

УДК 004.421

ББК 32.97

ЭКСПЕРИМЕНТАЛЬНАЯ ОЦЕНКА АЛГОРИТМОВ ИНТЕРПРЕТАЦИИ ПРОСОДИЧЕСКИХ ПРИЗНАКОВ РЕЧИ ПРИ ЕЕ СПЕЦИАЛЬНОЙ ОБРАБОТКЕ

Бессонов М.А.¹,

(ФГАОУ ВО «Российский университет дружбы народов»)

Фархадов М.П.²

(Учреждение Российской академии наук

Институт проблем управления РАН, Москва)

В статье в рамках решения задачи определения языка аудиосообщения на основе просодического подхода проводится экспериментальная оценка алгоритмов интерпретации просодических признаков речи – алгоритма на основе широких фонетических категорий и алгоритма на основе кросскорреляционной функции от мелодики речевого сигнала и последовательности кратковременных энергий. В качестве решающего правила используются нейронные сети.

Ключевые слова: идентификация языка, нейронные сети, просодические признаки речи, широкие фонетические категории.

1. Введение

Определение языка аудиосообщения является актуальной задачей, и выделяют 4 подхода ее решения – акустический, фонотактический, лексический и просодический [4-8]. Просодический подход использует такие параметра как мелодика речи,

¹ Бессонов Максим Александрович, аспирант (bessonovma@gmail.com).

² Фархадов Маис Паша Оглы, д.т.н., с.н.с., (mais@ipu.ru).

ритм, тембр и т.д. Ранее для интерпретации просодических параметров с целью их использования в обработке речи были предложены два алгоритма – на основе широких фонетических категорий [3] и на основе кросскорреляционной функции мелодии речи и последовательности кратковременных энергий.

Для проведения экспериментальной оценки формируется речевая база данных, далее создается и тестируется нейронная сеть.

2. Формирование речевой базы данных

Для проведения тестов в данной работе была сформирована база данных аудиозаписей, состав базы указан в таблице № 1.

Источник аудиозаписей – каналы интернет вещания – телевидение и радио, то есть речь, прошедшая обработку различными кодеками.

Для исключения влияния базы данных на эксперимент число дикторов по всем языкам выбрано одинаковым, суммарное время аудиозаписей выбрано одинаковым, также одинаков процент обучающей и тестовой выборок. Обучающая и тестовая выборки не перекрываются. Для проведения экспериментов все аудиозаписи обучающей и тестовой выборки разделялись на отрезки по 10 секунд. Выбор аудиозаписей для обучения и тестирования производился в случайном порядке.

Возрастной состав дикторов определить возможно приблизительно – мужчины и женщины от 20 до 50 лет.

Таблица 1. Характеристики базы данных для проведения экспериментальной оценки эффективности моделей описания просодических признаков

Язык	Число дикторов	Суммарное время аудиозаписей на каждого диктора, мин	Пол диктора (м-мужской, ж-женский)	Процент обучающей/тестовой выборки, %
Китайский	10	100	5м/5ж	80/20
Английский	10	100	5м/5ж	80/20
Финский	10	100	5м/5ж	80/20
Французский	10	100	5м/5ж	80/20

Немецкий	10	100	5м/5ж	80/20
Японский	10	100	5м/5ж	80/20
Персидский	10	100	5м/5ж	80/20
Португальский	10	100	5м/5ж	80/20
Русский	10	100	5м/5ж	80/20
Испанский	10	100	5м/5ж	80/20

3. Создание и настройка нейронной сети

Задача распознавания образов в большинстве случаев решается статистическими методами, но в случае речевых данных на различных языках достаточно сложно построить статистическое распределение рассматриваемых параметров, и поэтому в данной работе для классификации отрезков речи применены искусственные нейронные сети.

