УДК 004.822:514

ПРОГРАММНО-АППАРАТНЫЙ КОМПЛЕКС ГОЛОСОВОЙ ИДЕНТИФИКАЦИИ НА ОСНОВЕ НЕЙРОННОЙ СЕТИ КОХОНЕНА

П. А. МЕНЬШАКОВ, И. А. МУРАШКО
Учреждение образования «Гомельский государственный технический университет имени П. О. Сухого», Республика Беларусь

Введение

Большинство из средств контроля доступа имеет высокую цену. Причем большая часть расходов приходится на выделение персонального средства идентификации каждому пользователю. Решением данной проблемы может стать голосовая идентификация.

Биометрия предполагает систему распознавания людей по одной или более физической или поведенческой черт. В области информационных технологий биометрические данные используются в качестве формы управления идентификаторами доступа и контроля доступа. Также биометрический анализ используется для выявления людей, которые находятся под наблюдением [1] .

Задача голосовой идентификации или распознавания диктора по голосу сводится к тому, чтобы выделить, классифицировать и соответствующим образом отреагировать на человеческую речь из входного звукового потока [2]. При этом обычно выделяют три подзадачи: получение голосового отпечатка, идентификация и верификация [3].

Получение голосового отпечатка – процесс получения образца, представляющего вектор характеристик голоса диктора [3].

Идентификация – процесс определения личности по образцу голоса путем сравнения данного образца с шаблонами, сохраненными в базе [4].

Верификация – процесс, при котором с помощью сравнения представленного образца с хранимым в базе шаблоном проверяется запрошенная идентичность [4]. Результатом является подтверждение личности или отрицательный ответ системы.

Выполнение данных процедур занимает довольно длительное время, поэтому затруднена одноврменная идентификации нескольких лиц.

Целью работы является уменьшение времени получения отпечатка голоса, а также уменьшение времени идентификации и верификации голосового отпечатка. Для достижения поставленной цели предложено использовать самоорганизующиеся карты Кохонена (*SOM* – Self-Organized Map) [5], скорость обработки которой была увеличена за счет выделения нейронов с максимальной активностью, при этом получая минимальные потери точности.

Основная часть

Технические средства. Первоначальным этапом голосовой идентификации является получение голоса диктора. Для этого необходим микрофон, фильтр и аналого-цифровой преобразователь для дальнейшей работы с цифровой записью голоса.

В общем виде схема устройства представлена на рис. 1.

