УДК 004.822:514

МЕТОДИКА ГОЛОСОВОЙ ИДЕНТИФИКАЦИИ НА ОСНОВЕ НЕЙРОННЫХ СЕТЕЙ

П.А. МЕНЬШАКОВ, И.А. МУРАШКО

Гомельский государственный технический университет имени П.О. Сухого, Республика Беларусь

Поступила в редакцию 9 марта 2017

Аннотация. Рассмотрена проблема голосовой идентификации для применения в системах контроля доступа. Предложена методика быстрого получения отпечатка голоса диктора без потери данных, характеризующих голос. Предложено использовать самоорганизующиеся карты Кохонена для идентификации диктора, отличающиеся выделением нейронов с максимальной активностью, что позволило уменьшить время распознавания на 30–80 % по сравнению с существующими решениями.

Ключевые слова: голосовая идентификация, биометрия, система контроля доступа, нейронная сеть, быстрое преобразование Фурье.

Abstract. The problem of voice recognition for use in access control systems was considered. The technique of quick print announcer voice without loss of data characterizing the vote was offered. It proposed to use a Kohonen self-organizing map to identify the speaker, characterized by neuronal release of maximum activity, which reduced the recognition time by 30–80 % compared with existing solutions.

Keywords: voice recognition, biometrics, access control systems, neural network, fast Fourier transform.

Doklady BGUIR. 2017, Vol. 106, No. 4, pp. 12–18
Technique of voice recognition based on neural networks
P.A. Menshakou, I.A. Murashko

Введение

Проблема контроля доступа в настоящее время является достаточно актуальной. Большинство предлагаемых на рынке средств контроля доступа имеют высокую цену и предполагают использование некоторых технических средств персональной идентификации: удостоверение сотрудника, карты (магнитные, бесконтактные), брелоки, электронные ключи Touch Memory, специальные метки и т. п. [1]. При этом значительная часть стоимости эксплуатации системы приходится на создание персонального средства идентификации каждому пользователю. Решением данной проблемы может стать биометрия, то есть использование неотъемлемых характеристик человека (отпечаток пальца, мимика лица, голос, жесты и т. п.) для его идентификации [2]. Применение биометрии позволяет отказаться от технических средств идентификации, которые могут быть утеряны, украдены или переданы на хранение посторонним лицам.

В общем случае биометрия основана на измерении уникальных характеристик конкретного человека. В качестве таких характеристик могут быть использованы биологические признаки, такие как отпечатки пальцев, радужная оболочка глаза, ДНК и т. п., или поведенческие характеристики, полученные в процессе учебы или работы. К последним относятся голос, походка, жесты, почерк и т. п. До недавнего времени поведенческие характеристики мало применялись в системах идентификации в связи с очевидными недостатками. Со временем походка человека может меняться. Голос может измениться вследствие болезни, возрастных изменений или воздействия окружающей среды (например,

высокий уровень шума). Однако в настоящее время с появлением эффективных методов цифровой обработки сигналов интерес к данной тематике в мире значительно возрос [3].

В работе предлагается использовать голосовую идентификацию для устройств контроля доступа в служебные помещения. Преимуществом данного решения является простота аппаратно-программной реализации. Для получения голосового отпечатка требуется только микрофон и аналого-цифровой преобразователь. Данными устройствами оснащены практически все современные настольные и мобильные компьютеры.

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 [4]. При этом обычно выделяют три подзадачи: получение голосового отпечатка, идентификация и верификация [5].

Выполнение данных процедур занимает довольно длительное время, поэтому затруднена одновременная идентификация 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) [6], скорость обработки которых была увеличена за счет выделения нейронов с максимальной активностью, получая при этом минимальные потери точности.

Архитектура программно-аппаратных средств голосовой идентификации

Первоначальным этапом голосовой идентификации является получение речевых признаков диктора (рис. 1). Для этого необходимы микрофон, фильтр и аналого-цифровой преобразователь (АЦП) для дальнейшей работы с цифровой записью голоса.

