УДК 004.421
ББК 32.97

АЛГОРИТМЫ ИНТЕРПРЕТАЦИИ ПРОСОДИЧЕСКИХ ПРИЗНАКОВ РЕЧИ ПРИ ЕЕ ОБРАБОТКЕ НИЗКОСКОРОСТНЫМИ КОДЕКАМИ

Бессонов М. А.[1]**,**
(*ФГАОУ ВО «Российский университет дружбы народов», Москва*)
Фархадов М. П.[2]
(*ФГБУН Институт проблем управления им. В.А. Трапезникова РАН, Москва*)

В рамках решения задачи определения языка аудиосообщения на основе просодического подхода предложены два алгоритма интерпретации просодических признаков речи и методика их использования – алгоритм на основе широких фонетических категорий и алгоритм на основе кросскорреляционной функции от мелодики речевого сигнала и последовательности кратковременных энергий. Проводится экспериментальная оценка алгоритмов. 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 в связи с развитием множества сетевых человеко-машинных интерфейсов, при этом в данные системы закладывается поддержка множества языков. Выделяют четыре подхода ее

[1] *Максим Александрович Бессонов*, *аспирант* (*bessonovma@gmail.com*).
[2] *Маис Паша Оглы Фархадов*, *д.т.н., с.н.с.* (*mais@ipu.ru*).

решения – акустический, фонотактический, лексический и просодический. Так или иначе, первые три строятся на одних параметрах речевого сигнала – акустических – мел-кепстральных коэффициентах, смещенных мел-кепстральных коэффициентах и т.д. Просодический подход [5–9] использует такие параметра как мелодика речи, ритм, тембр и т.д. Просодические параметры сложно поддаются описанию и математической интерпретации. И поэтому в данной статье предлагаются два алгоритма для комплексного описания просодических признаков речи с целью их использования в системах автоматического определения языка аудиосообщения. Первый алгоритм основан на широких фонетических категориях [4], второй на кросскорреляционной функции мелодии речи и последовательности кратковременных энергий.

Данные алгоритмы отличаются от известных тем, что они применимы для определения языка аудиосообщения по речи, прошедшей через низкоскоростные кодеки. Это обусловлено тем, что в низкоскоростных кодеках в канал связи передаются такие параметры, как частота основного тона, сигнал тон-шум и усиление на квазипериодических отрезках.

2. Алгоритмы интерпретации просодических признаков

2.1. АЛГОРИТМ НА ОСНОВЕ ШИРОКИХ ФОНЕТИЧЕСКИХ КАТЕГОРИЙ

Пусть множество $L = \{L_1, L_2, \ldots, L_N\}$ есть множество языков, на котором осуществляется процедура определения языка аудиосообщения, где N – общее число языков. Пусть каждый язык L_i представляется множеством аудиозаписей различных дикторов этого языка $L_i = \{l_1, l_2, \ldots, l_{Mi}\}$, где M_i – общее число аудиозаписей языка L_i.

Аудиозапись разбивается на квазистационарные сегменты $s_i(m)$ длительностью K отсчетов, где i – номер сегмента речевого сигнала, $i = 1, 2, \ldots, P$, P – общее число сегментов в аудиозаписи речевого сигнала, $m = 1, \ldots, K-1$. На каждом сегменте i вычисляется признак в соответствии с природой сегмента – вокализованный, невокализованный или пауза

(1) $A_i = T\left(s_i\left(m\right)\right), i = 1, 2, ..., P$,

где T – операция вычисления типа сегмента, а также кратковременная энергия сегмента

(2) $E_{k_i} = E\left(s_i\left(m\right)\right), i = 1, 2, ..., P$,

где E – операция вычисления кратковременной энергии сегмента. При работе алгоритма без восстановления исходной формы речевого сигнала параметры A_i и Ek_i берутся из кадров вокодерной передачи. Соответственно формируются последовательности $\vec{A} = \left(A_1, A_2, ..., A_P\right)$ и $\vec{E}k = \left(Ek_1, Ek_2, ..., Ek_P\right)$. Если сегмент классифицирован как пауза, то $A_i = 0$, если классифицирован как невокализованный, то $A_i = 1$. На каждом вокализованном сегменте вычисляется частота основного тона