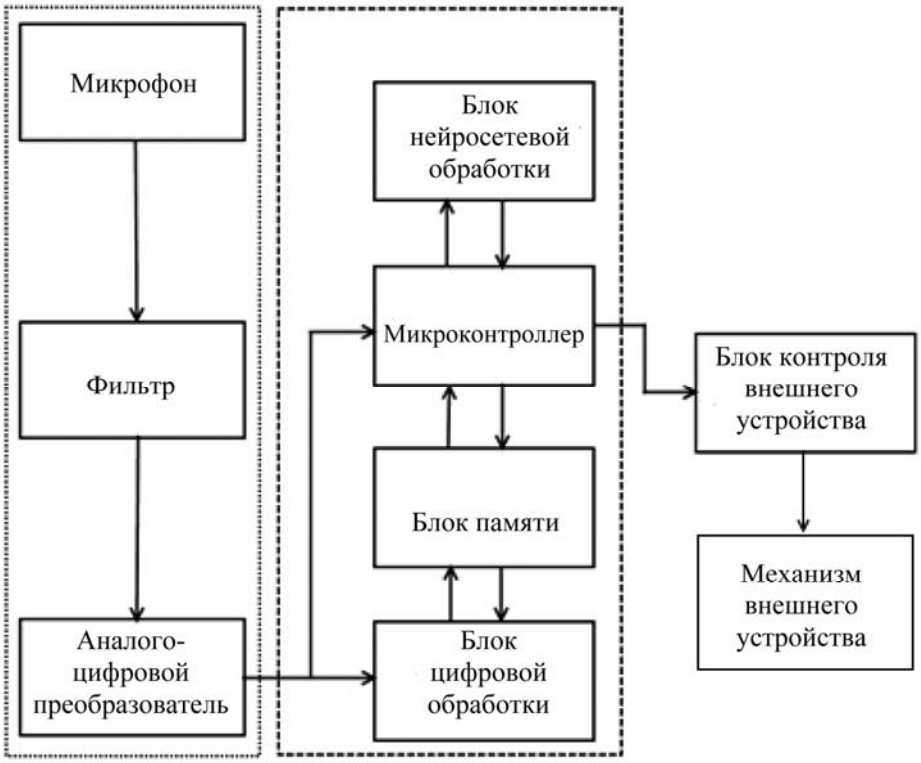

Рис. 1. Схема устройства

С выхода микрофона сигнал подается на вход блока фильтрации. Следующим этапом является прохождение АЦП. Когда АЦП используется для амплитудного анализа, число, получаемое на выходе АЦП, используется для адресации памяти и называется номером канала, а V – шириной канала. Номер канала несет информацию об амплитудном значении сигнала.

Далее оцифрованный сигнал попадает в блок цифровой обработки. В блоке цифровой обработки сигнал фильтруется и преобразуется в вектор, с которым в дальнейшем будет работать микропроцессор и нейросетевой обработчик.

Также полученный вектор заносится в энергонезависимую память. Это необходимо для последующего сравнения с полученным отпечатком.

После сравнения отпечатка в памяти с полученным отпечатком микроконтроллер подает команду на блок управления внешним устройством, к примеру, на магнитный дверной замок.

Сам процесс голосовой идентификации не требователен к ресурсам и состоит из двух этапов. Первым этапом является получение голосового отпечатка диктора и преобразование к виду, в котором его можно будет сравнить с другими. Вторым шагом является сравнение голосовых отпечатков при помощи обученной нейронной сети.

Принцип получения голосового отпечатка. Для реализации процесса преобразования аудиозаписи предлагается произвести определенный порядок действий.

При помощи микрофона получается запись голоса диктора. Наиболее оптимальным является получение WAV-файла ввиду простоты работы с ним [6].

Полученную запись голоса необходимо разделить на кадры. Разделение на кадры представлено на рис. 2. Данное действие необходимо для более простой работы с записанной звуковой дорожкой.

Далее, все вычисления будут производиться с каждым кадром в отдельности.

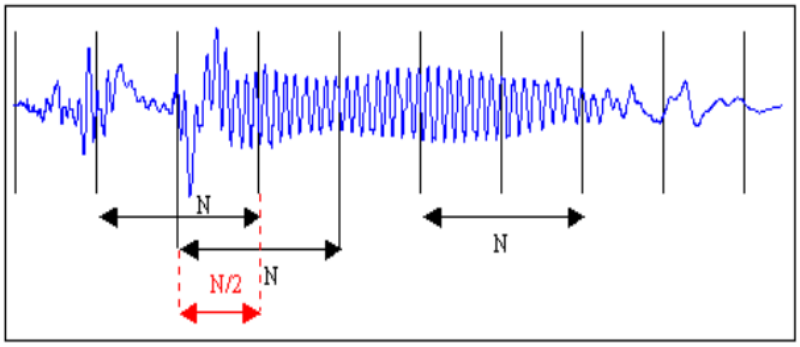

Рис. 2. График звуковой волны

Следующим этапом является устранение нежелательных эффектов и шумов. Это необходимо для того, чтобы записи, полученные в разное время, соответствовали друг другу независимо от сторонних факторов. Существует множество способов, при помощи которых можно уменьшить шумовые эффекты. Нами использовалось умножение каждого кадра на особую весомую функцию «Окно Хемминга» [7]:

$$\omega(n) = 0{,}53836 - 0{,}46164 \cdot \cos\left(\frac{2\pi n}{N-1}\right), \tag{1}$$

где n – порядковый номер элемента в кадре; N – длина кадра (количество значений сигнала, измеренных за период).

Полученные кадры преобразуется в их частотную характеристику при помощи «Быстрого Преобразования Фурье» [8]:

$$X_k = \sum_{i=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn}, \tag{2}$$

где N – длина кадра (количество значений сигнала, измеренных за период); x_n – амплитуда n-го сигнала; X_k – N комплексных амплитуд синусоидальных сигналов, слагающих исходный сигнал.

Важным аспектом оптимизации обработки является сегментация речи на полезные элементы и ее фильтрация. На образцах, записанных в реальных условиях, типовыми являются следующие случаи (рис. 3): наложение различных акустических помех на речь дикторов; наличие на фонограмме речи нескольких дикторов; наложение речи нескольких дикторов друг на друга.

