УДК 004.934

Е.Е. Федоров

Метод классификации вокальных звуков речи на основе саундлетной байесовской нейронной сети

Предложен метод классификации вокальных звуков речи, который базируется на авторской саундлетной байесовской нейронной сети и позволяет учитывать структуру квазипериодического сигнала и сопоставлять образцы вокальных звуков речи разной длины. Разработаны методы создания образцов, формирования опорных образцов и модель их классификации.

Ключевые слова: искусственная нейронная сеть, материнский саундлет, дочерний саундлет, саундлетное отображение, классификация вокальных звуков, байесовский подход.

Запропоновано метод класифікації вокальних звуків мовлення, який базується на авторській саундлетній байєсівській нейронній мережі та дозволяє враховувати структуру квазіперіодичного сигналу і зіставляти зразки вокальних звуків мовлення різної довжини. Розроблено методи створення зразків, формування опорних зразків і модель їх класифікації.

Ключові слова: штучна нейронна мережа, материнський саундлет, дочірній саундлет, саундлетне відображення, класифікація вокальних звуків, байєсівський підхід.

Общая постановка проблемы

Сегодня актуальна разработка программных компонент, предназначенных для распознавания, синтеза речи человека и других сигналов, используемых в интеллектуальных компьютерных системах.

В основе данной задачи лежит проблема построения эффективных методов, обеспечивающих высокую скорость обучения модели распознавания, а также высокую вероятность, адекватность и скорость распознавания речевых сигналов.

Анализ исследований

В современных системах распознавания речевых образов используются следующие подходы: логический, метрический, байесовский, нейросетевой, структурный. Существующие методы и модели распознавания речевых образов обычно основаны на скрытых марковских моделях (СММ) [1–3], алгоритме динамического программирования *DTW* [1–2], искусственных нейронных сетях [4–6] и имеют следующие недостатки [7]:

- время обучения несколько месяцев;
- хранение большого количества опорных образцов звуков или слов, а также весовых коэффициентов;
- длительное время распознавания;
- вероятность распознавания меньше 95 процентов;
- наличие сотен тысяч обучающих образцов.

Постановка задачи

Цель работы – разработка метода классификации вокальных звуков речи на основе саундлетной байесовской нейронной сети.

Решение задач и результаты исследований

Для достижения цели необходимо:

- разработать:
  - метод создания образцов вокальных звуков;
  - метод формирования опорных образцов на основе семейства дискретных саундлетов и саундлетных отображений;
  - модель классификации вокальных звуков на основе саундлетной байесовской нейронной сети и опорных образцов;
- создать критерии оценки эффективности модели;
- формализовать условия классификации вокального звука по опорным образцам на основе семейства саундлетов и саундлетных отображений для оценивания результатов классификации;
- разработать логико-формальные правила для оценивания результатов классификации по модели.

Метод создания образцов вокальных звуков. Согласно [8], образцом вокального звука речи назовем участок вокального звука в речевом сигнале, расположенный между соседними пиковыми значениями и имеющий длину, соответствующую квазипериоду.

При формировании образца в режиме обучения экспертом вводится левая и правая границы N^l, N^r вокального звука в сигнале f, а в режиме классификации автоматически определяется (на основе энергий последовательно идущих участков сигнала равной длины) левая и правая границы N^l, N^r вокальной части сигнала f

После задания или вычисления границ N^l, N^r на множестве $\{N^l,\ldots,N^r\}$ сигнала f вычисляется функция автокорреляции, с помощью которой определяется длина периода основного тона N^{FT} вокального звука.