(3) $F0_i = F\left(s_i\left(m\right)\right), i = 1, 2, ..., P$,

где F – операция вычисления частоты основного тона, и формируется последовательность $\overrightarrow{F0} = \left(F0_1, F0_2, ..., F0_P\right)$. При работе алгоритма без восстановления исходной формы речевого сигнала параметр $F0i$ берется их кадров вокодерной передачи. Диапазон изменения частоты основного тона аудиозаписей разбивается на 5 интервалов. Для вокализованных сегментов каждый сегмент обозначается цифрой в соответствии с тем, в какой интервал ЧОТ попадает значение частоты на данном сегменте

(4) $F0_{u_i} = UF\left(\overrightarrow{F0}\right), i = 1, 2, ..., P$,

где $F0u_i$ – уровень ЧОТ, UF – операция вычисления диапазона изменения ЧОТ и кодирования каждого сегмента цифровым обозначением, формируется последовательность $\overrightarrow{F0u} = \left(F0u_1, F0u_2, ..., F0u_P\right)$ – последовательность из значений ЧОТ на сегментах аудиозаписи. Далее вычисляются сегменты возрастания/убывания кратковременной энергии речевого сигнала

(5) $E_{u_i} = UE\left(\overrightarrow{E_k}\right), i = 1, 2, ..., P$,

Кодирующиеся Eu_i = (+/–)1 в зависимости от того, возрастает или убывает энергия соответственно, где UE – операция вычис-

ления возрастания/убывания кратковременной энергии речевого сигнала. Формируется последовательность $\overrightarrow{Eu} = (Eu_1, Eu_2, ..., Eu_P)$. Если данный сегмент относится к участку убыванию кратковременной энергии, цифровое значение ЧОТ умножается на (–1).

Для определения побочных и главных ударений определяется главный и побочный максимумы ЧОТ на отрезке между двумя паузами. Если положение максимума ЧОТ и кратковременной энергии совпадают во времени и максимальны на отрезке, то этот сегмент принимается за главный максимум, если максимумы во времени не совпадают, то сегмент принимается за побочный максимум $MAX_i = \Theta(\overrightarrow{F0u}, \overrightarrow{Eu})$, где Θ – операция определения главного и побочного максимумов ЧОТ и кратковременной энергии. Формируется последовательность

(6) $\overrightarrow{MAX} = (MAX_1, MAX_2, ..., MAX_P)$.

Таким образом, окончательная последовательность широких фонетических категорий (ШФК) аудиозаписи $\overrightarrow{X} = (X_1, X_2, ..., X_P)$ состоит из элементов X_i, где

(7) $$X_i = \begin{cases} 0, \text{ если } A_i - \text{пауза}, \\ 1, \text{ если } A_i - \text{невокализованный}, \\ 2, \text{ если } F0_{u_i} - \text{уровень 1}, \\ -2, \text{ если } F0_{u_i} - \text{уровень 1}, E_{u_i} = -1, \\ 3, \text{ если } F0_{u_i} - \text{уровень 2}, \\ -3, \text{ если } F0_{u_i} - \text{уровень 2}, E_{u_i} = -1, \\ 4, \text{ если } F0_{u_i} - \text{уровень 3}, \\ -4, \text{ если } F0_{u_i} - \text{уровень 3}, E_{u_i} = -1, \\ 5, \text{ если } F0_{u_i} - \text{уровень 4}, \\ -5, \text{ если } F0_{u_i} - \text{уровень 4}, E_{u_i} = -1, \\ 6, \text{ если } F0_{u_i} - \text{уровень 5}, \\ -6, \text{ если } F0_{u_i} - \text{уровень 5}, E_{u_i} = -1, \\ 7, \text{ если } MAX_i - \text{побочный максимум}, \\ 8, \text{ если } MAX_i - \text{главный максимум}. \end{cases}$$

На рис. 1 и рис. 2 приведены блок-схемы алгоритма кодирования сегментов речевого сигнала.

По последовательности широких фонетических категорий $\overrightarrow{X}$ вычисляется автокорреляционная функция $\vec{R} = \Psi(\overrightarrow{X})$, где Ψ – операция вычисления автокорреляционной функции.

В случае работы алгоритмов без восстановления исходной формы речевого сигнала значения ЧОТ берутся из кадров вокодерной передачи. В случае работы алгоритма с восстановлением исходной формы речевого сигнала требуется выбор алгоритма оценки частоты основного тона.

