УДК 004.421
ББК 32.97

АЛГОРИТМЫ ИНТЕРПРЕТАЦИИ ПРОСОДИЧЕСКИХ ПРИЗНАКОВ РЕЧИ ПРИ ЕЕ ОБРАБОТКЕ ВЫЧИСЛИТЕЛЬНЫМИ МАШИНАМИ

Бессонов М.А.[1],
(*ФГАОУ ВО «Российский университет дружбы народов»*)

В статье 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.

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 4 подхода ее 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 [3-7] использует такие параметра как мелодика речи, ритм, тембр и т.д. Просо-

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

дические 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 [2], второй на кросскорреляционной функции мелодии речи и последовательности кратковременных энергий.

2. Алгоритмы интерпретации просодических признаков

2.1. АЛГОРИТМ НА ОСНОВЕ ШИРОКИХ ФОНЕТИЧЕСКИХ КАТЕГОРИЙ

Пусть множество $L = \{L_1, L_2, ..., L_N\}$ 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 $L_i = \{l_1, l_2, ..., l_{Mi}\}$, где M_i – общее число аудиозаписей языка L_i.

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

$A_i = T(s_i(m))$, i=1,2…P,

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

$E_{ki} = E(s_i(m))$, i=1,2…P,

где E – операция вычисления кратковременной энергии сегмента. Соответственно формируются последовательности $\vec{A} = (A_1, A_2, ..., A_P)$ и $\vec{E}k = (Ek_1, Ek_2, ..., Ek_P)$. Если сегмент классифицирован как пауза, то $A_i = 0$, если классифицирован как невокализо-

ванный, то $A_i = 1$. На каждом вокализованном сегменте вычисляется частота основного тона

$F0_i = F(s_i(m))$, i=1,2…P,

где F – операция вычисления частоты основного тона, и формируется последовательность $\overrightarrow{F0} = (F0_1, F0_2, ..., F0_P)$. Диапазон изменения ЧОТ аудиозаписей разбивается на 5 интервалов. Для вокализованных сегментов каждый сегмент обозначается цифрой в соответствии с тем, в какой интервал ЧОТ попадает значение ЧОТ на данном сегменте.

$F0_{u_i} = UF(\overrightarrow{F0})$, i=1,2…P,

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

$E_{u_i} = UE(\vec{E}_k)$, i=1,2…P,

кодирующиеся $Eu_i = (+/-)1$ в зависимости от того, возрастает или убывает энергия соответственно, где *UE* - операция вычисления 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 $\overrightarrow{MAX} = (MAX_1, MAX_2, ..., MAX_P)$

Таким образом, окончательная последовательность ШФК аудиозаписи $\overrightarrow{X} = \left(X_1, X_2, ..., X_P\right)$ состоит из элементов Xi, где

$$X_i = \begin{cases} 0, если & A_i - пауза \\ 1, если & A_i - невокализованны \\ 2, если & F0u_i - уровень\ 1 \\ -2, если & F0u_i - уровень\ 1,\ Eu_i = -1 \\ 3, если & F0u_i - уровень\ 2 \\ -3, если & F0u_i - уровень\ 2,\ Eu_i = -1 \\ 4, если & F0u_i - уровень\ 3 \\ -4, если & F0u_i - уровень\ 3,\ Eu_i = -1 \\ 5, если & F0u_i - уровень\ 4 \\ -5, если & F0u_i - уровень\ 4,\ Eu_i = -1 \\ 6, если & F0u_i - уровень\ 5 \\ -6, если & F0u_i - уровень\ 5,\ Eu_i = -1 \\ 7, если & MAX_i - побочный\ максимум \\ 8, если & MAX_i - главный\ максимум \end{cases}$$

На рисунке 1 и рисунке 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.

Для определения ЧОТ существуют различные алгоритмы [1]. В данной работе были проведены испытания готовых алгоритмов, реализуюших определение ЧОТ по АКФ – алгоритм SIFT, по КФСР – алгоритм AMDF, а также алгоритм оценки ЧОТ из алгоритма кодирования речи MELP. Проценты отрезков речевого сигнала с показателями P(ОТ) - правильно определенным ОТ, 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.