Для формирования образца как структурообразующего элемента вокального звука множество $\{N^l,\ldots,N^r\}$ сигнала f разбивается на участки на основе вычисленной длины периода основного тона N^{FT} согласно следующему правилу

$$N_0^{\max} = \arg\max_n f(n),$$

$$n \in \left\{N^l - 0{,}1\cdot N_0^{FT},\ldots,N^l + 0{,}1\cdot N_0^{FT}\right\},$$

$$N_0^{FT} = N^{FT},$$

$$N_{i-1}^{\max} \le N^r \Rightarrow \left(N_i^{\min} = N_{i-1}^{\max}\right) \wedge \left(N_i^{\max} = \arg\max_n f(n)\right) \wedge$$

$$\wedge \left(N_i^{FT} = N_i^{\max} - N_i^{\min}\right),$$

$$n \in \left\{N_i^{\min} + 0{,}9\cdot N_{i-1}^{FT},\ldots,N_i^{\min} + 1{,}1\cdot N_{i-1}^{FT}\right\}.$$

На основе этого разбиения формируется конечная совокупность образцов, описываемых множеством вещественнозначных ограниченных финитных дискретных функций $\left\{x_i \mid i \in \{1,\ldots,I\}\right\}$ в виде

$$x_i(n) = \begin{cases} f(n), & n \in \left\{N_i^{\min},\ldots,N_i^{\max}\right\} \\ 0, & n \notin \left\{N_i^{\min},\ldots,N_i^{\max}\right\} \end{cases}, \quad i \in \{1,\ldots,I\},$$

$$A_i^{\min} = \min_n f(n),\ n \in \left\{N_i^{\min},\ldots,N_i^{\max}\right\},\quad i \in \{1,\ldots,I\},$$

$$A_i^{\max} = \max_n f(n),\ n \in \left\{N_i^{\min},\ldots,N_i^{\max}\right\},\quad i \in \{1,\ldots,I\}.$$

На основе введенных в [8] семейств непрерывных и дискретных саундлетов и саундлетных отображений сформируем опорные образцы вокальных звуков речи.

Метод формирования опорных образцов. Пусть дана конечная совокупность обучающих образцов вокального звука, которая описывается множеством целочисленных ограниченных финитных дискретных функций $X = \left\{x_i \mid i \in \{1,\ldots,I\}\right\}$, причем $A_i^{\min}, A_i^{\max}$ – минимальное и максимальное значения функции x_i на компакте $\left\{N_i^{\min},\ldots,N_i^{\max}\right\}$.

Для сопоставления элементов множества X между собой для каждой функции x_i, описывающей обучающий образец, формируется соответствующее ему конечное множество дочерних дискретных саундлетов S^c, находящихся в том же амплитудно-временном окне, что и эта функция, в виде

$$\forall x_i \in X\ \exists S^c = \left\{s_r^c \mid r \in \{1,\ldots,I\}\right\}\ :\ s_r^c = Fx_r.$$

Вычисляется нормированное расстояние между функцией, описывающей обучающий образец, и дочерним дискретным саундлетом в виде

$$\forall i, r \in \{1,\ldots,I\}$$

$$d_{ir} = \frac{\rho_p(x_i, s_r^c)}{(A_i^{\max} - A_i^{\min})\sqrt[p]{(N_i^{\max} - N_i^{\min} + 1)}},$$

$$\rho_p(x_i, s_r^c) = \sqrt[p]{\sum_{m\in Z} |\, x_i(m) - s_r^c(m)\,|^p}.$$

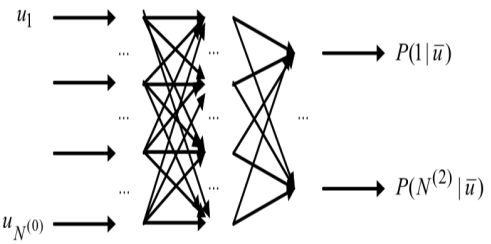
Осуществляется выбор множества функций H, описывающих опорные образцы, из множества функций X, описывающих обучающие образцы, на основе матрицы нормированных расстояний $[d_{ir}]$.

Каждая функция h_k множества H преобразуется к N-мерному вектору $\overline{v}_k$ множества V, причем все вектора $\overline{v}_k$ – одинаковые минимальные и максимальные значения *ноль* и A.

$$\forall h_k \in H\quad \exists \overline{v}_k : v_{kn} = (Fh_k)(n-1),\ \ n \in \{1,\ldots,N\}.$$

На основе введенных в [8] семейств саундлетов и саундлетных отображений и сформированного множества опорных образцов создадим модель классификации вокальных звуков.