С выхода микрофона сигнал подается на вход блока фильтрации. Следующим этапом является прохождение АЦП [7]. 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. Для последующего сравнения с сохраненным ранее голосовым отпечатком диктора полученный вектор заносится в энергонезависимую память. 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. На первом этапе осуществляется получение голосового отпечатка диктора и преобразование к виду, в котором его можно будет сравнить с другими. Второй этап заключается в сравнении голосовых отпечатков при помощи обученной нейронной сети [8].

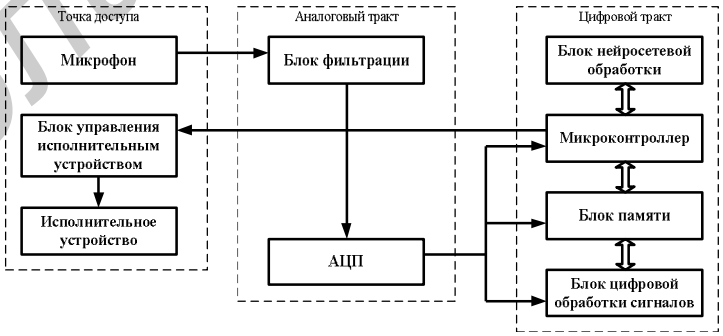


Рис. 1. Архитектура программно-аппаратных средств системы контроля доступа в помещение

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 [9].

Полученную запись голоса следует разделить на кадры. Результат такого разделения представлен на рис. 2. Данное действие необходимо для более простой работы с записанной звуковой дорожкой. Далее все вычисления будут производиться с каждым кадром в отдельности.

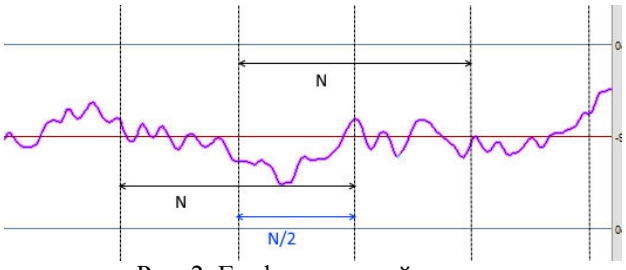


Рис. 2. График звуковой волны

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 [7]. В работе для этой цели использовалось окно Хемминга: $\omega(n) = 0,53836 - 0,36164\cos\left(\frac{2\pi n}{N-1}\right)$, где n – порядковый номер элемента в кадре, N – длина кадра (количество значений сигнала, измеренных за период). Полученные кадры преобразуются в их частотную характеристику при помощи быстрого преобразования Фурье $w_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N}kn}$, 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: наложение различных акустических помех на речь дикторов; наличие на фонограмме речи нескольких дикторов; наложение речи нескольких дикторов друг на друга [4]. Для решения перечисленных задач сегментации используются технологии, созданные в Центре речевых технологий (ЦРТ):

– выделение 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.;

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

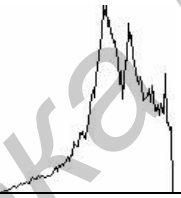
– разметка выделенных участков речевого сигнала по принадлежности различным дикторам (определение, кто и когда говорит), или диаризация речи дикторов.

На сегодняшний день наиболее успешными являются системы распознавания голоса, использующие знания об устройстве слухового аппарата. Ввиду данных особенностей, необходимо привести частотную характеристику каждого кадра к «мелам». Для перехода к «мел»-характеристике используется следующая зависимость: $m = 2595\log_{10}(1+\frac{f}{700}) = 1127\log_e(1+\frac{f}{700})$, где m – частота в мелах, f – частота в герцах.

Это последнее действие, необходимое для последующего преобразования в вектор характеристики, который, впоследствии, сравнивается с базой голосовых записей. Вектор будет состоять из мел-кепстральных коэффициентов, получить которые можно по следующей формуле: $\tilde{c_n} = \sum_{j=1}(\log S_k)\left[n(k-\frac{1}{2})\frac{\pi}{k}\right]$, где c_n – мел-кепстральный коэффициент под номером n, S_k – амплитуда k-го значения в кадре в мелах. Полученный вектор характеристик добавляется в базу данных для последующего сравнения с ним. В ходе проведенных экспериментов было выявлено, что более оптимальным вариантом является использование нескольких записей одного и того же голоса. Заранее определенное количество образцов голоса можно использовать для обучения нейронной сети.

