УДК 510.676, 519.7

НЕЙРОННЫЕ СЕТИ В ЗАДАЧЕ ИДЕНТИФИКАЦИИ ДИКТОРА ПО ГОЛОСУ

Бучнева Т.И., Кудряшов М.Ю.
Кафедра информационных технологий

Поступила в редакцию 01.06.2015, после переработки 18.06.2015.

В статье представлена система идентификации диктора по голосу, разработанная на основе многослойной нейронной сети. Рассматривается структура нейронной сети для решения задачи идентификации по голосу. Для обучения нейронной сети предлагается последовательно использовать генетический алгоритм и алгоритм обратного распространения ошибки. Разработан алгоритм принятия решения о распознавании на основе выходных данных нейронной сети. Приведены результаты работы системы на реальной речевой базе.

Ключевые слова: биометрическая идентификация, текстонезависимая идентификация диктора, распознавание по голосу, нейронные сети, генетические алгоритмы, алгоритм обратного распространения ошибки.

Вестник ТвГУ. Серия: Прикладная математика. 2015. № 2. С. 119–126.

Введение

Решение проблемы идентификации диктора по голосу занимает значительное место при разработке биометрических систем распознавания личности ввиду удобства и простоты использования, широкого спектра возможностей применения и отсутствия необходимости в дорогостоящем оборудовании. Однако при реализации автоматических систем текстонезависимой идентификации диктора по голосу возникают значительные сложности, связанные с неустойчивостью речевого сигнала [4]. Один из вариантов решения данной проблемы – реализация процесса распознавания на основе искусственных нейронных сетей [4].

Наличие реальных систем распознавания личности по голосу, в основу которых положены нейронные сети [1, 2], свидетельствует о том, что данный подход является перспективным с точки зрения усовершенствования как структуры нейронных сетей, так и алгоритмов обучения. Повышение показателей распознавания является актуальной задачей, поскольку ни одна система в настоящий момент не обеспечивает 100% точности, в силу чего практические применения таких систем немногочисленны.

В статье рассматривается архитектура системы текстонезависимой автоматической идентификации диктора по голосу. Составной частью предложенной архитектуры является многослойная нейронная сеть, разработанная структура которой подробно излагается. Статья содержит описание обучающей процедуры, основанной на последовательном применении генетического алгоритма и алгоритма

обратного распространения ошибки. Разработан алгоритм принятия решения для системы с предложенной архитектурой.

1. Архитектура системы текстонезависимой идентификации диктора

На Рис. 1 представлены основные этапы обработки в рассматриваемой архитектуре системы текстонезависимой идентификации диктора и их взаимосвязи.

На этапе предобработки выполняется преобразование произнесенной диктором фразы в цифровой сигнал, из которого удаляются шум, паузы и невокализованные фрагменты.

Вектора признаков формируются с использованием кепстральных коэффициентов [7].

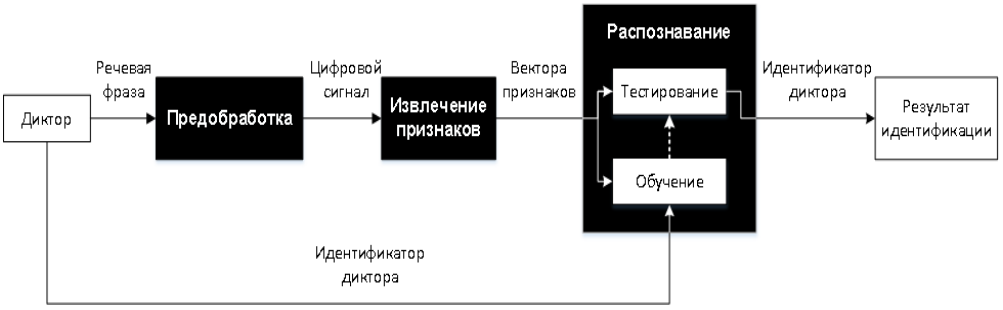

Рис. 1: Архитектура системы текстонезависимой идентификации диктора

Для каждого фрагмента речевого сигнала длительностью 20 мс рассчитываются значения мел-частотных кепстральных коэффициентов [7], формирующих вектор признаков. Каждый вектор состоит из n вещественных коэффициентов, количество которых варьируется от 12 до 24.