Модель классификации вокальных звуков на основе саундлетной байесовской нейронной сети. Структура модели авторской двухслойной саундлетной байесовской нейронной сети (*SBNN*) представлена на рисунке.



Особенности предложенной модели *SBNN*:

- нейронам входного слоя соответствуют компоненты вектора, описывающего тестовый образец;
- нейроны первого (скрытого) слоя соответствуют опорным образцам;
- нейроны второго слоя соответствуют звукам;
- адаптация к голосовым особенностям конкретного оператора осуществляется путем добавления в модель векторов опорных образцов;
- каждый нейрон первого (скрытого) слоя обрабатывает информацию на основе нормированного расстояния между опорным и тестовым образцами звука;
- веса связей между нейронами первого (скрытого) и второго (выходного) слоев равны единице или нулю, для этих весов не требуется процедура обучения;
- агрегирование выходов нейронов первого (скрытого) слоя выполняется на основе максимума;
- во втором (выходном) слое вычисляются апостериорные вероятности по формуле Байеса, что позволяет определить вероятность принадлежности тестового образца звуку.

Дадим общую математическую постановку задачи классификации, которая может служить основой для построения моделей метрической классификации. Пусть x – функция, описывающая подлежащий классификации образец, y – номер класса образца (вокального звука речи). Задача заключается в том, чтобы по значению x определить значение величины y. Тогда построение модели нейросетевой классификации сводится к определению зависимости между номером класса y от значения x на основе нейронной сети.

Модель классификации вокальных звуков на основе *SBNN* представлена в виде

$$\bar{u} = (u_1,...,u_{N^{(0)}}),\ u_n = (Fx)(n-1),\quad n \in \left\{1,...,N^{(0)}\right\},$$

$$y = \arg\max_j P(j\,|\,\bar{u}),\quad \tilde{y} = \arg\max_j P(j\,|\,\bar{u}),$$

$$j \in \left\{1,...,N^{(2)}\right\},$$

$$P(j\,|\,\bar{u}) = \frac{P(j)P(\bar{u}\,|\,j)}{\sum_{j=1}^{N^{(2)}} P(j)P(\bar{u}\,|\,j)},\quad j \in \left\{1,...,N^{(2)}\right\},$$

$$P(\bar{u}\,|\,j) = \max_z w_{zj} P(\bar{u}\,|\,\bar{v}_z),\quad z \in \left\{1,...,N^{(1)}\right\},$$

$$j \in \left\{1,...,N^{(2)}\right\},$$

$$P(\bar{u}\,|\,\bar{v}_z) = 1 - \frac{\rho_p(\bar{v}_z,\bar{u})}{A\sqrt[p]{N^{(0)}}},\quad z \in \left\{1,...,N^{(1)}\right\},$$

$$\rho_p(\bar{v}_z,\bar{u}) = \sqrt[p]{\sum_{k=1}^{N^{(0)}} |\,v_{zk} - u_k\,|^p},$$

где y – номер звука; $\tilde{y}$ – максимум апостериорной вероятности; x – целочисленная ограниченная финитная дискретная функция, описывающая тестовый образец дискретного речевого сигнала; $\bar{u}$ – целочисленный вектор, полученный в результате преобразования функции x к единому амплитудно-временному окну на основе саундлетов и саундлетных отображений, поступающий на вход искусственной нейронной сети; $\bar{v}_z$ – целочисленный вектор, связанный с z-м нейроном скрытого слоя, соответствующий z-му опорному образцу; $P(j\,|\,\bar{u})$ – апостериорная вероятность (условная вероятность появления тестового образца j-го звука при наблюдении $\bar{u}$), вычисляемая для каждого j-го нейрона второго (выходного) слоя; $P(\bar{u}\,|\,j)$ – эмиссионная вероятность (вероятность наблюдения $\bar{u}$ при условии, что тестовый образец соответствует j-му звуку); $P(\bar{u}\,|\,\bar{v}_z)$ – условная вероятность (вероятность наблюдения $\bar{u}$ при условии, что тестовый образец соответствует вектору опорных образцов $\bar{v}_z$), вычисляемая для каждого z-го нейрона первого (скрытого) слоя; $P(j)$ – априорная вероятность появления образца j-го звука (безусловная вероятность), которая равновероятна в силу ограниченной статистики, т.е. $P(j) = \frac{1}{N^{(2)}}$;

