, и в сигналах речи, что обеспечивает более высокий уровень унификации обучения при решении разных задач экспертизы. При этом, несмотря на общность такого подхода, сохраняются особенности решаемых задач.

Например, задача выявления монтажа основана на модели пауз двух видов как составной части моделей звуков речи, построенной на нейронных сетях глубокого обучения. Паузы как объекты классификации разбиваются на два вида: «чистые паузы» и «паузы с монтажом». Для обозначения фрагментов «чистых пауз» в базе данных далее использован специальный символ [pau], для пауз с монтажом — [paue].

Для обучения сети, решающей задачу идентификации диктора, классификация входных сигналов проводится другим способом. При обучении сетей для обеих экспертных задач можно создать общую первичную базу данных звуков и пауз (Dataset Sounds). На первом этапе формирования базы данных используется «нарезка» звуков и пауз, выполненная в звуковом редакторе, обеспечивающем возможность реализации операции в «ручном режиме». В первичную базу входят фрагменты фонограмм с различным контекстом, записанных разными дикторами на украинском, русском, английском и китайском языках.

Для «нарезки» сигналов речевой информации отбирались сигналы, выбранные из ограниченного перечня звуков, который приведен далее в транскрипции, принятой в системе API Международной фонетической организации. Из фонограмм, записанных на украинском, русском и английском языках, отбирались:

- гласные звуки — [a], [e], [i], [i:], [o], [u];
- согласные звуки — [b], [c], [d], [f], [g], [j], [k], [l], [m], [n], [p], [r], [s], [t], [v], [w], [x], [z], [ʃ], [ð];
- мягкие согласные звуки — [b’], [d’], [l’], [n’], [p’], [t’], [r’], [s’].

Из фонограмм, записанных на китайском языке, в первичную базу отбирались только гласные звуки.

Данный перечень не охватывает всех звуков, содержащихся в исследуемых фонограммах. Поэтому звуки, отличающиеся от входящих в перечень, при классификации в реальных фонограммах будут отнесены к какому-либо из классификационных объектов. Однако вероятность их правильной классификации в рамках принятой модели мала. Это позволит отбрасывать такие звуки при последующем применении программы, созданной на основе изложенных принципов, в автоматическом режиме.

Отметим, что «нарезка» звуковых фрагментов фонограмм осуществлялась при их прослушивании экспертом, исходя из его субъективного восприятия пауз и звуков, входящих в приведенный перечень. Первичные фрагменты гласных и ряда согласных звуков, воспринимаемые на слух, имели различную длительность, в ряде случаев до 100 и более мс.

На основе полученной первичной базы фрагментов звуков и пауз формировалась исходная база данных для обучения, тестирования и исследования свойств рассматриваемых моделей. Дальнейшее формирование исходной базы

данных обучения для разных задач выполнялось в автоматическом режиме с помощью специального программного модуля.

В частности, для формирования в исходной базе данных пауз с монтажом [paue] модуль осуществлял автоматическое формирование и моделирование монтажа на основе «чистых пауз» при большом объеме статистического материала. База данных [paue] содержала сотни тысяч фрагментов пауз с монтажом. Модель автоматического формирования базы данных пауз с монтажом обеспечивала большую вариабельность характеристик видов монтажа.

Сигналы исходной базы подвергались предварительной обработке. Каждый ранее обозначенный фрагмент из первичной базы разбивался на фрагменты длительностью 20 мс. Преобразование сигналов этих фрагментов из временной в частотную область выполнялось вейвлет-преобразованием на основе вейвлета Морле. Затем в сигнале выделялось 70 наибольших по спектру локальных максимумов [8]. Отобранные сигналы подвергались частотной фильтрации полосовым фильтром с частотной характеристикой, обратной характеристике кривых равной громкости (фильтр-ухо) [9], поскольку экспериментально установлено, что такая фильтрация способствует повышению вероятности правильной классификации.

Из фрагментов длительностью 20 мс, прошедших предварительную обработку, были сформированы три независимых массива данных: обучающий массив (Train — для обучения сети), тестовый массив (Test — для оценки эффективности выбранной модели) и верификационный массив (на его основе строились графики ошибок первого и второго рода исследуемой модели).