Для определения ЧОТ существуют различные алгоритмы [2]. В данной работе были проведены испытания готовых алгоритмов, реализуюших определение ЧОТ по автокорреляционной функции (АКФ) – алгоритм SIFT, по кратковременной функции средней разности (КФСР) – алгоритм AMDF, а также алгоритм оценки ЧОТ из алгоритма кодирования речи MELP. Проценты отрезков речевого сигнала с показателями *P*(ОТ) – правильно определенным основным тоном, *P*(НВ/В) – принятия вокализованного отрезка за невокализованный, *P*(В/НВ) – принятия невокализованного за вокализованный приведены в таблице 1.

Таблица 1.Показатели правильности оценки основного тона

Алгоритм	SIFT	AMDF	MELP
P(ОТ), %	87±1	89±1	95±1,5
P(НВ/В), %	7±1	6±1	3±0,5
P(В/НВ), %	0,5	0,5	0,5

Как следует из экспериментального сравнения представленных алгоритмов, наилучшим оказался MELP. Данный алгоритм был выбран для оценки основного тона.

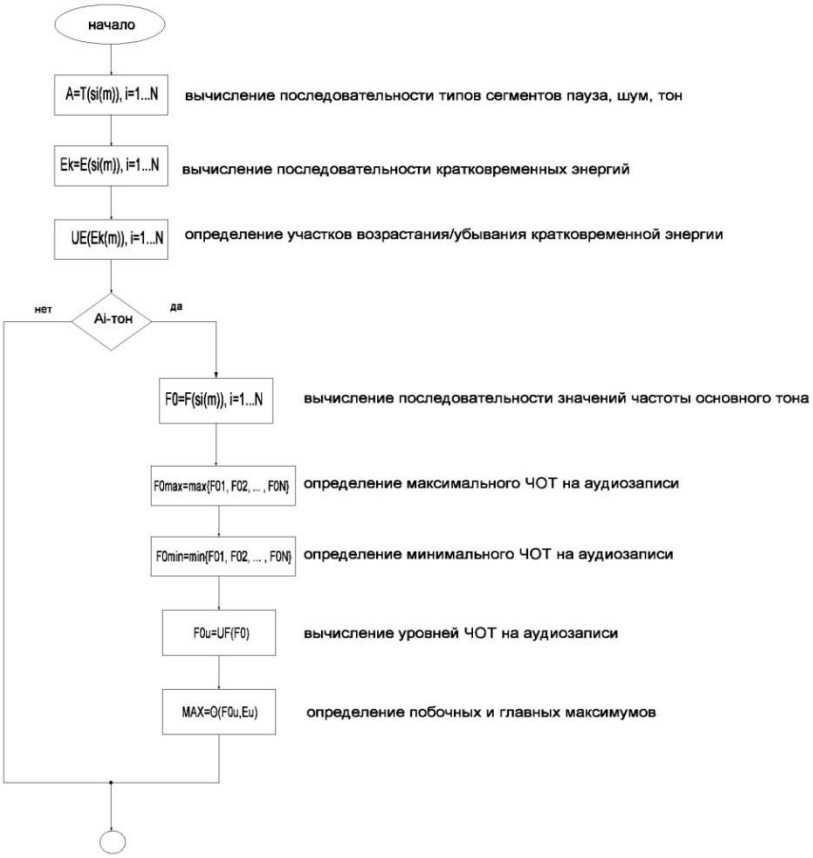

Рис. 1. Блок-схема алгоритма кодирования сегментов речевого сигнала

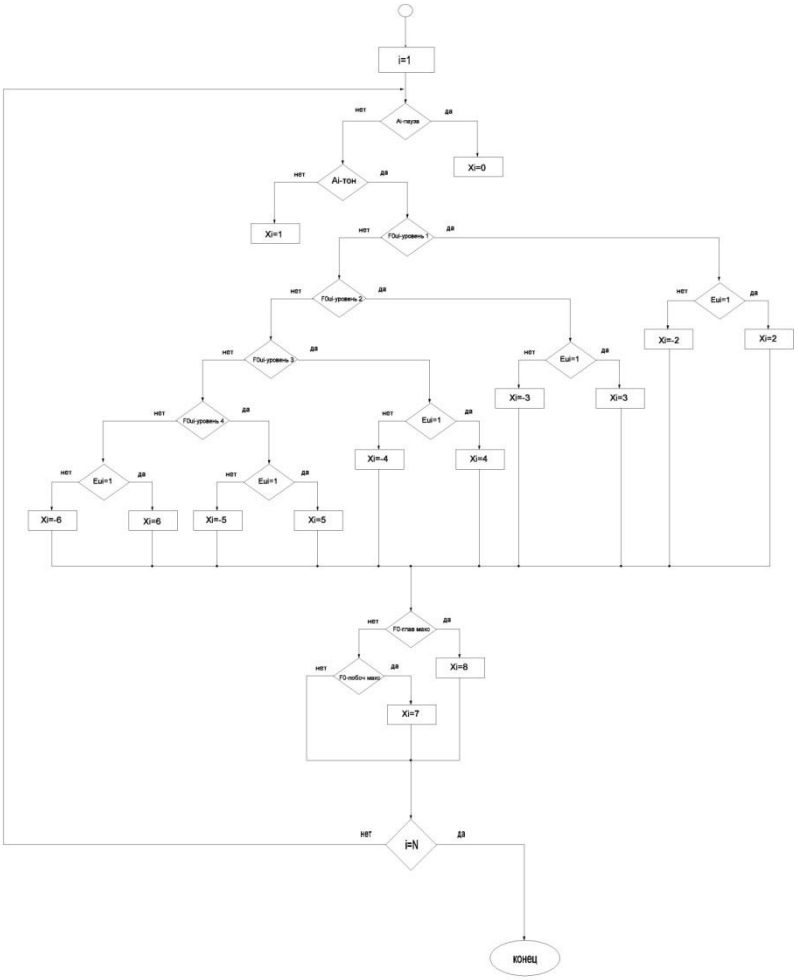

Рис. 2. Блок-схема алгоритма кодирования сегментов речевого сигнала (продолжение)

2.2. АЛГОРИТМ НА ОСНОВЕ КРОССКОРРЕЛЯЦИОННОЙ ФУНКЦИИ МЕЛОДИИ ОСНОВНОГО ТОНА И ПОСЛЕДОВАТЕЛЬНОСТИ КРАТКОВРЕМЕННЫХ ЭНЕРГИЙ

Для реализации просодической классификации предлагается использование кросскорреляционной функции мелодии основного тона и последовательности кратковременных энергий сигналов аудиозаписей. Аудиозапись разбивается на квазистационарные сегменты $s_i(m)$ длительностью K отсчетов, где i – номер сегмента речевого сигнала, $i = 1, 2, \ldots, P$, P – общее число сегментов в аудиозаписи речевого сигнала, $m = 1, \ldots, K - 1$. На каждом сегменте i вычисляется признак в соответствии с природой сегмента – вокализованный, невокализованный или пауза