A – максимальное значение вектора $\overline{u}$; $N^{(0)}$ – количество нейронов входного слоя, соответствующее длине вектора $\overline{u}$; $N^{(1)}$ – количество нейронов скрытого слоя, соответствующее количеству опорных образцов всех звуков; $N^{(2)}$ – количество нейронов выходного слоя, соответствующее количеству звуков; w_{zj} – вес z-го опорного образца, $w_{zj} \in [0, 1]$, причем если z-й нейрон скрытого слоя не связан с j-м нейроном выходного слоя (z-й опорный образец не соответствует j-му звуку), то $w_{zj}=0$. Если вес не учитывается, то $w_z \in \{0, 1\}$.

Для созданной модели сформулируем критерии эффективности.

Критерии оценки эффективности модели

• *Критерий скорости классификации* означает выбор из заданного набора метрик такой метрики, которая *на стадии обучения модели* требует наименьшего количества опорных образцов

$$F = T \to \min_{p}.$$

• *Критерий оценки пороговой вероятности классификации* означает выбор такого множества опорных образцов *на стадии опытной эксплуатации модели*, чтобы для тестового образца номер звука, вычисленный по модели, совпадал с тестовым номером звука этого тестового образца

$$F = \frac{1}{I}\sum_{i=1}^{I} \phi(y_i^{\text{model}}, y_i^{\text{test}}) \to \max_{\{H_j\}},$$

$$\phi(a,b) = \begin{cases} 1, & a = b \\ 0, & a \neq b \end{cases},$$

$$y_i^{\text{model}} = \arg\max_j P(j \mid \overline{u}_i), \quad j \in \{1,...,J\},$$

где $\overline{u}_i$ – i-й тестовый образец; y_i^{test} – тестовый номер звука для i-го тестового образца; I – количество тестовых образцов.

• *Для оценки готовности модели к эксплуатации* используется *критерий ее адекватности*, основанный на минимуме среднеквадратичной ошибки

$$F = \frac{1}{I}\sum_{i=1}^{I} (\tilde{y}_i^{\text{model}} - \tilde{y}_i^{\text{test}}) \to \min_{\{H_j\}},$$

$$\tilde{y}_i^{\text{model}} = \max_j P(j \mid \overline{u}_i); \quad j \in \{1,...,J\},$$

где $\overline{u}_i$ – i-й тестовый образец; $\tilde{y}_i^{\text{test}}$ – тестовый максимум апостериорной вероятности для i-го тестового образца.

Для оценивания результатов классификации вокальных звуков необходимо сформулировать условия их классификации.

Условия классификации тестового образца вокального звука по опорным образцам. Пусть дан тестовый образец вокального звука, который описывается целочисленным вектором $\overline{u}$.

Пусть для каждого j-го вокального звука вычислена эмиссионная вероятность $P(\overline{u} \mid j)$, т.е. вероятность наблюдения вектора $\overline{u}$, описывающего тестовый образец, при условии, что тестовый образец соответствует j-му звуку.

Необходимое условие классификации тестового образца. Тестовый образец классифицирован, если

$$\forall n \in \{1,...,J\} \quad \forall m \in \{1,...,J\}$$

$$\left(P(\overline{u} \mid n) = \max_j P(\overline{u} \mid j)\right) \wedge \left(P(\overline{u} \mid m) = \max_j P(\overline{u} \mid j)\right) \to$$

$$\to (n = m) \wedge (P(\overline{u} \mid n) > \tilde{\varepsilon}), \quad j \in \{1,...,J\},$$

где $\tilde{\varepsilon}$ – заданная эмиссионная вероятность классификации, $0 < \tilde{\varepsilon} \le 1$.