В первый слой нейронной сети вводились параметры 70 локальных максимумов — 70 амплитуд нормированных спектров и значений их частоты (всего 140 параметров). Применялась библиотека keras (backend tensorflow) и использовалась полносвязная нейронная сеть, где число слоев доходило до 50.

Эффективность обучения нейронных сетей, предназначенных для решения задач разделения сигналов, которые выделены из фонограммы в автоматическом режиме и основаны на предложенных методах формирования обучающей базы из фрагментов звуковых сигналов длительностью 20 мс, проиллюстрирована на рис. 1 при множественной классификации и на рис. 2 — при бинарной классификации. На рис. 3 приведены графики эффективности обучения сети, предназначенной для автоматического выявления пауз, содержащих и не содержащих признаки монтажа.

Таким образом, предложенный подход позволил осуществить обучение и проверку качества обучения нейронной сети для множественной и бинарной классификации фрагментов фонограмм длительностью 20 мс.

Отметим, что для множественной классификации эффективность распознавания звуков недостаточно высока. Однако при работе с реальными фонограммами

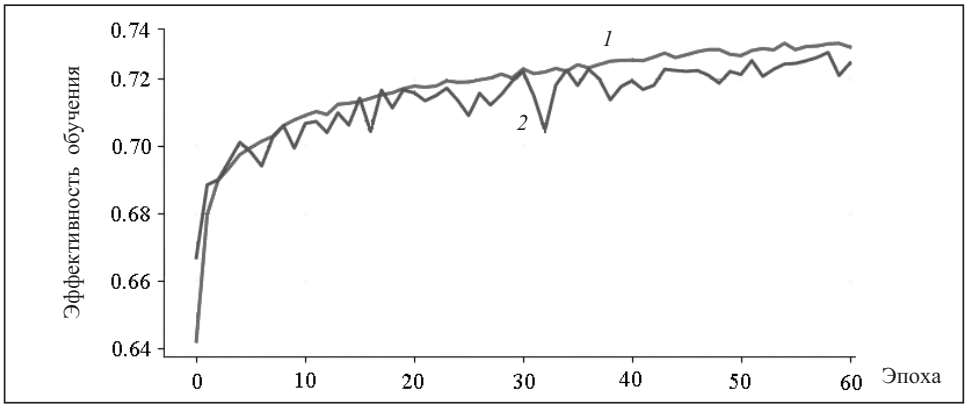

Рис. 1. Графики эффективности обучения нейронной сети при множественной классификации звуков и сигналов пауз по фрагментам длительностью 20 мс: тренировочный массив (*1*); тестовый массив (*2*), на котором максимум эффективности за 59 эпох составляет 0.7277

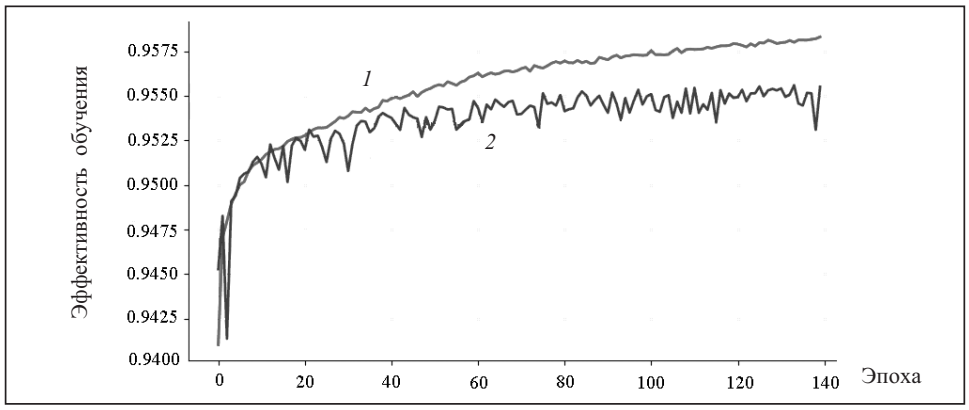

Рис. 2. Графики эффективности обучения нейронной сети при бинарной классификации звуков речи и пауз по фрагментам длительностью 20 мс: тренировочный массив (*1*); тестовый массив (*2*), на котором максимум эффективности за 134 эпохи составляет 0.9556