Для решения перечисленных случаев сегментации используются созданные ЦРТ (Центр Речевых Технологий) технологии:

– выделения в фонограмме речи диктора на фоне акустических помех, где для подавления помехи и выделения речи используется образец соответствующей помехи, взятый из Интернета, компакт-диска и т. д.;

– разделения речи дикторов в голосовом коктейле по частоте основного тона;

– разметки выделенных участков речевого сигнала по принадлежности различным дикторам (определение кто и когда говорит), так называемая диаризация речи дикторов [9].

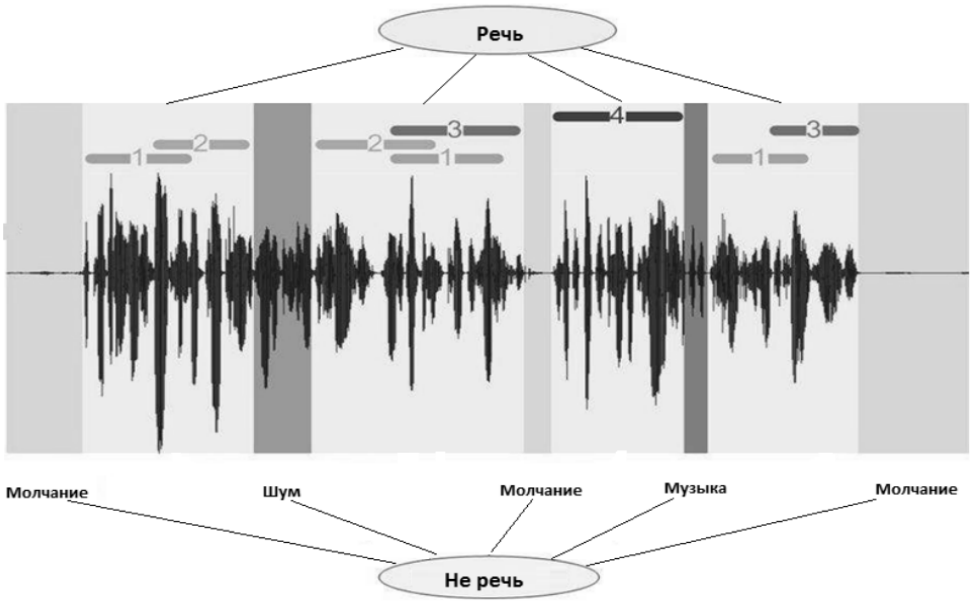

Рис. 3. Схема предварительной обработки речевого сигнала (сегментация дикторов):
1–4 – номера дикторов, речь которых содержится в фонограмме

На сегодняшний день наиболее успешными являются системы распознавания голоса, использующие знания об устройстве слухового аппарата [10]. Ввиду данных особенностей необходимо привести частотную характеристику каждого кадра к «мелам» [11], [12].

Для перехода к «мел»-характеристике используется следующая зависимость:

$$m = 1127 \log_e \left(1 + \frac{f}{700} \right), \tag{3}$$

где m – частота в «мелах»; f – частота в герцах.

Это последнее действие, необходимое для последующего преобразование в вектор характеристики, который впоследствии сравнивается с базой голосовых записей. Вектор будет состоять из «мел»-кепстральных коэффициентов, получить которые можно по следующей формуле:

$$c_n = \sum_{k=1}^{K} (\log S_k) \left[n(k - \frac{1}{2}) \frac{\pi}{K} \right], \tag{4}$$

где c_n – «мел»-кепстральный коэффициент под номером n; S_k – амплитуда k-го значения в кадре в «мелах»; K – наперед заданное количество «мел»-кепстральных коэффициентов $n \in [1, K]$.

Полученный вектор характеристик добавляется в базу данных для последующего сравнения с ним.

Однако более оптимальным вариантом является использование нескольких записей одного и того же голоса. Заранее определенное количество образцов голоса можно использовать для обучения нейронной сети.

Нейросетевое сравнение. В работе использовалось обучение без учителя, так как оно является намного более правдоподобной моделью обучения в биологической системе. Развитая Кохоненом и многими другими, она не нуждается в целевом век-

торе для выходов и, следовательно, не требует сравнения с предопределенными идеальными ответами, а обучающее множество состоит лишь из входных векторов. Обучающий алгоритм подстраивает веса сети так, чтобы получались согласованные выходные векторы, т. е. чтобы предъявление достаточно близких входных векторов давало одинаковые выходы. Процесс обучения, следовательно, выделяет статистические свойства обучающего множества и группирует сходные векторы в классы. Предъявление на вход вектора из данного класса даст определенный выходной вектор [3], [13], [14].