Достаточное условие классификации тестового образца. Тестовый образец классифицирован, если

$$\forall n \in \{1,...,J\}, \quad \forall m \in \{1,...,J\},$$

$$\left(P(\overline{u} \mid n) = \max_j P(\overline{u} \mid j)\right) \wedge \left(P(\overline{u} \mid m) = \max_j P(\overline{u} \mid j)\right) \to$$

$$\to (n = m) \wedge (P(\overline{u} \mid n) = 1), \quad j \in \{1,...,J\}.$$

На основе полученных условий можно сформировать логико-формальные правила оценивания результатов классификации.

Логико-формальные правила оценивания результата классификации. Для оценивания результатов классификации формируются следующие логико-формальные правила:

если $P(\overline{u} \mid y) > \tilde{\varepsilon}$, то $q = y$,

если $(P(\overline{u} \mid y) \le \tilde{\varepsilon})$, то $q = 0$,

где q – номер звука, $\tilde{\varepsilon}$ – вычисленный максимум эмиссионных вероятностей $P(\overline{u} \mid y)$ для

множества векторов тестовых образцов невокальных (непериодических) звуков.

Численное исследование метрического метода классификации вокальных звуков. В таблице приведено сравнение предложенного метода и существующих нейросетевых методов на основе базы данных TIMIT. Классификации подлежали все вокальные звуки.

Оценка нейросетевых методов классификации

Искусственные нейронные сети	**Ошибка классификации (%)**
Трехслойный *MLP*	0,80
RBFNN	0,81
GRNN	0,82
PNN	0,84
Трехслойный *RMLP*	0,90
Авторская *SBNN*	0,95

В неавторских методах в качестве образцов использовались вектора мел-частотные кепстральные коэффициенты (*MFCC*), вычисленные на участках равной длины, т.е. фреймах. Ошибка классификации представляет собой отношение количества правильно классифицированных образцов, содержащих вокальные звуки, к их общему количеству в процентах, при этом образцы, содержащие конец первого вокального звука и начало вокального второго звука, не учитывались. Приведенные в таблице стандартные нейросетевые методы реализованы автором статьи посредством пакета *Matlab*. Исследование позволяет сделать вывод, что авторский метод обеспечивает высокую вероятность классификации.

Заключение. В статье впервые предложено использовать саундлеты и саундлетные отображения применительно к искусственным нейронным сетям. Нейросетевой подход к классификации вокальных звуков, который отличается возможностью учитывать квазипериодическую структуру вокальных звуков и обобщать образцы одного звука различной длины с различным размахом амплитуд, что повышает эффективность классификации вокальных звуков речи, усовершенствован. Дальнейшее развитие получил метод создания множества опорных образцов, который основан на семействе саундлетов и саундлетных отображений, что повышает эффективность процедуры формирования опорных образцов. В рамках предложенных саундлетов и саундлетных отображений усовершенствована модель нейронной сети, которая позволяет сопоставлять образцы различной длины и использовать адаптивный нормированный порог в логико-формальных правилах, что повышает вероятность классификации полезных звуков.

Практическое значение состоит в том, что разработан метод построения модели классификации вокальных звуков на основе саундлетной байесовской нейронной сети, позволяющий сократить количество опорных образцов. Предложенный адаптивный нормированный порог для логико-формальных правил оценивания классификации речевых сигналов позволяет с большей вероятностью выделять полезные звуки. В результате численного исследования установлено, что алгоритм метрической классификации вокальных звуков на основе саундлетной байесовской нейронной сети дает вероятность классификации 0,95. Созданные алгоритмы можно использовать для решения задач, связанных с распознаванием речи оператора, синтезом речи, анализом вибрационного сигнала.