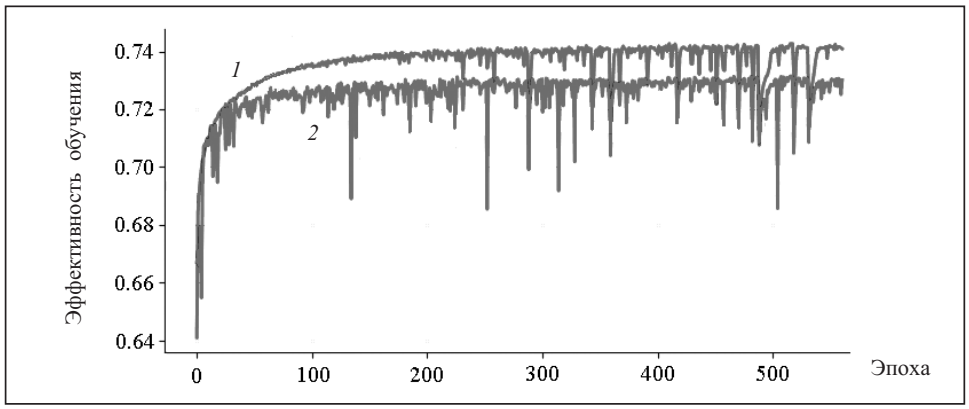

Рис. 3. Графики эффективности обучения нейронной сети при бинарной классификации пауз по фрагментам длительностью 20 мс: тренировочный массив без признаков монтажа (*1*); тестовый массив с признаками монтажа (*2*), на котором максимум эффективности за 517 эпох составляет 0.7315

эффективность модели можно существенно увеличить. Повышение эффективности классификации сигналов при решении поставленной задачи обусловлено некоторыми, рассмотренными ниже, особенностями речи. Увеличение эффективности при множественной классификации фрагментов звуков важно при решении экспертной задачи идентификации диктора по голосу.

Результатом работы модели является прогноз вероятности классификации для каждого объекта множественной классификации. Как правило, при принятии решения выбирается объект с максимальной вероятностью классификации (например, звук [a]). Это использовалось для классификации фрагментов длительностью 20 мс. Проверка модели на реальных фонограммах осуществлялась ее сканированием окном длительностью 20 мс с шагом сканирования 2 мс, что приблизительно соответствует длительности паузы между импульсациями нервной клетки слухового аппарата человека [10].

Запишем, например, английское слово this в транскрипции так, как оно реально воспринимается при его сканировании сдвигающимся окном длительностью 20 мс в идеальной модели с вероятностью правильной классификации, равной единице:

[ð][ð][ð][ð][i:][i:][i:][i:][i:][i:][i:][i:][i:][i:][s][s][s][s][s][s].

В случае использования модели с реальной классификацией в рамках данной методологии при сканировании наблюдается существенно отличающаяся транскрипция, например:

[ð][z][ð][ð][i:][i][i:][i][i:][e][i:][i:][i:][i:][s][s][s][c][s][s].

Для повышения эффективности классификации применяется усреднение во времени сканирования на интервале до 30–35 мс. Для большинства реальных фрагментов речи образуются еще более сложные структуры.

Рассмотрим данный эффект на примере английского слова mother. В транскрипции разработанной модели оно может иметь следующий вид:

[m][m][m][m][pau][pau][paue][a][a][o][a][a][a][o][a][a][a][pau][paue][ð][ð][z][ð][ð][ð][ð][pau][pau][e][i][e][e][i:][e][i][e][e].

