УДК 621.317+681.849

В.И. Соловьев, О.В. Рыбальский, В.В. Журавель

СИСТЕМА АВТОМАТИЧЕСКОЙ СЕГМЕНТАЦИИ ПАУЗ В ФОНОГРАММАХ НА ОСНОВЕ НЕЙРОННЫХ СЕТЕЙ ГЛУБОКОГО ОБУЧЕНИЯ

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

Введение и постановка задачи

В работах [1, 2] показана возможность высокоэффективного использования нейронных сетей глубокого обучения при построении всего комплекса инструментария для проведения экспертиз материалов и аппаратуры цифровой звукозаписи.

Одна из интересных особенностей применения нейронных сетей глубокого обучения в такой экспертизе — возможность их применения для любого из ее подвидов:

- идентификации диктора по физическим параметрам речевых сигналов, зафиксированных на фонограмме;
- идентификации аппаратуры записи по характеристикам ее собственных шумов, содержащихся в фонограмме;
- выявления следов монтажа в таких фонограммах [3, 4].

Эта особенность вытекает из возможности автоматизации обучения таких сетей и, как следствие, возможности получения в автоматическом режиме огромного количества образцов, необходимых для построения кривых ошибок первого и второго рода, используемых при проведении любого из подвидов такой экспертизы [5]. Отметим, что в экспертизе вероятностью в точке пересечения таких кривых первого и второго рода (далее на графиках обозначается ER) определяется минимальная точность (т.е. минимальная эффективность) инструмента, предназначенного для проведения конкретного вида экспертизы. Как правило, под ошибкой первого рода понимают вероятность того, что правильно принятая в качестве основной, гипотеза H_0 о принадлежности двух распределений (или случайных величин) к одной совокупности, окажется ложно отвергнутой (отклоненной). Тогда под ошибкой второго рода понимают вероятность того, что ложно будет отвергнута правильно принятая альтернативная гипотеза H_1 о принадлежности двух распределений к разным совокупностям [5].

Для криминалистической идентификации это, например, означает, что ошибка первого рода является вероятностью того, что принятая в качестве основной гипотеза H_0 о принадлежности идентификационных признаков к одному объекту их происхождения, будет ложно отвергнута (объекты будут неправильно идентифицированы как разные). Соответственно ошибка второго рода — это вероят-

ность того, что правильно принятая в качестве альтернативной гипотеза H_1 о принадлежности идентификационных признаков к разным объектам их происхождения будет ложно отвергнута (разные объекты будут идентифицированы, как один) [5].

В [1, 2] показано, что использование нейронных сетей глубокого обучения для построения инструментария проведения экспертиз материалов и аппаратуры цифровой звукозаписи позволяет решить «проклятую» проблему такой экспертизы — проблему выявления следов монтажа в цифровых фонограммах. Метод, предложенный в работе [2], обеспечивает высокую вероятность выявления таких следов в паузах речевой информации, записанной на фонограмме. Из этого следует, что перед проведением поиска следов монтажа в исследуемой фонограмме необходимо выделить паузы (произвести ее сегментацию). При этом показано, что инструментарий, построенный на основе нейронных сетей глубокого обучения, требует его работы в автоматическом режиме. В настоящее время в экспертной практике используются две основные модели выделения пауз речевой информации в фонограммах, приспособленные, в той или иной степени, к применению в автоматическом режиме, в частности:

- выделение пауз по установленному порогу [3];
- выделение пауз, основанное на использовании фрактальной размерности Хаусдорфа [6].

Основное требование к системе автоматического выявления следов монтажа в фонограммах — ее высокая эффективность. Она прямо связана с эффективностью выделения пауз в условиях постоянного изменения уровня шумов внутри фонограмм. Эта эффективность при использовании нейронных сетей глубокого обучения также может быть определена кривыми ошибок первого и второго рода, возникающих при автоматической классификации сигналов, содержащихся в фонограмме (речь или пауза). Необходимость использования таких сетей обусловлена тем, что сегментация, основанная на фиксированном пороге, установленном для сигналов фонограммы во временной области, требует корректировки в ручном режиме, а сегментация, основанная на использовании размерности Хаусдорфа, требует значительных вычислительных затрат и не обладает достаточной точностью классификации сигналов фонограммы. Авторы полагают, что эти недостатки можно устранить в результате применения для сегментации фонограмм нейронных сетей глубокого обучения.