(8) $A_i = T\left(s_i\left(m\right)\right), i = 1, 2, ..., P$,

где T – операция вычисления типа сегмента, а также кратковременная энергия сегмента

(9) $E_{k_i} = E\left(s_i\left(m\right)\right), i = 1, 2, ..., P$,

где E– операция вычисления кратковременной энергии сегмента. Соответственно формируются последовательности $\vec{A} = \left(A_1, A_2, ..., A_P\right)$ и $\vec{E}k = \left(Ek_1, Ek_2, ..., Ek_P\right)$. Если сегмент классифицирован как пауза, то $A_i = 0$, если классифицирован как невокализованный, то $A_i = 1$. На каждом вокализованном сегменте вычисляется частота основного тона (ЧОТ)

(10) $F0_i = F\left(s_i\left(m\right)\right), i = 1, 2, ..., P$,

где F – операция вычисления частоты основного тона, и формируется последовательность $\overrightarrow{F0} = \left(F0_1, F0_2, ..., F0_P\right)$.

При работе алгоритма без восстановления исходной формы речевого сигнала параметры A_i и Ek_i, $F0$i берутся из кадров вокодерной передачи.

По последовательности значений частоты основного тона и последовательности кратковременных энергий вычисляется их кросс-корреляционная функция

(11) $\vec{B} = \Phi(\overrightarrow{F0}, \overrightarrow{Ek})$,

где Ф – операция вычисления кросскорреляционной функции мелодии основного тона и последовательности кратковременных энергий. Вектор значений кросскорреляционной функции последовательности широких фонетических категорий подается на вход нейронной сети, которая принимается решение по отнесению данного вектора к какой-либо группе языков.