1. *Осовский С.* Нейронные сети для обработки информации. – М.: Финансы и статистика, 2002. – 344 с.
2. *Хайкин С.* Нейронные сети: Полный курс. – М.: Вильямс, 2006. – 1104 с.
3. *Каллан Р.* Основные концепции нейронных сетей. – М.: Там же, 2001. – 288 с.
4. *Rabiner L.R., Jang B.H.* Fundamentals of speech recognition. – Englewood Cliffs, NJ: Prentice Hall PTR, 1993. – 507 p.
5. *Потапова Р.К.* Речь: коммуникация, информация, кибернетика. – М.: Радио и Связь, 1997. – 528 с.
6. *Винцюк Т.К.* Анализ, распознавание и интерпретация речевых сигналов. – Киев: Наук. думка, 1987. – 261 с.
7. *Федоров Е.Е.* Методология создания мультиагентной системы речевого управления. – Донецк: Ноулидж, 2011. – 356 с.
8. *Федоров Е.Е.* Метод синтеза вокальных звуков речи по эталонным образцам на основе саундлетов // Наук. пр. Донецьк. нац. техн. ун-ту, 2014. – Т. 2. – С. 127–137.

E-mail: *fedorovee75@mail.ru*

UDC 004.934

Ye.Ye. Fedorov

The Classification Method of Vocal Speech Sounds on the Basis of Soundlet Bayesian Neural Network

Keywords: artificial neural network, mother soundlet, child soundlet, soundlet mapping, classification of vocal sounds, Bayesian approach.

The urgent task of developing a software component is a human speech recognition, using the intelligent computer systems. The basis of this problem is the construction of the effective methods, providing the high speed of the learning pattern recognition models as well as high probability, the adequacy and speed of speech signals recognition.

The existing speech recognition system images using the following approaches: logical, metric, Bayesian, artificial neural network, structural. The existing methods and models is usually based on hidden Markov models, dynamic programming algorithm DTW. The artificial neural networks have the following disadvantages: while learning a few months; the retention of the large amount of reference patterns (sounds and words), as well as weighting coefficients; big time recognition; probability of detection is less than 95 %; the presence of hundreds of thousands of training patterns.

To remedy these shortcomings, this article describes a method for the classification of vocal speech sounds on reference patterns on the basis of soundlet. The work improves the approach to detection of the vocal sounds, which allows to generalize the single sound patterns of different lengths and different swing amplitudes, which increases the efficiency of the classification of vocal speech sounds.

The author introduces the notion of a vocal sound sample and the method of its creation. Further development of the generating the plurality of the reference patterns method, which is characterized based on the soundlet and soundlet mappings collections, which increases the efficiency of the procedure of generating the reference patterns.

On the basis of the soundlet and soundlet mappings collections, the method of the vocal sounds classification is improved, based on soundlet Bayesian neural network (SBNN). The proposed model SBNN has the following characteristics: the neurons of the input layer correspond to the components of the vector that describes the test pattern; the neurons of the first (hidden) layer correspond to the reference patterns; the neurons of the second layer correspond to the sounds; adaptation to the voice characteristics of the particular operator is carried out by adding to the model vectors of the reference patterns; each neuron of the first (hidden) layer processes information based on normalized distances between the reference sound pattern and a test pattern of the sound; the weight of connections between neurons in the first (hidden) and second (output) layer is equal to 1 or 0 for these balances do not require the procedure of training; aggregation of outputs of neurons in the first (hidden) layer is performed on the basis of the maximum; in the second (output) layer are calculated posterior probabilities by Bayes formula, which allows to determine the probability that a test pattern vocal sound.

The numerical studies are conducted on the vocal sounds of the TIMIT database. Were use such artificial neural networks like MLP, RBFNN, GRNN, PNN, RMLP and author SBNN. The study allows to conclude that the author's method provides the highest probability of classification.

Algorithms can be used to solve the problems associated with the speech recognition in information systems, analysis of the vibration signal in intelligent systems technical diagnostics, the speaker identification in security systems and for the phonoscope examination.

●