Появление [pau] и [paue] вне фрагментов пауз на стыке звуков имеет физическое объяснение. При сканировании скользящим окном длительностью 20 мс на стыке согласных и гласных звуков при частичном перекрытии окном двух звуков спектр в окне является средневзвешенным спектром двух звуков. Он часто приближается к спектру модели пауз. Поскольку пауз столь малой длительности (не более 12 мс) в речевом сигнале не бывает, эти фиктивные паузы легко удаляются применением логического анализа. Иногда такой спектр приближается к спектру других звуков для рассматриваемой множественной классификации. Этот эффект можно наблюдать и при прослушивании соответствующего фрагмента речи.

Предложенная модель достаточно эффективна при решении задачи идентификации диктора, что проиллюстрировано кривыми ошибок первого и второго рода, показанными на рис. 4. Данная модель обеспечивает также эффективное решение задачи сегментации пауз в автоматическом режиме, что необходимо при локализации потенциальных точек цифрового монтажа в фонограммах (рис. 5), и эффективное автоматическое выявление пауз с монтажом (рис. 6). В экспертизе материалов и аппаратуры звукозаписи точку пересечения кривых ошибок принято считать минимальной эффективностью экспертного инструментария. На рис. 4 проиллюстрирована эффективность применения предложенной модели для распознавания звуков в реальных речевых фонограммах (на примере звука [e]). Полученные кривые ошибок первого рода (*1*) и второго рода (*2*) построены в координатах: вероятность ошибки — вероятность прогноза классификации (модели). Графики на рис. 4–6 построены на основе большого объема статистического материала. Графики на рис. 5 иллюстрируют эффективность разделения звуков и пауз речи в реальных фонограммах.

Как видно на рис. 6, минимальная эффективность выявления и локализации потенциальных мест монтажа равна 0.0591, что является удовлетворительным результатом для системы, работающей в автоматическом режиме.

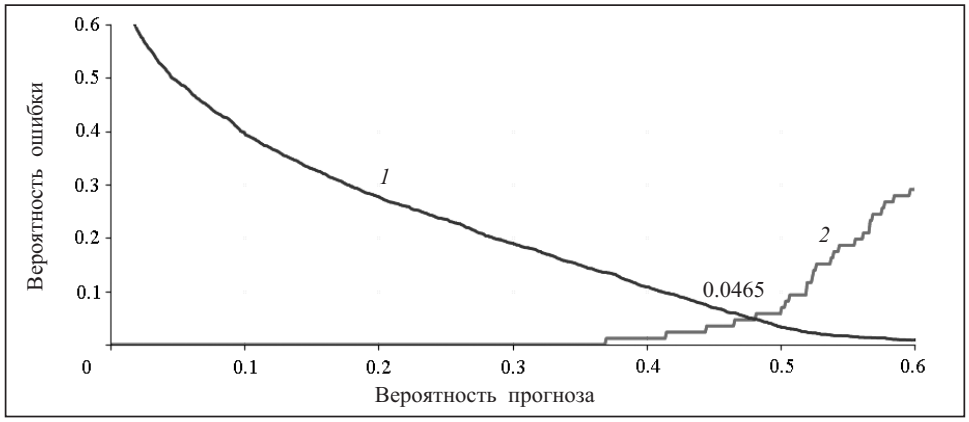

Рис. 4. Графики вероятностей ошибок первого рода (кривая *1*) и второго рода (кривая *2*) распознавания звука [e] в реальных речевых фонограммах