Таким образом, для фразы, произнесенной диктором d, участвующим в распознавании, рассчитывается набор векторов $V_d = \{v_1, \ldots, v_{m_d}\}$ мел-частотных кепстральных коэффициентов. Далее, полученное множество векторов сужается с помощью алгоритма кластеризации k-внутригрупповых средних [6].

Распознавание в системе идентификации диктора по голосу основано на использовании многослойной нейронной сети, для работы с которой требуется две основные фазы — обучение и тестирование [3].

Во время обучения выполняется формирование обучающих наборов, соответствующих каждому диктору, участвующему в распознавании. Для этого записывается речь каждого зарегистрированного пользователя, и по данной записи вычисляются вектора признаков с помощью вышеупомянутой процедуры. Обучающий набор состоит из вектора признаков и соответствующему ему эталонного выхода нейронной сети. Полученные наборы используются для обучения нейронной сети.

В фазе тестирования вектора признаков, вычисленные по речевому сигналу неизвестного диктора, подаются на вход обученной сети. Решение о распознавании данного диктора принимается на основе полученного выходного сигнала нейронной сети.

2. Структура многослойной нейронной сети

Для решения задачи распознавания диктора по голосу использовалась многослойная нейронная сеть с одним скрытым слоем. Количество нейронов n во входном слое $i_1, i_2, \ldots, i_n$ соответствует размерности вектора признаков. Параметры скрытых слоев (количество и размерность) определялись экспериментально по критерию минимизации ошибок обучения нейронной сети. Наименьшая ошибка обучения получена в случае, когда единственный скрытый слой состоит из 10 нейронов $h_1, h_2, \ldots, h_{10}$. Количество нейронов выходного слоя сети $o_1, o_2, \ldots, o_k$ соответствует размерности k множества дикторов G, зарегистрированных в системе. Элементы множества G есть числа от 1 до k.

Для формирования обучающих наборов необходимо установить соответствие между эталонным выходным вектором сети и одним из классов, на которые нужно распределить входные вектора.

Эталонный выходной сигнал сети в обучающем наборе устроен таким образом, что на одном из выходов должен присутствовать признак принадлежности входного вектора определенному классу (зарегистрированному диктору $d \in G$). И наоборот, на остальных выходах есть признак отсутствия принадлежности соответствующему классу. Признаком принадлежности классу для обучающих векторов является число 1, и напротив, число 0 означает, что образ классу не принадлежит.

Таким образом, для обучающего вектора $\{i_{d_1}, i_{d_2}, \ldots i_{d_n}\}$ диктора d эталонный выход сети должен иметь вид

$$\{o_j\}, j \in G, o_j = \begin{cases} 0, j \neq d, \\ 1, j = d. \end{cases}$$

3. Алгоритм принятия решения о распознавании на основе выходных данных нейронной сети

Решение о распознавании диктора $d \in G$ принимается на основе анализа выходных сигналов $O_t = \{o_{t_1}, o_{t_2}, \ldots, o_{t_k}\}, t = \overline{1, m_d}, k = |G|$ нейронной сети, полученных по входным сигналам $v_1, \ldots, v_{m_d}$ (вектора признаков диктора).

Суть анализа заключается в вычислении значений вектора $S_d = \{s_{d_1}, \ldots, s_{d_k}\}$ по следующей формуле:

$$s_{d_s} = \sum_{t=1}^{m_d} s_t, \text{где } s_t = \begin{cases} \min\limits_{i=1\ldots k}\{o_{t_i} : o_{t_i} \geq e\}, & o_{t_g} \geq e, \exists i : i \neq g, o_{t_i} \geq p, \\ o_{t_g}, & o_{t_g} \geq p, \forall i : i \neq g, o_{t_i} < e, \\ 0, & \text{иначе}, \end{cases}$$

где $g \in [1, k]$, $e \in [0.17, 0.5]$ и $p \in [0.5, 1)$ – заранее заданные параметры.