Алгоритм вычисления признаков представлен на рис. 3

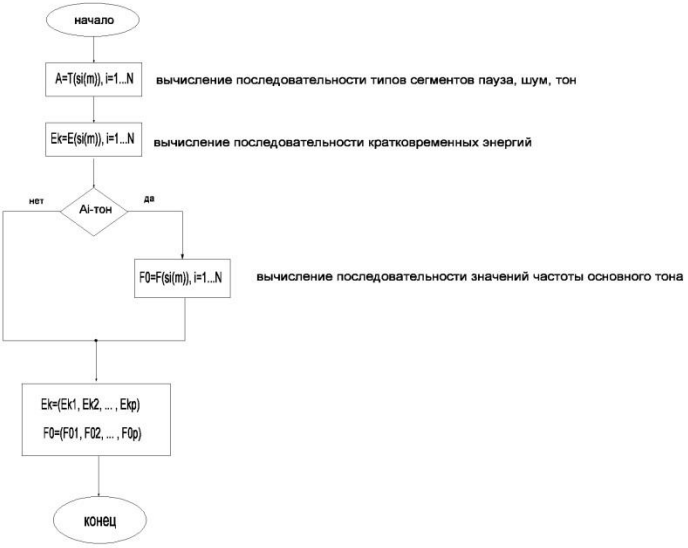

Рис. 3. Блок-схема алгоритма кодирования сегментов речевого сигнала

3. Методика применения алгоритмов интерпретации просодических признаков в задаче определения языка аудиосообщения

Для применения указанных алгоритмов была разработана следующая методика. Она заключается в последовательности ряда этапов.

Этап 1. Формирование обучающей речевой базы данных. Обучающая база данных должна удовлетворять следующим условиям: если N – общее число языков, d_m^i – число дикторов мужского пола языка i, d_f^i – число дикторов женского пола языка i, то $V_i(d_m^i, d_f^i) = V_j(d_m^j, d_f^j)$, где i, j – номера языков, $i, j = 1, \ldots, N$. То есть все возрастные группы должны быть представлены в равной пропорции дикторами мужского и женского пола, объемы речевых данных дикторов различных возрастных групп должны быть одинаковы. Объем речевых данных должен быть достаточен со статистической точки зрения для описания всех вариативностей произношения на данном языке. Общие объемы речевых баз по языкам должны быть равны.

Шаг 1. Получение от источника аудиосообщения в цифровом виде $S_t(f_d, m, p, f_r)$ с параметрами – формат f_r = «wav», частота дискретизации f_d = 8 кГц, режим m = моно, p = 16 бит, t – номер аудиосообщения.

Шаг 2. Фильтрация аудиосообщения $S_t(f_d, m, p, f_r)$ – удаление посторонних шумов. Получение фильтрованного аудиосообщения $S_t^f(f_d, m, p, f_r) = P[S_t(f_d, m, p, f_r)]$, где P – операция фильтрации.

Шаг 3. Формирование обучающих и тестовых данных. Для каждого языка L_i формируется база аудиосообщений $Z_{Li}\{S_1^f{}_{Li}(f_d, m, p, f_r), \quad S_2^f{}_{Li}(f_d, m, p, f_r), \ldots, S_{Mi}^f{}_{Li}(f_d, m, p, f_r)\}$, где Mi – общее число аудиосообщений языка Li.

Общая база аудиосообщений $Z = \{Z_{L1}, Z_{L2}, \ldots, Z_{LN}\}$.

Шаг 4. Обработка всех аудиосообщений всех языков заданным вокодером.

$Z^{vok} = \mathrm{VOK}(Z)$, где VOK – операция обработки базы аудиосообщений вокодером, $Z^{vok} = \{Z^{vok}_{L1}, Z^{vok}_{L2}, \ldots, Z^{vok}_{LN}\}$.

Шаг 5. Вычисление параметров из аудиосообщений в соответствии с разработанными алгоритмами – формирование базы параметров $Z^{vok}_{Li}{}^{Mod1} = \mathrm{Mod1}(Z^{vok}_{Li})$, $Z^{vok}_{Li}{}^{Mod2} = \mathrm{Mod2}(Z^{vok}_{Li})$, где Mod1, Mod2 –операции вычисления параметров в соответствии с разработанными алгоритмами описания просодических параметров речи.

Этап 2. Обучение искусственной нейронной сети, в процессе обучения происходит настройка различных параметров нейронной сети. Нейронные сети с различной топологией опи-

сываются различными математическими моделями, поэтому в каждом конкретном случае нейронная сеть будет описываться своей формулой. Для формирования групп языков строятся нейронные сети, число которых равно сочетанию из N по 2.