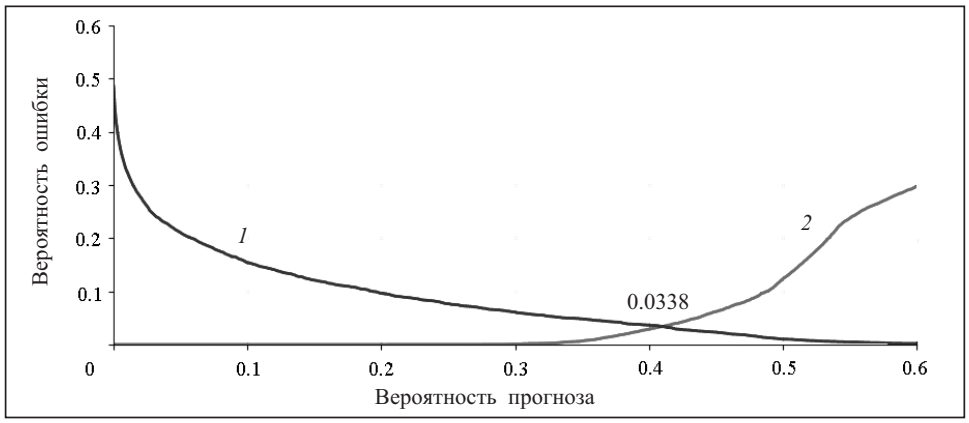

Рис. 5. Графики вероятностей ошибок первого рода (кривая *1*) и второго рода (кривая *2*) разделения звуков и пауз в реальных речевых фонограммах

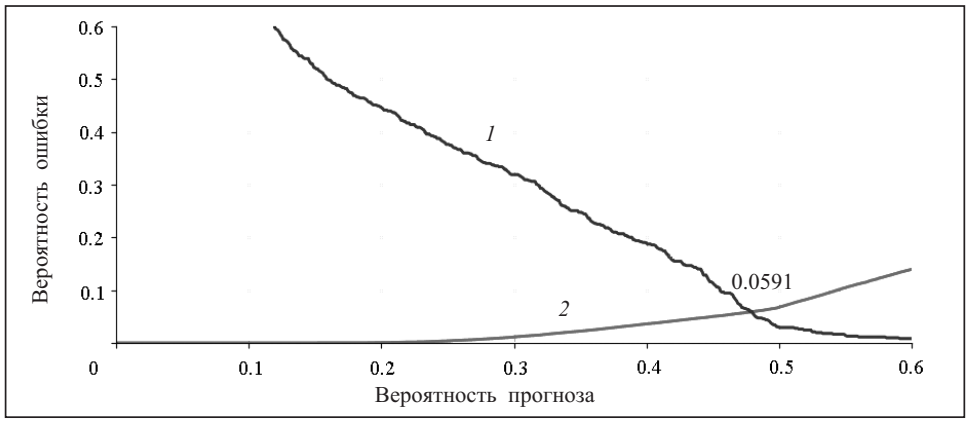

Рис. 6. Графики вероятностей ошибок первого рода (кривая *1*) и второго рода (кривая *2*) выявления пауз с признаками монтажа в реальных речевых фонограммах

В работе предложены методы предварительной обработки входных сигналов, обеспечивающие высокоэффективное обучение нейронных сетей, предназначенных для решения различных задач экспертизы цифровых фонограмм и аппаратуры записи. При этом обеспечивается унификация подхода для построения моделей на основе нейронных сетей глубокого обучения.

Показано, что такие методы эффективны при решении задач автоматической сегментации фонограмм для идентификации диктора и выявления пауз речевой информации, содержащих признаки монтажа.

СПИСОК ЛИТЕРАТУРЫ

1. Solovyov V.I., Rybalskiy O.V., Zhuravel V.V. Verification of fundamental fitness of neuron networks of the deep educating for the construction of the system of exposure of editing of digital phonograms. *Cybernetics and Systems Analysis.* 2020. Vol. 56. N 2. P. 326–330. https://doi.org//10.1007/s10559-020-00249-2.
2. Рибальський О.В., Соловйов В.І., Чернявський С.С., Журавель В.В. Особливості сучасних імовірносних технологій судової експертизи. *Право і правоохорона*. 2019. № 4. С. 212–215. https://doi.org/10.36486np.2019.4.
3. Рыбальский О.В., Соловьев В.И., Журавель В.В. Системы инструментария экспертизы аудио и видеозаписи в Украине. *Вестн. Полоцк. гос. ун-та.* Сер. С. 2018. № 4. С. 15–19.