Результатом этапа получения голосового отпечатка является сохраненный, для определенного диктора, набор мел-кепстральных коэффициентов. В предложенном алгоритме получения отпечатка использовались наиболее быстрые методы обработки голосового сигнала. Для определенных задач данный набор действий может не подойти ввиду нехватки точности вычислений. Однако для данной задачи информация о голосе является достаточной. В таблице представлены графические отображения мел-кепстральных векторов, полученные от разных дикторов для слова «один».

Таблица. **Результаты получения вектора характеристик**

Голос	Спектральная характеристика
Диктор №1	
Диктор №2	
Диктор №3	

Анализ представленных в таблице результатов показывает, что для одинакового слова, используемого при получении голосового отпечатка, результирующие векторы характеристик различны и могут быть использованы для идентификации дикторов.

Нейросетевое сравнение

Для реализации процесса сравнения голосовых отпечатков предлагается использовать нейросетевое сравнение при помощи самоорганизующейся сети Кохонена, так как данная нейронная сеть обучается без учителя, с применением модификации, позволяющей ускорить ее работу с минимальной потерей точности.

Обучающее множество, для используемой нейронной сети, состоит лишь из входных векторов, а 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 [10].

Методика работы нейронной сети в стандартном варианте состоит из трех действий.

1. Обучение:

– нумерация входного вектора;

– подача вектора на вход сети;

– вычисление сумм в узлах слоя Кохонена;

– активация узла с суммой 1, если такового нет – повтор пункта «с»;

– коррекция веса узла по формуле: $w_n = w_o + \alpha(x - w_o)$, где w_n – новое значение веса; w_o – старое значение; α – скорость обучения; x – величина входа.

2. Идентификация:

– нумерация входного вектора;

– подача вектора на вход сети;

– вычисление сумм в узлах слоя кохонена;

– активация узла с суммой 1, если такового нет – повтор пункта «c»;

– получение классифицируемого вектора.

3. Верификация:

– нахождение среднеквадратического отклонения между точками введенного вектора и имеющегося в памяти (классифицируемого);

– сверка полученных значений отклонения с пороговым значением.

Ввиду большого количества итераций, вычисления занимают длительное время.

В методике «SOM: Leading Neurons» предлагается сократить количество итераций путем выбора лидирующего нейрона. Таким образом, срабатывает нейрон, для которого вектор входа ближе всего к вектору весов связей $\Delta w(x) = \sum_{j=1}^{n} w_j x^j - w_o - \alpha(x^j - w_o)$, где $\Delta w(x)$ – расстояние между векторами входа и вектором весов связей; w_j – новое значение веса; $x = (x^1, \ldots, x^n)$ – n-мерный вектор входа; w_o – начальный порог активации; α – скорость обучения. В качестве активационной функции используется сигмоид, который имеет следующий вид: $f(x) = \frac{1}{1 + e^{-\alpha \cdot x}}$, где α – параметр наклона.

Результирующая методика:

1. Обучение:

– нумерация входного вектора;

– подача вектора на вход сети;

– вычисление сумм в узлах слоя кохонена;

– нахождение узла с максимальной суммой;

– активация найденного узла;

– коррекция веса узла по формуле: $w_n = w_o + \alpha(x - w_o)$, где w_n – новое значение веса; w_o – старое значение; α – скорость обучения; x – величина входа.

2. Идентификация:

– нумерация входного вектора;

– подача вектора на вход сети;

– вычисление сумм в узлах слоя кохонена;

– нахождение узла с максимальной суммой;

– активация найденного узла;

– получение классифицируемого вектора.

3. Верификация:

– нахождение среднеквадратического отклонения между точками введенного вектора и имеющегося в памяти (классифицируемого);

– сверка полученных значений отклонения с пороговым значением.