Этап 3. Тестовая оценка нейронной сети.

Шаг 1. Получение от источника аудиосообщения в цифровом виде $S_t(f_d, m, p, f_r)$ с параметрами: формат f_r = «wav», частота дискретизации f_d = 8 кГц, режим m = моно, p = 16 бит, t – номер аудиосообщения.

Шаг 2. Фильтрация аудиосообщения $S_t(f_d, m, p, f_r)$ – удаление посторонних шумов. Получение фильтрованного аудиосообщения $S_t^f(f_d, m, p, f_r) = P[S_t(f_d, m, p, f_r)]$, где P – операция фильтрации.

Шаг 3. Тестирование нейронных сетей. На вход нейронной сети для каждой пары языков L_i и L_j подаются аудиосообщения языков i и j, на выходе – оценка того, какому языку принадлежит данное аудиосообщение, t – номер аудиосообщения.

$$\hat{L}_t = NET\left(S_t^f\left(f_d, m, p, f_r\right)\right). \quad (12)$$

Шаг 4. Вычисление числа правильно распознанных аудиосообщений в каждой паре языков. Получение вектора $D = (d_{12}, d_{21}, d_{13}, d_{31}, \ldots, d_{N(N-1)}, d_{(N-1)N})$, где d_{ij} – число правильно определенных аудиосообщений для пары языков L_iL_j, $i \neq j$.

Шаг 5. Построение иерархического дерева языков на основе агломеративного иерархического алгоритма

$$\rho_{min}\left(\omega_i, \omega_j\right) = \min_{x_k \in \omega_i, x_l \in \omega_j} d\left(X_k, X_l\right), \quad (13)$$

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

На основе иерархического дерева строятся группы языков.

4. Формирование речевой базы данных

Для проведения тестов в данной работе была сформирована база данных аудиозаписей, состав базы указан в таблице 2.

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

Таблица 2. 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

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 выборок разделялись на отрезки по 10 с.

Возрастной состав дикторов определить возможно приблизительно – мужчины и женщины от 20 до 50 лет, объем обучающей выборки – 80% от времени аудиозаписей каждого диктора, объем тестовой выборки – 20% от времени аудиозаписей каждого диктора. Деление аудиозаписей на обучающую и тестовую выборки произведено в случайном порядке.

5. 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 [3], а число нейронов выходного слоя зависит от того, какая задача решается и какое применяется правило интерпретации выходных значений [3]. Для оценки числа нейронов в скрытых слоях применяют формулу [3]

$$(14)\quad \frac{N_y N_p}{1+log_2\left(N_p\right)} \le N_w \le N_y\left(\frac{N_p}{N_x}+1\right)\left(N_x+N_y+1\right)+N_y,$$

где N_y – размерность выходного вектора нейросети (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 [3]. Для выбора числа нейронов выделяют два метода: 1) чем больше нейронов, тем надежнее работа сети; 2) 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 Элмана, 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 гради-

ентов Полака-Ривьера; метод сопряженных градиентов Пауэлла-Беаля; базовый метод градиентного спуска; метод градиентного спуска с переменным шагом обучения; алгоритм Левенберга-Марквартда, 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 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 использовать единую архитектуру для сетей каждой пары языков.

6. 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 $117 \le N_w \le 2015$ для модели ШФК.

Поскольку Nw 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 $D = (d_{1,2}, d_{2,1}, d_{1,3}, d_{3,1}, d_{i,j}, d_{j,i}, d_{N,N-1}, d_{N-1,N})$, 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 $D_k = (d^k_{1,2}, d^k_{2,1}, d^k_{1,3}, d^k_{3,1}, d^k_{i,j}, d^k_{j,i}, d^k_{N,N-1}, d^k_{N-1,N})$ для текущей архитектуры НС имеет также длину 90 и расстояние между D и D_k определяется как

$$(15)\ D_r = \sqrt{\left(d_{1,2} - d^k_{1,2}\right)^2 + \left(d_{2,1} - d^k_{2,1}\right)^2 + \left(d_{i,j} - d^k_{i,j}\right)^2 + \ldots}$$
$$\ldots + \left(d_{j,i} - d^k_{j,i}\right)^2 \ldots + \left(d_{N,N-1} - d^k_{N,N-1}\right)^2 + \left(d_{N-1,N} - d^k_{N-1,N}\right)^2 .$$

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 3.