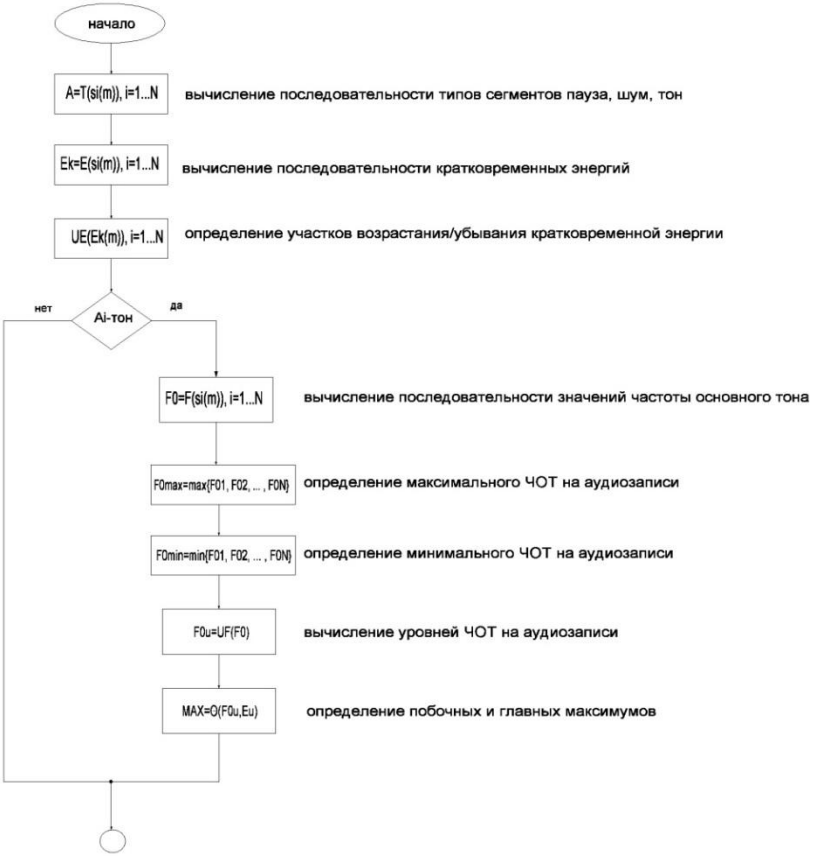

Рис. 1. Блок-схема алгоритма кодирования сегментов речевого сигнала

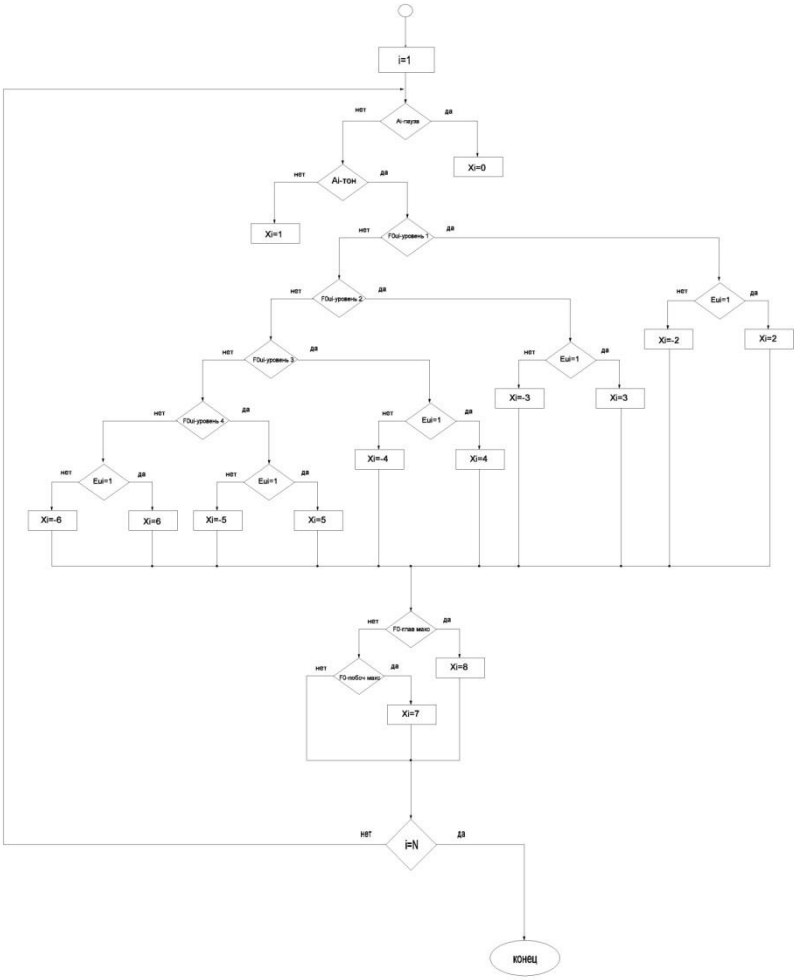

Рис. 2. Блок-схема алгоритма кодирования сегментов речевого сигнала (продолжение)