Распространение сигнала в такой сети происходит следующим образом: входной вектор нормируется на 1.0 и подается на вход, который распределяет его дальше через матрицу весов W. Каждый нейрон в слое Кохонена вычисляет сумму на своем входе и в зависимости от состояния окружающих нейронов этого слоя становится активным или неактивным (1.0 и 0.0). Нейроны этого слоя функционируют по принципу конкуренции, т. е. в результате определенного количества итераций активным остается один нейрон или небольшая группа. Этот механизм называется латеральным. Так как отработка этого механизма требует значительных вычислительных ресурсов, в нашей модели он заменен нахождением нейрона с максимальной активностью и присвоением ему активности 1.0, а всем остальным нейронам 0.0. Таким образом, срабатывает нейрон, для которого вектор входа ближе всего к вектору весов связей.

Если сеть находится в режиме обучения, то для выигравшего нейрона происходит коррекция весов матрицы связи по формуле

$$w_{\text{н}} = w_{\text{с}} + \alpha(x - w_{\text{н}}), \tag{5}$$

где $w_{\text{н}}$ – новое значение веса; $w_{\text{с}}$ – старое значение; α – скорость обучения; x – величина входа.

Геометрически это правило иллюстрирует рис. 4.

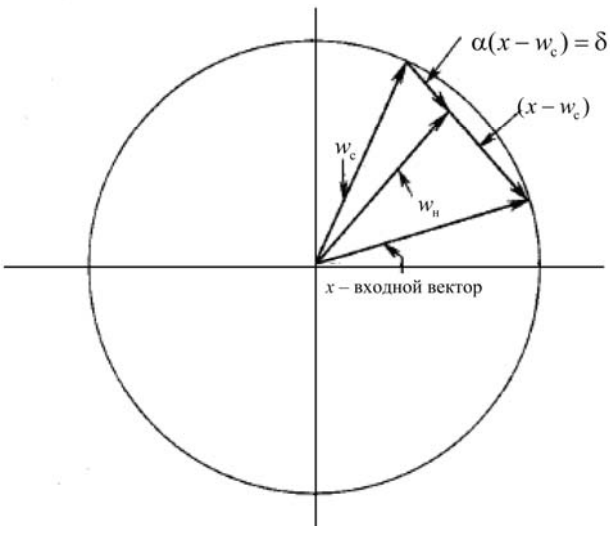

Рис. 4. Коррекция весов нейрона Кохонена

Так как входной вектор x нормирован, т. е. расположен на гиперсфере единичного радиуса в пространстве весов, то при коррекции весов по этому правилу происходит поворот вектора весов в сторону входного сигнала, что позволяет произвести статистическое усреднение входных векторов, на которые реагирует данный нейрон.

Сравнение с существующей аналогичной методикой. Для сравнения алгоритма использовались методики, описанные в *«Analysis of Kohonen's Neural Network with application to speech recognition*» [15] на базе самоорганизующихся карт Кохонена. Для проведения эксперимента были записаны слова, состоящие из цифр. Набор данных включает в себя речевые данные 14 дикторов. Для каждого диктора было записано 30 слов (10 различных слов, по 3 образца на каждое). В табл. 1 показаны выбранные слова.

Таблица 1

Слова, взятые для эксперимента

Один	Два	Три	Четыре	Пять
Шесть	Семь	Восемь	Девять	Ноль

Так же как и в описанном в статье исследовании, все слова были записаны в закрытом помещении, и в качестве источника шума использовался кондиционер. Привлеченные дикторы (11 мужчин и 3 женщины) говорили свободно, сохраняя свои соответствующие акценты и дефекты произношения. Это было необходимо для усложнения задач классификации, поскольку даже те же высказывания имели разную длительность после обнаружения конечной точки.

Сети *SOM* и *TS-SOM*, используемые при моделировании, имели 10 входов и 256 нейронов, расположенных в 16 × 16 массиве.

Результат выполнения операций приведен в табл. 2. Алгоритм с использованием лидирующих нейронов приведен в конце таблицы.