Как известно, для задач типа классификации число нейронов во входном слое вычисляется исходя из вектора признаков, который подается на вход [2], а число нейронов выходного слоя зависит от того, какая задача решается и какое применяется правило интерпретации выходных значений [2]. Для оценки числа нейронов в скрытых слоях применяют формулу

$$\frac{N_y N_p}{1 + \log_2(N_p)} \leq N_w \leq N_y \left(\frac{N_p}{N_x} + 1 \right) (N_x + N_y + 1) + N_y$$

где N_y – размерность выходного вектора НС, N_p – число элементов обучающей выборки, N_x – размерность входного вектора, N_w – общее число нейронов.

Выбор класса и архитектуры НС является нетривиальной задачей, для решения которой точных методов не существует [2]. Для выбора числа нейронов выделяют два метода: тем больше нейронов, тем надежнее работа сети, чем больше число нейронов, тем хуже создаваемая нейронная сеть аппроксимирует функцию.

Для реализации классификатора на базе нейронной сети был сделан выбор в пользу пакета MATLAB, который включает в себя функционал по нейронным сетям.

В работе экспериментальные исследования проводились со следующими сетями: сеть Кохонена, каскадная НС, сеть Элма-

на, многослойный персептрон, сеть Хопфилда, вероятностная сеть, сеть с радиальными базисными функциями RBF, НС встречного распространения – LVQ сети.

Алгоритмы, стандартные в MATLAB, использованные при обучении сетей [1]: квазиньютоновский алгоритм; алгоритм Левенберга-Марквардта с регуляризацией Байеса; метод сопряженных градиентов Флетчера-Ривса; метод сопряженных градиентов Полака-Ривьера; метод сопряженных градиентов Пауэлла-Беаля; базовый метод градиентного спуска; метод градиентного спуска с переменным шагом обучения; алгоритм Левенберга-Марквардта, метод масштабированных сопряженных градиентов; метод градиентного спуска с моментом; метод градиентного спуска с моментом и переменным шагом обучения; метод «One Step Secant»; метод случайных приращений; эластичный алгоритм обратного распространения ошибки.

На первом этапе для построения сокращенных групп из 10 языков эксперименты проводились с отдельной сетью для каждой пары языков, то есть было построено $\binom{10}{2} = 45$ нейронных сетей.

Наилучшие показатели были получены при создании многослойного персептрона, поэтому было принято решение провести более точную настройку данного типа НС.

Но поскольку заранее неизвестно, какой язык подается на вход НС, было принято решение строить одну общую НС для всех пар языков.

4. Оценка алгоритма на основе широких фонетических категорий

Согласно формуле и исходным параметрам для тестирования НС: $N_y=2$, $N_p=600$, $N_x=399$, число нейронов в скрытых слоях $117 \leq N_w \leq 2015$ для модели ШФК.

Поскольку N_w лежит в пределах от 117 до 2015, то при создании архитектуры НС число нейронов в слое варьировалось от 100 до 2000, соответственно число слоев от 1 (1 слой от 100 до 2000 нейронов) до 20 (20 слоев по 100 нейронов) в следующих конфигурациях: со 100 до 1000 нейронов с шагом в 10 нейронов

в слое, с 1000 до 2000 с шагом в 100 нейронов. Максимальное число нейронов в слое 800

При построении различных архитектур многослойного персептрона для 45 пар языков формировался вектор целевых показателей достоверности распознавания $D = (d_{1,2}, d_{2,1}, d_{1,3}, d_{3,1}, d_{i,j}, d_{j,i}, \dots, d_{N,N-1}, d_{N-1,N})$, где N – общее число языков в САОЯ. Таким образом, длина вектора $D = 90$. Каждый элемент $d_{i,j}, d_{j,i} = 100$.