Расчетная формула для s_{d_g} была получена эвристическим методом на основе вычислительных экспериментов на реальной речевой базе, которая подробно описана в разделе 5.

Индекс максимального значения в векторе S_d, превышающего заранее заданную пороговую величину p_1, соответствует распознанному диктору:

$$g^* = \arg\max_{g \in G}\{s_{d_g} : s_{d_g} \geq p_1\}.$$

Параметр p_1 рассчитывается заранее для каждого набора G зарегистрированных дикторов. Если все значения вектора S_d меньше p_1, то диктор d считается чужим, незарегистрированным пользователем.

4. Обучение нейронной сети

Обучение нейронной сети заключается в настройке ее весов таким образом, чтобы на заданном обучающем множестве входных сигналов на выходе сети формировались требуемые значения.

Для обучения многослойной нейронной сети при решении задачи идентификации диктора по голосу предлагается использовать подход, основанный на последовательном применении генетического алгоритма и алгоритма обратного распространения ошибки.

Использование генетического алгоритма при обучении сети обусловлено тем, что он позволяет избежать попадания в локальные минимумы и тем самым производит глобальный поиск в пространстве весов нейронной сети заданной топологии [3]. Как показали проведенные вычислительные эксперименты, среднеквадратичная ошибка обучения сети генетическим алгоритмом уменьшается с очень медленной скоростью и не доходит до минимально допустимой величины. В этой связи, для достижения значительного уменьшения ошибки обучения необходимо последующее применение стандартной процедуры обучения многослойной нейронной сети алгоритмом обратного распространения ошибки [3]. Полученные результаты показали, что такой подход позволяет достичь значительно лучших показателей качества обучения сети, нежели генетический алгоритм и алгоритм обратного распространения ошибки в отдельности.

В начале обучения веса сети инициализируются значениями случайной величины, имеющей нормальное распределение $N(0, 1)$, что позволяет избежать некорректных случаев обучения, возникающих, например, в результате насыщения сети большими весами, или при получении одинаковых начальных значений.

Далее, последовательно применяются два алгоритма: генетический алгоритм и алгоритм обратного распространения ошибки. Смена алгоритма осуществляется в результате анализа ошибки обучения: как только ошибка обучения перестает уменьшаться в заданной динамике, происходит переход к алгоритму обратного распространения ошибки.

Стандартная процедура обучения многослойной нейронной сети алгоритмом обратного распространения ошибки описана, например, в [3]. В следующем разделе рассмотрена процедура предварительной оптимизации весов нейронной сети эволюционным методом.

4.1 Метод предварительной оптимизации весов нейронной сети с помощью генетического алгоритма

Эволюционный подход к обучению нейронных сетей состоит из выбора представления весов сети и самого процесса эволюции, основанного на генетическом алгоритме [3]. Все веса многослойной нейронной сети представляют собой множество

действительных чисел. Таким образом, каждая особь в популяции содержит информацию обо всех весах нейронной сети и кодируется в виде последовательности действительных чисел $a_1a_2\dots a_qa_{q+1}\dots a_la_{l+1}\dots a_{m-1}a_m$. Реализация эволюционного процесса заключается в применении генетического алгоритма к популяции особей и состоит из следующих шагов.

1. Применение генетических операторов к текущему поколению особей. Основные генетические операторы, такие как двухточечный кроссовер, арифметический кроссовер и оператор мутации, подробно рассматриваются ниже.

2. Восстановление множества весов из каждой особи нового поколения, полученного после применения генетических операторов, и конструирование соответствующей этому множеству нейронной сети с рассмотренной выше архитектурой.

3. Вычисление общей среднеквадратической погрешности между фактическими и заданными значениями на всех выходах сети при подаче на ее входы обучающих образов [3]. Таким образом, функция приспособленности особей вычисляется как
$$\sum_{i=1}^{N}\sum_{j=1}^{k}(o_{ij}-o'_{ij})^2,$$
где N – число обучающих векторов, o_{ij} – j-й выход сети для i-го входного вектора, o'_{ij} – эталонное значение j-го выхода сети для i-го входного вектора, k – количество нейронов в выходном слое сети.