Таблица 3. 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
португальский	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 качестве расстояния между классами используем расстояние по принципу ближайшего соседа:

(16) $\rho_{min}\left(\omega_i,\omega_j\right)=\min\limits_{x_k\in\omega_i,x_l\in\omega_j} d\left(X_k,X_l\right)$,

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.

7. 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 $117 \le N_w \le 2806$ для модели кросскорреляционной функции от последовательности значений основного тона и кратковременной энергии речевого сигнала.

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	

8. Заключение

Целью описанных в статье алгоритмов является комплексное описание просодических признаков речи для того, чтобы эти признаки можно было использовать при специальной обработке данных, в частности в задаче определения языка аудиосообщения. Как следует из приведенных таблиц, описание просодических параметров моделью ШФК дает большую достоверность определения языка, но незначительно в сравнении с кросскорреляционной функцией. Показатель близости текущих результатов определения языка к целевым D_r составил D_r = 72,5358 для модели АКФ от ШФК и D_r = 89,1449 для модели кросскорреляционной функции от последовательности значений основного тона и кратковременной энергии речевого сигнала.

Отличительной особенностью данных алгоритмов является то, что они применимы при определении языка аудиосообщения

по речи, преобразованной вокодерами, без восстановления исходной формы речевого сигнала.

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. ИМАМВЕРДИЕВ Я.Н., СУХОСТАТ Л.В. *Подходы для оценки периода основного тона речевого сигнала в зашумлённой среде* // Речевые технологии. – 2014. – №1-2. – С. 84–102.
3. 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.
4. МИЛОШЕНКО А.А. *Разработка методики использования широких фонетических категорий в задачах верификации диктора*: Автореф. дис. канд. техн. наук. – Москва, 2010. – 94 с.
5. AMBIKAIRAJAH E., LI H., WANG L., YIN B. *Language Identification: A Tutorial* // IEEE Circuits and Systems Magazine. – 2011. – Vol. 11, Iss. 2. – P. 82–108.
6. BHATTACHARJEE U., SARMAH K. *Language identification system using MFCC and prosodic features* // Int. Conference on Intelligent Systems and Signal Processing (ISSP), Gujarat, 2013. – P. 194–197.
7. LEE R., LEUNG C.-C., MA B. *Spoken Language Recognition with prosodic features* // IEEE Trans. on Audio, Speech, and Language Processing. – 2013. – Vol. 21, Iss. 9. – P. 1841–1853.
8. MARTINEZ D., JEIDA E., ORTEGA A. *Prosodic features and formant modeling for an ivector-based language recognition system* // Proc. ICASSP, Vancouver, Canada, May 2013. – P. 6847–6851.
9. MARTÍNEZ D., BURGET L., FERRER L., SCHEFFER N. *iVector-based prosodic system for language identification* // IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 2012. – P. 4861–4864.

ALGORITHMS FOR INTERPRETATION OF PROSODIC FEATURES IN LOW-BITRATE SPEECH PROCESSING

Maxim Bessonov, «Russian peoples' friendship university», Moscow, graduate student (bessonovma@gmail.com).
Mais Pasha Farhadov, Institute of Control Sciences of RAS, Moscow, Doctor of Science, Senior Researcher, (Moscow, Profsoyuz naya st., 65, *mais@ipu.ru*).

Abstract: *We study the language identification problem using prosodic features. Prosodic features such as melody, rhythm, timbre and others are difficult to formalize mathematically. Two algorithms for a complex description of prosodic features are proposed in the paper. The first is based on the broad phonetic categories, and the second is based on the cross-correlation of the speech melody and the short-term energy sequence. The fundamental frequency was estimated by MELP algorithm. The performance of the proposed algorithms was evaluated experimentally on a database of speech recordings obtained from Internet and therefore encoded by low-bitrate vocoders. The database includes ten different languages. The proposed algorithms provide a feature description and a multi-layer neural network was used as a language classifier. Both algorithms show satisfactory classification performance, but the broad phonetic categories approach performs slightly better than the cross-correlation function. These algorithms can be applied to a speech signal processed by low-bitrate vocoders without decoding to the original signal.*

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

Статья представлена к публикации членом редакционной коллегии Н.И. Базенковым.

Поступила в редакцию 09.11.2016.
Опубликована 31.03.2017.