Вектор показателей достоверности распознавания $D_k = (d^k_{1,2}, d^k_{2,1}, d^k_{1,3}, d^k_{3,1}, d^k_{i,j}, d^k_{j,i}, \dots, d^k_{N,N-1}, d^k_{N-1,N})$ для текущей архитектуры НС имеет также длину 90 и расстояние между D и D_k определяется как

$$D_r = \sqrt{(d_{1,2} - d^k_{1,2})^2 + (d_{2,1} - d^k_{2,1})^2 + (d_{i,j} - d^k_{i,j})^2 + \dots + (d_{j,i} - d^k_{j,i})^2 + \dots + (d_{N,N-1} - d^k_{N,N-1})^2 + (d_{N-1,N} - d^k_{N-1,N})^2}$$

Таким образом, тем меньше расстояние D_r , тем лучше настроена НС. В результате исследования D_r колебалась в пределах от 59.1861 до 532.4106. Наилучший показатель $D_r=72.5358$ был получен для конфигурации НС – общее число нейронов 1400, 1 слой 800 нейронов, 2 слой 600 нейронов. Результаты определения языка представлены в таблице 2.

Таблица 2. Средние значения достоверности определения языка

	китайский	английский	финский	французский	немецкий	японский	персидский	португальский	русский	испанский
китайский		94,5	95,1	96,2	95,9	97,5	96,6	95,2	94,4	97,9
английский	93,8		97,4	92,8	93,8	93,6	98,1	94,5	94,0	97,8
финский	93,8	93,7		93,2	93,4	93,9	93,9	96,1	93,7	94,3
французский	94,2	93,6	93,2		93,9	93,4	94,0	94,8	93,8	94,4
немецкий	94,5	92,6	93,7	92,5		94,6	94,0	97,5	96,3	93,9
японский	83,6	94,1	74,0	98,3	93,3		94,0	84,9	94,4	98,0

персидский	84,4	94,0	74,6	93,3	93,8	83,6		92,7	84,3	93,2
португаль- ский	94,2	93,6	93,5	93,9	94,2	94,5	93,5		93,9	98,4
русский	94,4	95,1	94,1	95,3	93,4	94,0	94,4	94,3		94,5
испанский	93,9	94,3	93,4	93,2	94,2	93,8	94,1	94,5	93,2	

Для группировки языков на группы применим агломеративный иерархический алгоритм. В качестве образов выступают пары языков, в качестве расстояния между образами – средние значения достоверности определения языка в паре при фиксированной вероятности ошибки первого и второго рода. В расстояния между классами используем расстояние по принципу ближайшего соседа:

$$\rho_{\min}(\omega_i, \omega_j) = \min_{x_k \in \omega_i, x_l \in \omega_j} d(x_k, x_l),$$

где ω_i, ω_j - языки L_i и L_j , $\rho(\omega_i, \omega_j)$ - расстояние между L_i и L_j .

Таким образом, получаем граф иерархической классификации. Исходя из полученного графа иерархической классификации, получаем группы схожести языков.

5. Оценка алгоритма на основе кросскорреляционной функции мелодии основного тона и последовательности кратковременных энергий

Согласно формуле и исходным параметрам для тестирования НС: $N_y=2$, $N_p=600$, $N_x=797$, число нейронов в скрытых слоях $117 \leq N_w \leq 2806$ для модели ВКФ от ОТ и КЭ.

Поскольку N_w лежит в пределах от 117 до 2806, то при создании архитектуры НС число нейронов в слое варьировалось от 100 до 3000, соответственно число слоев от 1 (1 слой от 100 до 3000 нейронов) до 20 (30 слоев по 100 нейронов) в следующих конфигурациях: со 100 до 1000 нейронов с шагом в 10 нейронов в слое, с 1000 до 3000 с шагом в 100 нейронов. Максимальное число нейронов в слое 800. $D_r=89.1449$.

Результаты определения языка представлены в таблице 4.