4. Селекция особей согласно их приспособленности. В новое поколение попадают 10 особей с наибольшими значениями функции приспособленности.

При двухточечном кроссовере выбираются две хромосомы (особи) и случайные позиции q и l. Хромосомы разрезаются по этим позициям и обмениваются средними частями. Пусть выбраны исходные хромосомы:

$$A:\ a_1a_2\dots a_qa_{q+1}\dots a_la_{l+1}\dots a_{m-1}a_m \text{ и } B: b_1b_2\dots b_qb_{q+1}\dots b_lb_{l+1}\dots b_{m-1}b_m.$$

Тогда результирующими хромосомами являются

$$A_1: a_1a_2\dots a_qb_{q+1}\dots b_la_{l+1}\dots a_{m-1}a_m \text{ и } B_1: b_1b_2\dots b_qa_{q+1}\dots a_lb_{l+1}\dots b_{m-1}b_m.$$

Арифметический кроссовер работает следующим образом. Пусть C_1, C_2 – предки, тогда потомки вычисляются как $H_{1i}=\alpha C_{1i}+(1-\alpha)C_{2i}, H_{2i}=(1-\alpha)C_{1i}+\alpha C_{2i}$, где α – константа из $[0,1]$. Используется так называемая линейная нормализация, когда хромосомы сортируются по лучшему значению функции приспособленности, а затем допускаются к кроссоверу с вероятностью, пропорциональной позиции в списке.

Оператор мутации заключается в том, что необходимо сначала случайным образом выбрать из текущей популяции особь, выбрать значение случайной величины, равномерно распределенной на интервале $[0,1]$, и если данное число меньше p, нужно выбрать две точки мутации и значения, находящиеся между этими точками, заменить случайными значениями из интервала значений весов. В результате мутации, выбранная особь помещается в новое поколение.

5. Результаты работы системы идентификации диктора по голосу

Для оценки результатов работы системы с предложенной архитектурой были использованы следующие критерии оценки качества работы систем идентификации – вероятности ошибок первого и второго рода.

Ошибка первого рода возникает в случае, если принимается решение о том, что диктор не принадлежит списку зарегистрированных пользователей, в то время как на самом деле он там присутствует. Ошибка второго рода, напротив, возникает при ложном допуске незарегистрированного пользователя. Значения ошибок рассчитывались как отношения числа неудачных тестов к общему числу тестов.

Помимо этих критериев рассчитывается такой показатель качества распознавания как процент точной верификации. Под данным критерием понимается процентное соотношение количества правильно допущенных пользователей с количеством допущенных и правильно распознанных пользователей. Для вычисления статистических критериев было проведено 15 экспериментов. В каждом эксперименте, для обучения сети использовались 10 зарегистрированных в системе пользователей (дикторов). Для каждого диктора был сформирован аудио файл, в который с частотой дискретизации 12 кГц записана произнесенная им фраза, либо на русском, либо на английском языке, состоящая из нескольких (от одного до пяти) произвольных слов. Для тестирования таким же образом было записано 20 аудио файлов, по одному для каждого диктора. Половина тестовых записей принадлежит зарегистрированным пользователям, в то время как оставшиеся 10 представляют записи речи неизвестных дикторов. Записи аудио-файлов зарегистрированных пользователей для тестирования системы были сделаны на несколько месяцев позже, чем записи тех же пользователей для обучения системы. В ходе исследования использовалась речь 10 мужчин и 10 женщин в возрастном диапазоне от 20 до 55 лет. Запись производилась с помощью микрофона Soundking EH031 на компьютере Lenovo в операционной системе Windows 7.

В ходе проведенных экспериментов наилучшие результаты были получены при следующих значениях параметров: $p = 0.99$, $e = 0.18$, $p_1 = 33.7$. В частности, величина ошибки первого рода составила 0.027, значение ошибки второго рода оказалось равной 0.033, а коэффициент точной верификации достиг 93.97%.