Таблица 2

Результаты классификации различными сетями

Алгоритмы	**Классификация, %**			**Время выполнения, мс**	
	Среднее	**Мин**	**Макс**	**Обучение**	**Выполнение**
SOM original	87,7	82,9	92,9	2,911,873,47	8,14
SOM: *SWS+PDS+Rect*	85,8	79,4	92,2	2,275,220,35	4,48
SOM: *PDS*	89,6	85,1	92,9	2,860,639,27	3,97
SOM: *PDS+Rect*	88,2	84,4	92,9	2,275,061,98	4,07
SOM: *PDS+Trunc.Gauss*	87,9	83,7	93,6	2,661,322,49	5,80
TS-SOM	82,5	75,9	88,6	3,310,01	7,71
SOM:*Leading neurons*	85,5	81,8	89,1	2,113,022,83	1,44

Заключение

Итогом данного исследования стало модульное приложение, осуществляющее голосовую идентификацию пользователя, использующее модернизированный алгоритм вычисления нейронов в слое Кохонена. Программа состоит из трех основных частей. Первая выполняет добавление пользователей, вторая выполняет идентификацию и третья – хранение голосовых записей.

Как показало исследование, полученный алгоритм позволяет значительно ускорить работу программы голосовой идентификации. Данная модернизация позволяет использовать программу на предприятиях с большим потоком пользователей.

Также программный комплекс очень гибок и имеет большое пространство для дальнейшего усовершенствования и добавления новых функций, что делает его не только выгодным программным продуктом, но и перспективным проектом для развития и получения прибыли.

Литература

1. Гарафутдинова, Ф. М. Истоки дактилоскопии / Ф. М. Гарафутдинова // Публичное и частное право. – 2014. – № II (XXII). – С. 173.
2. Ing-Jr, Ding. Developments of Machine Learning Schemes for Dynamic Time-Wrapping-Based Speech Recognition / Ding Ing-Jr, Yen Chih-Ta, Hsu Yen-Ming // Mathematical Problems in Engineering, 2013.
3. Bosi, M. Introduction to digital audio coding and standards / M. Bosi, R. E. Goldberg. – Springer Science+Business, Media USA. – 2010. – 434 p.
4. You, Y. AudioCoding: Theory and Applications / Y. You. – NY : Springer, 2010. – 349 p.
5. Manalili, S. Design of a structured 3D SOM as a music archive / S. Manalili // Springer Verlag Lecture Notes Series : Proceedings of the 8th international conference on advances in selforganizing maps. – P. 188–197.
6. Keyword Extraction for Very High Dimensional Datasets using Random Projection as Key Input Representation Scheme. Master's thesis, De La Salle University. – Manila.
7. Harris, F. J. On the use of windows for harmonic analysis with the discrete Fourier transform / F. J. Harris // Proceedings of the IEEE. – Jan. 1978. – Vol. 66. – P. 51–83.
8. Сергиенко, А. Б. Цифровая обработка сигналов / А. Б. Сергиенко. – 2-е изд. – СПб. : Питер, 2006. – 751 с.
9. Using self-organizing maps to cluster music files based on lyrics and audio features / Research Congress 2013 De La Salle University. – Manila, March 7–9, 2013.
10. Азаров, И. С. Применение мгновенного гармонического анализа для антропоморфической обработки речевых сигналов / И. С. Азаров, А. А. Петровский // Доклады БГУИР. – 2011. – № 4. – С. 59–70.
11. Ghitza, O. Auditory Models and Human Performance in Tasks Related to Speech Coding and Speech Recognition / O. Ghitza // IEEE Transactions on Speech and Audio Processing. – 1994. – Vol. 2, № 1. – P. 115–132.
12. Ivanov, A. V. Analysis of the IHC Adaptation for the Anthropomorphic Speech Processing Systems / A. V. Ivanov, A. A. Petrovsky // EURASIP Journal on Applied Signal Processing. – 2005. – № 9. – P. 1323–1333.
13. Kohonen, T. Self-Organizing Maps / T. Kohonen. – Berlin : Springer, 2001.
14. Mwasiagi, I. Self Organizing Maps – Applications and Novel Algorithm Design / Josphat Igadwa Mwasiagi. – InTech. – Jan. 21, 2011. – P. 91.
15. Carlos Alejandro de Luna-Ortega. Analysis of Kohonen's Neural Network with application to speech recognition / Carlos Alejandro de Luna-Ortega // Mexican International Conference on Artificial Intelligence (MICAI 2009), Nov. 9 to 13, 2009, Guanajuato, Mexico.

Получено 16.11.2016 г.