Таким образом, в ходе исследования латеральный подход был заменен на активацию лидирующих нейронов.

Результаты и их обсуждение

Эффективность предлагаемого подхода к идентификации сотрудника по голосу, реализованного в системе контроля доступа в помещение, была оценена на основании методик, предложенных в [11]. 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).

Так же как и в описанном в статье исследовании, все слова были записаны в закрытом помещении, 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. Для моделирования были использованы SOM и TS-SOM, имеющие следующую конфигурацию: 10 входов и 256 нейронов, расположенных в 16×16 массиве.

Результат выполнения операций приведен на рис. 3. Значения в графике указаны относительно SOM original.

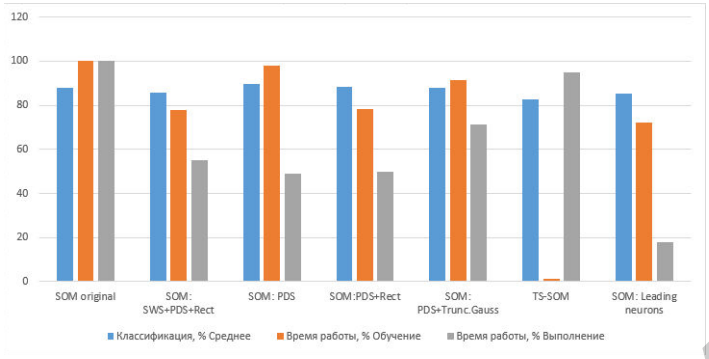


Рис. 3. Результаты классификации различными сетями

На рис. 3 описаны следующие методики [12]:

SOM (Self-organized map) – самоорганизующаяся карта Кохонена без использования модификаций.

SOM SWS (Shortcut Winner Search) + PDS (Partial Distance Search) + Rect (Rectangular SOM) – методика содержащая три модификации:

1) SWS (Shortcut Winner Search) – модификация была предложена Кохоненом (1997). Этот метод требует, частично упорядоченную карту. Таким образом, увеличивается вероятность того, что выигрышный нейрон для входного вектора x будет в окрестности последнего выигравшего нейрона;

2) PDS (Partial Distance Search) – модификация вычисления расстояний. Во время вычисления суммы накопленного расстояния происходит проверка – если квадрат частичного расстояния превышает наименьшее расстояние до ближайшего соседа, вычисление останавливается;

3) Rect (Rectangular SOM) – двумерная SOM, в которой центроиды представлены узлами, которые организованы в прямоугольной решетке. Каждому центроиду присваивается пара координат (*I*, *J*).

SOM PDS – использует только модификацию PDS.

SOM PDS + Rect – методика содержащая две модификации PDS и Rect.

SOM PDS + Trunc. Gauss – содержит две модификации: PDS и Gauss (Gaussian Neighborhood Function) – модификация, при которой в качестве функцией окрестности использована функция Гаусса.

TS-SOM (Tree structured SOM) – самоорганизующаяся карта, представленная в виде древовидной структуры, была предложена Koikkalainen и Oja (1990) [12] в качестве быстрой альтернативы процедур подготовки / тестирования SOM. Дерево поиска уменьшает сложность поиска из O (M) до O (logM). TS-SOM состоит из нескольких слоев SOM сетей с различными разрешениями, при котором нейроны соединены в боковом направлении.

SOM Leading Neurons – предложенная методика работы нейронной сети.

Анализ рис. 3 показывает, что полученная модификация позволила повысить эффективность выполнения в два раза по сравнению с SOM: PDS, при падении точности на 3 %.

Заключение

В статье предложена методика текстонезависимой идентификации человека по голосу, предназначенная для применения в системах контроля доступа. Основу методики составляет выделение речи из фонограмм и последующее формирование биометрических признаков. В качестве идентификационных речевых признаков использовались векторы мел-кепстральных коэффициентов. В качестве метода идентификации предложено использовать самоорганизующиеся карты Кохонена. Отличительной особенностью методики является выбор лидирующего нейрона, что в результате позволяет сократить количество

итераций и, как следствие, уменьшить время обучения и время выполнения по сравнению с известными решениями.