4. Рыбальский О.В., Жариков Ю.Ф. Современные методы проверки аутентичности магнитных фонограмм в судебно-акустической экспертизе. К.: НАВСУ, 2003. 300 с.
5. Solovyov V.I., Rybalskiy O.V., Zhuravel V.V. Method of exposure of signs of the digital editing in phonograms with the use of neuron networks of the deep learning. *Journal of Automation and Information Sciences.* 2020. Vol. 52, N 1. P. 22–28. https://doi.org/10.1615/JAutomatInfScien.v52.i1.30.
6. Рыбальский О.В., Соловьев В.И., Журавель В.В. Автоматическая сегментация фонограмм по паузам речевого потока. *Сучасна спеціальна техніка.* 2018. № 1. С. 58–64.
7. Семенова Н. В., Колечкина Л. Н., Нагорная А.Н. Векторные задачи оптимизации с линейными критериями на нечетко заданном комбинаторном множестве альтернатив. *Кибернетика и системный анализ.* 2011. № 2. С. 88–99.
8. Малла С. Вэйвлеты в обработке сигналов. Москва: Мир, 2005. 670 с.
9. Сапожков М.А. Электроакустика. Москва: Связь, 1978. 272 с.
10. Александрова Ю.И. Психофизиология. Москва; СПб.: Наука, 2006. 463 с.

Надійшла до редакції 26.05.2020

В.І. Соловйов, О.В. Рибальський, В.В. Журавель, Н.В. Семенова
ДОСЛІДЖЕННЯ МОДЕЛЕЙ РОЗПІЗНАВАННЯ ЗВУКІВ МОВИ НА ОСНОВІ НЕЙРОННИХ МЕРЕЖ ГЛИБОКОГО НАВЧАННЯ ДЛЯ ЕКСПЕРТИЗИ ЦИФРОВИХ ФОНОГРАМ

Анотація. Досліджено моделі на основі нейронних мереж глибокого навчання, що базуються на загальному підході до пауз і сигналів мови як різних видів зафіксованої у фонограмі звукової інформації, які відрізняються деякими характеристиками. Такий підхід дає змогу формувати бази даних навчання з використанням загальних для пауз і сигналів мови методів попереднього оброблення інформації, що забезпечує вищий рівень уніфікації методів навчання мереж, призначених для розв'язання різних задач експертизи.

Ключові слова: апаратура цифрового звукозапису, база даних навчання, нейронна мережа глибокого навчання, цифрове оброблення фонограм, цифрова фонограма, експертиза.

V.I. Solovyov, O.V. Rybalskiy, V.V. Zhuravel, N.V. Semyonova
ANALYZING THE MODELS OF SPEECH RECOGNITION ON THE BASIS OF NEURAL NETWORKS OF DEEP LEARNING FOR EXAMINATION OF DIGITAL PHONOGRAMS

Abstract. The authors analyze the models based on deep learning neural networks, on the basis of the general approach to pauses and speech signals as different types of voice information fixed in a phonogram, different in some characteristics. It is shown that such an approach allows generating the learning databases with the use of the general for pauses and signals of speech methods of preliminary processing of information. This provides a high level of unification of network learning methods intended for solution of various examination problems.

Keywords: digital audio recording devices, learning database, deep learning neural network, digital treatment of phonograms, digital phonogram, examination.

Соловьев Виктор Иванович,
кандидат техн. наук, доцент, заместитель заведующего кафедрой Восточноукраинского национального университета имени Владимира Даля, Северодонецк, e-mail: edemsvi@gmail.com.

Рыбальский Олег Владимирович,
доктор техн. наук, профессор, главный научный сотрудник научно-исследовательской лаборатории, профессор кафедры Национальной академии внутренних дел, Киев, e-mail: rov_1946@ukr.net.

Журавель Вадим Васильевич,
кандидат техн. наук, заведующий лабораторией Киевского научно-исследовательского экспертно-криминалистического центра МВД Украины, e-mail: fonoscopia@ukr.net.

Семенова Наталия Владимировна,
доктор физ.-мат. наук, старший научный сотрудник, ведущий научный сотрудник Института кибернетики имени В.М. Глушкова НАН Украины, Киев, e-mail: nvsemenova@meta.ua.