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 – i-й сегмент речевого сигнала, i=1,2…P, P – общее число сегментов в аудиозаписи речевого сигнала, $m = 1,..., K-1$. На каждом сегменте i вычисляется признак в соответствии с природой сегмента – вокализованный, невокализованный или пауза

$A_i = T(s_i(m))$, i=1,2…P,

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

$E_{ki} = E(s_i(m))$, i=1,2…P,

где E – операция вычисления кратковременной энергии сегмента. Соответственно формируются последовательности $\vec{A} = (A_1, A_2,..., A_P)$ и $\vec{E}k = (Ek_1, Ek_2,..., Ek_P)$. 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

$F0_i = F(s_i(m))$, i=1,2…P,

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

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

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

где Φ - операция вычисления кросскорреляционной функции мелодии основного тона и последовательности кратковременных энергий. Вектор значений кросскорреляционной функ-

ции 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 рисунке 3

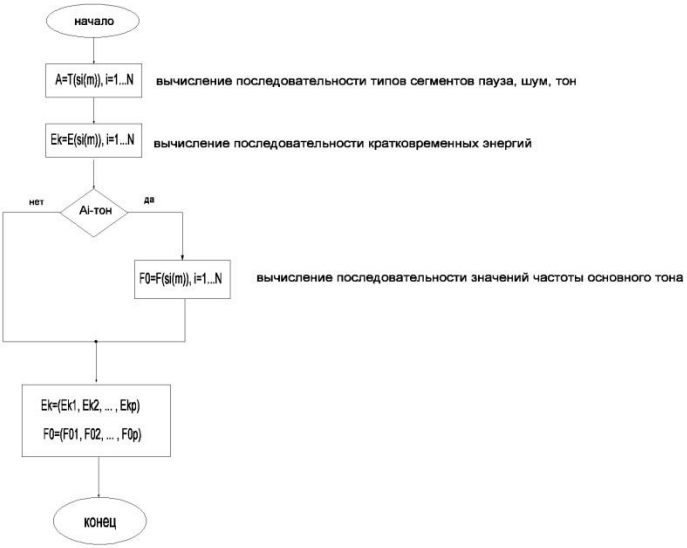

Рис. 3. Блок-схема алгоритма кодирования сегментов речевого сигнала

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…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 $Z_{Li}\{S_1^f{}_{Li}(f_d,m,p,f_r), S_2^f{}_{Li}(f_d,m,p,f_r),\ldots, S_{Mi}^f{}_{Li}(f_d,m,p,f_r)\}$, где M_i – общее число аудиозаписей языка L_i.

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

Шаг 4. Вычисление параметров из аудиосообщений в соответствии с разработанными моделями – формирование базы параметров $Z_{Li}^{Mod1}=Mod1(Z_{Li})$, $Z_{Li}^{Mod2}=Mod2(Z_{Li})$, 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 описываются различными математическими моделями, поэтому в каждом конкретном случае нейронная сеть будет описываться своей формулой. Для формирования групп языков строятся $\binom{N}{2}$ нейронных сетей.

Этап 3.Формирование групп языков

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 Li и Lj подаются аудиосообщения языков i и j, на выходе – оценка того, какому языку принадлежит данное аудиосообщение, t – номер аудиосообщения.

$$\hat{L}_t = NET(S_t^f(f_d, m, p, f_r))$$

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≠j.

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

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

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

На основе иерархического дерева строятся группы языков.

4. 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. В качестве классификатора могут выступать смеси гауссовых распределений, модели опорных векторов или нейронные сети.

Литература

1. ИМАМВЕРДИЕВ, Я.Н., СУХОСТАТ, Л.В. *Подходы для оценки периода основного тона речевого сигнала в за-*

шумлённой среде // Речевые технологии, 1-2/2014, стр.84-102
2. 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.
3. AMBIKAIRAJAH, 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
4. 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.
5. 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
6. 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.
7. 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.

INTERPRETATION ALGORITHMS OF 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 the 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…