Эффективность предложенной методики оценивалась путем сравнения с известными решениями на основании эксперимента, включающего 30 слов (10 слов в трех вариантах), произносимых 14 дикторами. 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). Для распознавания использовалась нейронная сеть Кохонена, имеющая 10 входов и состоящая из 256 нейронов, организованных в массив 16×16. Анализ результатов показал, что при использовании данной методики время обучения снижается на 10–30 % по сравнению с известными решениями (за исключением самоорганизующихся карт с древовидной структурой, которые имеют очень высокую скорость обучения). Время распознавания снижается практически в два раза, при этом падение точности не превышает 3 %.

Список литературы / References

1. Adeyemo Z.K, Oyeyemi O.J., Akanbi I.A. Development of Hybrid Radio Frequency Identification and Biometric Security Attendance System // Int. J. of Applied Science and Technology. 2014. Vol. 4, № 5. P. 190–197.
2. Biometrics, Computer Security Systems and Artificial Intelligence Applications / Ed. K. Saeed, J. Pejas, R. Mosdorf. Springer, 2006. 345 p.
3. You Y. Audio Coding: Theory and Applications. New York: Springer, 2010. 349 p.
4. Herbig T., Gerl F., Minker W. Self-Learning Speaker Identification: A System for Enhanced Speech Recognition. Berlin: Springer, 2011. 172 p.
5. Al-Shayea Q., Al-Ani M. Speaker Identification: A Novel Fusion Samples Approach // Int. J. of Computer Science and Information Security (IJCSIS). 2016. Vol. 14, № 7. P. 423–427.
6. Kohonen T. Self-Organizing Maps. Berlin: Springer, 1997. 425 p.
7. Bosi M., Goldberg R.E. Introduction to digital audio coding and standards. Springer, 2010. 434 p.
8. Menshakou P.A., Murashko I.A. Voice User Identification in Access Control Systems // Proc. Int. Conf. Open Semantic Technologies for Intelligent Systems (OSTIS-2017). Minsk: BSUIR, 2017. P. 175–178.
9. Petrovsky A., Azarov E. Instantaneous harmonic analysis: techniques and applications to speech signal processing // Speech and computer, Lecture notes in computer science. 2014. Vol. 8773. P. 24–33.
10. Alejandro C. Analysis of Kohonen's Neural Network with application to speech recognition // Mexican International Conference on Artificial Intelligence. Mexico: Guanajuato, 2009. P. 8.
11. Data Mining Algorithms In R/Clustering/Self-Organizing Maps (SOM) [Electronic resource]. – Access mode: https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Clustering/Self-Organizing_Maps (SOM). – Date of access: 15.02.2017.
12. Koikkalainen P., Oja E. Self-Organizing hierarchical feature maps // Proc. Int. Joint Conference on Neural Networks. Vol. II. San Diego, 1990. P. 279–284.

Сведения об авторах

Меньшаков П.А., аспирант кафедры информационных технологий Гомельского государственного технического университета им. П.О. Сухого.

Мурашко И.А., д.т.н., доцент, профессор кафедры информационных технологий Гомельского государственного технического университета им. П.О. Сухого.

Адрес для корреспонденции

246746, Республика Беларусь,
г. Гомель, пр-т Октября, д. 48,
Гомельский государственный
технический университет им. П.О. Сухого
тел. +375-44-567-25-28;
e-mail: pmenshakov@gmail.com;
Меньшаков Павел Алексеевич

Information about the autors

Menshakou P.A., postgraduate student of information technologies department of Sukhoi State Technical University of Gomel.

Murashko I.A. D.Sci., associate professor, professor of information technologies department of Sukhoi State Technical University of Gomel.

Address for correspondence

246746, Republic of Belarus,
Gomel, Octiabria ave., 48,
Gomel State Technical University
named after P.O. Sukhoi
tel. + 375-44-567-25-28;
e-mail: pmenshakov@gmail.com;
Menshakou Pavel Alekseevich