Основная цель данной работы — показать подход к построению системы автоматической сегментации фонограмм, основанной на нейронных сетях глубокого обучения, и оценить ее эффективность при автоматическом выделении пауз из потока речевой информации.

При этом система должна быть независимой от уровня шумов в каждой конкретной паузе, а также языка, контекста и диктора, чья речь зафиксирована в фонограмме.

Принципы построения системы

В основе рассматриваемой модели системы лежит подход к паузам, как одному из видов звуковой информации, отличающейся по своим характеристикам от речевой информации, зафиксированной в фонограмме. Мы считаем, что такой подход потенциально обеспечивает не только высокую эффективность сегментации пауз, но и высокую эффективность сегментации речевой информации по отдельным произносимым формантам звуков, что необходимо при идентификации диктора.

Таким образом, рассматривается подход, при котором сигналы в паузах и сигналы звуков речи составляют общую модель звуковой информации, построенной на нейронной сети глубокого обучения.

Для обучения такой сети необходимо создать первичную базу данных звуков и пауз (DataSet Sounds). Для этого используется «нарезка» звуков и пауз, выпол-

ненная в звуковом редакторе, обеспечивающим возможность выполнения этой операции ручным способом. В такую базу вошли фрагменты фонограмм с различным контекстом, записанных разными дикторами на русском, украинском, английском и китайском языках.

Для «нарезки» отбирались сигналы в паузах и сигналы речевой информации, выбранные из ограниченного перечня звуков. Этот перечень приведен ниже в транскрипции, принятой в системе API Международной фонетической организации. Из фонограмм, записанных на русском, украинском 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.

Для обозначения фрагментов пауз в базе данных далее использован специальный символ — [pau].

Разумеется, что вышеприведенный перечень не может охватить всех звуков, содержащихся в исследуемых фонограммах. Все звуки, отличающиеся от входящих в представленный перечень, при классификации в реальных фонограммах будут отнесены к какому-либо из классификационных объектов. Однако вероятность правильной классификации в рамках принятой модели будет мала.

«Нарезка» осуществлялась прослушиванием фрагментов фонограмм на основе субъективного восприятия оператором 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 мсек.

На основе полученной первичной базы фрагментов звуков и пауз формировалась исходная база данных для обучения, тестирования и исследования свойств модели. При формировании сигналов исходной базы они подвергались предварительной обработке. Каждый ранее обозначенный фрагмент из первичной базы разбивался на фрагменты длительностью 20 мс, и производилось преобразование сигналов этих фрагментов из временной в частотную область путем вейвлет преобразования на основе вейвлета Морле. Затем в сигнале выделялось 40 наибольших по спектру локальных максимумов [7].

Отобранные сигналы подвергались частотной фильтрации полосовым фильтром, поскольку экспериментально было установлено, что такая фильтрация способствует повышению вероятности правильной сегментации.

Из фрагментов длительностью 20 мс, прошедших предварительную обработку, были сформированы три независимых массива данных:

- обучающий массив (далее на графике обозначается как Train — использовался для обучения сети);
- тестовый массив (далее на графике обозначается как Test — использовался для оценки выбранной модели по ее эффективности);
- верификационный массив (на его основе осуществлялось построение графиков ошибок первого и второго рода исследуемой модели).

Все три массива представляют собой непересекающиеся множества, что обеспечивает их независимость. В частности, независимость обучающего и тестового массивов позволяет найти конфигурацию сети, обеспечивающую наибольшую вероятность правильного распознавания при наименьшем количестве циклов обучения (эпох), т.е. наибольшую эффективность.

Независимость верификационного массива от массивов, используемых для обучения сети, позволяют обеспечить корректность оценки эффективности предложенной системы с использованием реальных фонограмм.

Каждый из массивов представляет собой смесь фрагментов сигналов длительностью 20 мсек, выделенных из фонограмм, содержащих речь разных дикторов, записанных на разных языках.

В первый слой нейронной сети подавались параметры 40 локальных максимумов — 40 амплитуд нормированных спектров и значений их частоты (всего 80 параметров). Применялась библиотека keras (backend tensorflow) и использовалась полносвязная нейронная сеть, где число слоев доходило до 50. При обучении использовалась множественная классификация на основе выбранной базы звуков.

Таким образом, предложенный подход позволил произвести обучение и проверку качества обучения нейронной сети в условиях множественной классификации фрагментов фонограмм. Этот процесс проиллюстрирован рис. 1 (maximum y on test data — 0,5885, Epoch — 96).