Таблица 4. Средние значения достоверности определения языка

	китайский	английский	финский	французский	немецкий	японский	персидский	португальский	русский	испанский
китайский		97,7	94,7	92,8	97,8	97,9	93,8	91,7	93,1	92,1
английский	91,2		91,4	92,3	92,9	94,8	92,7	97,7	90,3	92,0
финский	90,9	91,5		95,8	94,7	94,6	95,4	90,9	93,6	95,9
французский	92,1	92,9	92,4		93,9	96,7	97,5	92,1	91,8	91,8
немецкий	92,5	90,2	91,4	90,4		91,8	92,2	92,4	93,0	95,4
японский	80,6	91,8	90,1	82,3	71,9		90,7	90,5	94,7	97,2
персидский	71,1	91,5	82,3	91,6	82,6	78,2		97,5	92,5	91,5
португальский	90,7	91,0	92,0	92,0	93,2	93,4	92,1		94,5	92,5
русский	91,0	91,7	90,6	92,6	92,3	92,4	91,7	91,6		96,2
испанский	90,5	92,9	90,9	92,8	91,2	93,1	91,4	92,1	93,6	

6. Заключение

Как следует из приведенных таблиц, описание просодических параметров моделью ШФК дает большую достоверность определения языка, но незначительно в сравнении с ВКФ. Показатель близости текущих результатов определения языка к целевым D_T составил $D_T=72.5358$ для модели АКФ от ШФК и $D_T=89.1449$ для модели ВКФ от ОТ и КЭ.

Литература

1. ДЬЯКОНОВ, В.П., КРУГЛОВ, В.В. *Matlab 6.5 SP1/7/7 SP1/7 SP2 + Simulink 5/6. Инструменты искусственного интеллекта и биоинформатики.* – М.: СОЛОН-ПРЕСС, 2006. – 456 с.
2. КОМАРЦОВА, Л.Г., МАКСИМОВ, А.В. *Нейрокомпьютеры: Учебное пособие для вузов.* – 2-е изд., перераб. и

- доп. – М.: изд-во МГТУ им. Н.Э. Баумана, 2004. – 400 с.: ил. – (информатика в техническом университете)
3. МИЛОШЕНКО, А.А. *Разработка методики использования широких фонетических категорий в задачах верификации диктора* : диссертация кандидата технических наук : 05.13.17 /Милошенко Алексей Анатольевич; [место защиты: Моск. гос. ун-т путей сообщ. (МИИТ) МПС РФ], Москва, 2010. - 94 с. : ил.
 4. AMBIKAIRAJAN, E., LI, H., WANG, L., YIN, B. *Language Identification: A Tutorial IEEE Circuits and Systems Magazine (Volume:11 , Issue: 2)* pp. 82 – 108 IEEE, 2011
 5. BHATTACHARJEE, U., SARMAH, K. *Language identification system using MFCC and prosodic features*. International Conference on Intelligent Systems and Signal Processing (ISSP), Gujarat, 2013. – pp. 194 – 197.
 6. LEE, R., LEUNG, C.-C., MA, B. *Spoken Language Recognition with prosodic features*. IEEE Transactions on Audio, Speech, and Language Processing (Volume:21 , Issue: 9). pp. 1841 – 1853. IEEE Signal Processing Society, 2013
 7. MARTINEZ, D., JEIDA, E., ORTEGA, A. *Prosodic features and formant modeling for an ivector-based language recognition system*. in Proc. ICASSP, Vancouver, Canada, May 2013, pp. 6847-6851.
 8. MARTÍNEZ, D., BURGET, L., FERRER, L., SCHEFFER, N. *iVector-based prosodic system for language identification*. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp. 4861 – 4864, Kyoto, 2012.

EXPERIMENTAL EVALUATION OF ALGORITHMS DESCRIBES PROSODIC FEATURES IN SPEECH PROCESSING

Bessonov Maxim, «Russian peoples' friendship university», graduate student (bessonovma@gmail.com).

Farhadov Mais Pasha, Institute of Control Sciences of RAS, Moscow, Doctor of Science, Senior Researcher, (Moscow, Profsoyuznaya st., 65, mais@ipu.ru).

Abstract: This article describes experimental evaluation of algorithms algorithms that describe the prosodic features in language identification systems.

Keywords: identification of language, neural networks, prosodic features of speech, broad phonetic categories

*Статья представлена к публикации
членом редакционной коллегии ...заполняется редактором...*

*Поступила в редакцию ...заполняется редактором...
Опубликована ...заполняется редактором...*