Заключение

Проведенное исследование показало, что предложенная система на основе многослойной нейронной сети может быть использована при решении задачи идентификации диктора по голосу как вспомогательная подсистема, гарантирующая высокий процент точной верификации, либо как самостоятельная система, но с учетом вероятностей возникновения ошибок первого и второго рода. Результаты исследования могут быть использованы для дальнейшего изучения и усовершенствования алгоритмов голосовой идентификации.

Список литературы

[1] Kamruzzaman S.M., Karim R., Islam S., Haque E. Speaker identification using MFCC-domain support vector machine // International Journal of Electrical and Power Engineering. 2007. Vol. 1, № 3. Pp. 274–278. doi:10.3923/ijepe.2007.274.278

[2] Yee C.S., Ahmad A.M. Mel frequency cepstral coefficients for speaker recognition using gaussian mixture model-artificial neural network model // Proc. of International Conference on Electronic Design (ICED 2008). 2008. Vol. 1. Pp. 1–5.

[3] Рутковская Д., Пилиньский М., Рутковский Л. Нейронные сети, генетические алгоритмы и нечеткие системы. М.: Горячая линия, 2006. 452 с.

[4] Макаревич О.Б., Федоров В.М., Тумоян Е.П. Применение сетей функций радиального базиса для текстонезависимой идентификации диктора // Нейрокомпьютеры: разработка, применение. 2001. № 7-8. С. 82–86.

[5] Davis S.B., Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences // IEEE Transactions on Acoustics, Speech, and Signal Processing. 1980. Vol. 28. Pp. 357–366.

[6] Лепский А.Е., Броневич А.Г. Математические методы распознавания образов: Курс лекций. Таганрог: ТТИ ЮФУ, 2009. 155 с.

[7] Матвеев Ю.Н. Технологии биометрической идентификации личности по голосу и другим модальностям // Вестник Московского государственного технического университета им. Н.Э. Баумана. Серия: Приборостроение. Специальный выпуск. Биометрические технологии. 2012. № 3(3). С. 46–61.

Библиографическая ссылка

Бучнева Т.И., Кудряшов М.Ю. Нейронные сети в задаче идентификации диктора по голосу // Вестник ТвГУ. Серия: Прикладная математика. 2015. № 2. С. 119–126.

Сведения об авторах

1. **Бучнева Татьяна Игоревна**
аспирант кафедры информационных технологий Тверского госуниверситета.

Россия, 170100, г. Тверь, ул. Желябова, д. 33, ТвГУ, ПМиК.

2. **Кудряшов Максим Юрьевич**
доцент кафедры информационных технологий Тверского госуниверситета.

Россия, 170100, г. Тверь, ул. Желябова, д. 33, ТвГУ, ПМиК.

NEURAL NETWORK IN THE TASK OF SPEAKER IDENTIFICATION BY VOICE

Buchneva Tatyana Igorevna
PhD student at Information Technologies department, Tver State University.
Russia, 170100, Tver, 33 Zhelyabova str., TSU

Kudryashov Maxim Yuryevich
Associate professor at Information Technologies department, Tver State University.
Russia, 170100, Tver, 33 Zhelyabova str., TSU

Received 01.06.2015, revised 18.06.2015.

The article presents the speaker identification system developed on the basis of multilayer neural network The article provides the structure of a neural network for solving the problem of voice identification. It is dealt with the sequential training method, which include back-propagation and genetic algorithm. The method of decision-making is proposed. This method bases on the output data of the neural network. Data are given about the results of identification for real speech base.

Keywords: biometric identification, text-independent speaker identification, speaker recognition, neural network, genetic algorithm, error back propagation algorithm.

Bibliographic citation

Buchneva T.I., Kudryashov M.Yu. Neural network in the task of speaker identification by voice. *Vestnik TvGU. Seriya: Prikladnaya matematika* [Herald of Tver State University. Series: Applied Mathematics], 2015, no. 2, pp. 119–126. (in Russian)