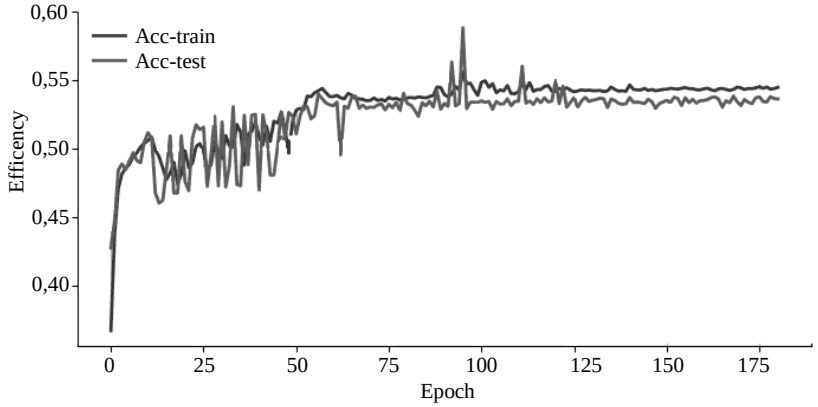

Рис. 1

Как видно на рис. 1, первоначальная эффективность обучения сети при множественной классификации обучающих элементов (фрагментов фонограмм длительностью 20 мсек) достаточно низкая. Но этого и следовало ожидать, поскольку вероятность правильной классификации для каждого отдельного элемента при таких условиях также мала. Можно сказать, что в целом модель не обладает высокой эффективностью распознавания отдельных фрагментов длительностью 20 мсек. Однако при работе модели, полученной при обучении сети, с реальными фонограммами, эффективность выделения пауз можно существенно увеличить (до 90–95 %).

Такое увеличение эффективности классификации сигналов при выделении пауз обусловлено некоторыми, рассмотренными ниже, особенностями речи.

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 [а]). Это использовалось для классификации фрагментов длительностью 20 мсек. Проверка модели на реальных фонограммах осуществлялась путем ее сканирования окном длительностью 20 мсек с шагом сканирования 2 мсек, что примерно соответствует длительности паузы между импульсациями нервной клетки слухового аппарата человека [8].

Запишем, например, русское слово «мама» в транскрипции так, как оно реально воспринимается при сканировании его сдвигающимся окном длительностью 20 мсек в идеальной модели с вероятностью правильной классификации, равной 1:

[m][m][m][m][a][a][a][a][a][a][a][a][a][a][m][m][m][m][a][a][a][a][a][a][a][a].

 ISSN 0572-2691

При использовании модели с реальной классификацией в рамках данной методологии при сканировании наблюдается существенно отличающаяся транскрипция, например:

[m][m][m][m][a][a][o][a][a][a][o][a][a][a][m][m][m][m][a][o][a][a][a][a][o][a][a].

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

Рассмотрим это на примере английского слова «mother». В транскрипции разработанной модели оно может иметь вид:

[m][m][m][m][pau][pau][a][a][o][a][a][a][o][a][a][a][pau][pau][ð][ð][z][ð][ð][ð]

[ð][pau][pau][e][i][e][e][e][e][i][e][e].

Появление [pau] вне фрагментов пауз на стыке звуков имеет физическое объяснение. При сканировании скользящим окном длительностью 20 мсек 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. Эти фиктивные паузы легко удаляются логическим анализом. Пауз столь малой длительности (не более 12 мсек) в речевом сигнале не бывает. Иногда этот спектр весьма близок к другим звукам принятой классификации. Этот эффект можно наблюдать и при прослушивании соответствующего фрагмента речи.

В результате учета некоторых особенностей речи построена система, обеспечивающая в автоматическом режиме эффективную сегментацию пауз в фонограммах, что проиллюстрировано кривыми ошибок первого и второго рода, показанными на рис. 2 и рис. 3. На рис. 2 проиллюстрирована эффективность применения модели для распознавания звуков в реальных речевых фонограммах (звук [a]). Полученные кривые ошибок первого (Error 1) и второго рода (Error 2) построены в координатах: вероятность ошибки — вероятность прогноза классификации. Эти графики построены на основе большого объема статистического материала.

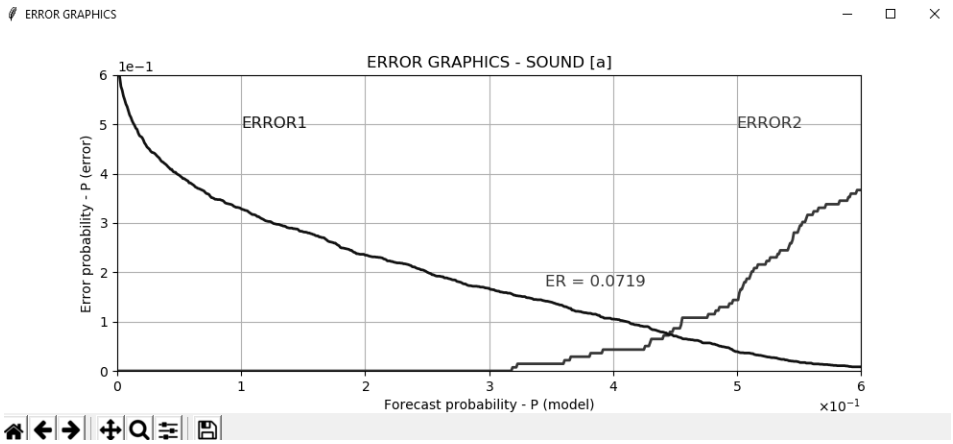

Рис. 2

Как видно из графиков, приведенных на рис. 2, эффективность автоматического выделения из звукового потока звука [a], определяемая вероятностью в точке пересечения кривых ошибок первого и второго рода, весьма высока (ER = 0,0719).

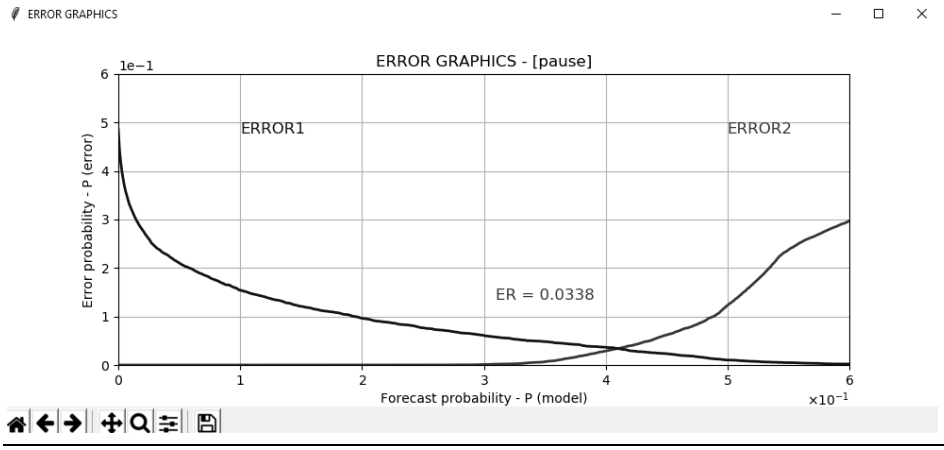

Рис. 3

Графики ошибок первого и второго рода, представленные на рис. 3, иллюстрируют эффективность автоматического выделения пауз из сплошного звукового потока, воспроизводимого с реальных фонограмм. Как видно на рис. 3, эффективность, определяемая величиной вероятности в точке пересечения этих кривых для разработанной системы ER = 0.0338, что является более чем удовлетворительным результатом для системы, работающей в автоматическом режиме.

Заключение

Проведены исследования, позволившие на основе модели нейронной сети глубокого обучения создать эффективную систему автоматического выделения пауз из речевого потока, зафиксированного в фонограмме.

Предлагаемый подход обеспечивает независимость системы от уровня шумов в каждой конкретной паузе, а также языка, контекста и диктора, чья речь зафиксирована в фонограмме.

При построении системы предложено рассматривать паузы в речевом потоке как один из видов звуковой информации, отличающейся по своим характеристикам от речевой информации, зафиксированной в фонограмме.

В.І. Соловйов, О.В. Рибальський, В.В. Журавель

СИСТЕМА АВТОМАТИЧНОЇ СЕГМЕНТАЦІЇ ПАУЗ У ФОНОГРАМАХ НА ОСНОВІ НЕЙРОННИХ МЕРЕЖ ГЛИБОКОГО НАВЧАННЯ

Використання нейронних мереж глибокого навчання для побудови інструментарію для проведення експертиз матеріалів та апаратури цифрового звукозапису дозволяє розв'язати «кляту» проблему такої експертизи — виявлення слідів монтажу в цифрових фонограмах. Ці мережі забезпечують високу ймовірність виявлення таких слідів в паузах мовної інформації, записаної на фонограмі. Перед проведенням пошуку слідів монтажу в досліджуваній фонограмі необхідно виділити паузи (зробити її сегментацію), а інструментарій, побудований на основі нейронних мереж глибокого навчання, вимагає працювати в автоматичному режимі. Основною вимогою автоматичної сегментації є висока ефективність виділення пауз в умовах постійної зміни рівня шумів у фонограмах. Вона визначається ймовірністю у точці перетину кривих помилок I і II роду. На основі нейронних мереж глибокого навчання запропоновано створити автоматизовану систему сегментації фонограм, що має високу ефективність виділення пауз у мовленнєвій інформації. При цьому система має бути незалежною від рівня шумів у кожній конкретній паузі, а також мови, контексту та диктора, чия мова зафіксована у фонограмі. Запропоновано розглядати паузи як один з видів звукової інформації, що відрізняється своїми характеристиками від мовної інформації, зафіксованої у фонограмі. Для навчання такої мережі потрібно створити

первинну базу таких звуків і пауз. На її основі створено три масиви даних, призначених для навчання, тестування та визначення кривих помилок I і II роду. Після навчання та тестування система пройшла перевірку на реальних фонограмах. У результаті врахування деяких особливостей мови на нейронних мережах глибокого навчання побудовано систему, що забезпечує в автоматичному режимі ефективну сегментацію пауз у фонограмах. Отримані результати задовольняють вимогам експертизи, що підтверджено приведеними кривими помилок I і II роду.

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

V.I. Solovyov, O.V. Rybalskiy, V.V. Zhuravel

SYSTEM OF AUTOMATIC SEGMENTATION OF PAUSES IN PHONOGRAMS ON THE BASIS OF NEURON NETWORKS OF THE DEEP LEARNING

The use of neuron networks of the deep learning for the construction of tool for realization of examinations of materials and apparatus of the digital audio recording allows to solve the «frigging» problem of such examination — problem of exposure of tracks of editing in digital phonograms. These networks provide high probability of exposure of such tracks in the pauses of speech information writtenin on a phonogram. Before man-hunting of tracks of editing in the investigated phonogram it is necessary to distinguish pauses (to perform its segmentation), and tool built on the basis of neuron networks of the deep learning, requires its work to be done in automatic mode. The basic requirement of automatic segmentation is high efficiency of selection of pauses in the conditions of permanent change of level of noises in phonograms. It is determined by probability of errors of I and II kinds. It is offered on the basis of neuron networks of the deep learning to create CAS of segmentation of phonograms, possessing high efficiency of selection of pauses in speech information. Thus the system must be independent of level of noises in every concrete pause, and also language, context and announcer, whose speech is fixed in a phonogram. It is suggested to examine pauses as one of the types of voice information, which characteristics differ from characteristics of speech information fixed in a phonogram. For educating of such network it was required to create the primary base of these sounds and pauses. On its basis three arrays of the data, intended for learning, testing and determination of the crooked errors of I and II kinds, are created. After learning and testing the system passed verification on the real phonograms. As a result taking into account some features of speech on the neuron networks of deep learning there has been built the system providing effective segmentation of pauses in phonograms in the automatics mode. The obtained results suit examination that is conformed by given curves over of errors of I and II kinds.

Keywords: apparatus of the digital audio recording, base of learning, neuron network of the deep learning, digital treatment of phonograms, digital phonogram, examination.

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. **56**, N 2. P. 326–330. https://doi.org//10.1007/s10559-020-00249-2.
2. 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. **52**, N 1. P. 22–28. https://doi.org/10.1615/JAutomatInfScien.v52.i1.30.
3. Рыбальский О.В., Жариков Ю.Ф. Современные методы проверки аутентичности магнитных фонограмм в судебно-акустической экспертизе. Киев: НАВСУ, 2003. 300 с.
4. Соловьев В.И. Идентификация аппаратуры аудиозаписи по статистическим характеристикам аудиофайлов. *Реєстрація, зберігання і обробка даних*. 2013. **14**, № 1. С. 59–70.
5. Малла С. Вэйвлеты в обработке сигналов. М. : Мир, 2005. 670 с.
6. Сапожков М.А. Электроакустика. М. : Связь, 1978. 272 с.
7. Александрова Ю.И. Психофизиология. М.-С.П.: Наука, 2006. 463 с.

Получено27.04.2020
После доработки 02.08.2020