# Модель системы идентификации дикторов, использующей сегментацию диапазона основного тона голоса

***Леднов Д.А.,***
*кандидат технических наук,*
***Хацкевич А.В.,***
***Широкова А.М.***

**В статье описывается метод идентификации дикторов, основанный на выделении гармонической структуры спектра вокализованных звуков, пригодный для использования в условиях телефонного канала. Для определения гармонического спектра используется свёртка спектра с вейвлет-функцией типа «Мексиканская шляпа». На основе анализа значения этих свёрток в различных областях спектра вводится классификация типов спектра. Для каждого спектрального типа определяется свой метод выделения характеристик речи. При реализации процедур обучения и распознавания для каждого спектрального класса вводится своя модель гауссовой смеси.**

**The speaker identification method based on the extraction of the spectrum harmonical structure of vocalized speech, suitable for the use in the telephone channel conditions, is described. For the detection of the harmonical spectrum the spectrum convolution with wavelet-function «Mexican hat» is used. On the basis of the convolution value analysis in the different spectrum area a classification of the spectra types is made. For each class of the harmonical spectrum an individual Gaussian Mixed Model GMM for training and identification procedures is created. Experimental results are presented.**

## Введение

Системы идентификации дикторов, основанные на модели гауссовых смесей (GMM), в которых в качестве метода предварительной обработки используется метод melfrequency cepstral coefficients (MFCC) [1], на практике применяются достаточно давно и демонстрируют высокую точность [2] для «чистых» речевых сигналов. В случае фиксации речи на фоне шумов эффективность MFCC падает из-за использования процедуры интегрирования взвешенных компонент спектра, которая является неотъемлемой частью метода. Этот факт заставляет исследователей искать модели обработки, обладающие устойчивостью к шумам. В работе [3] был описан метод идентификации дикторов, построенный на основе выделения гармонической структуры сигнала [4]. Такой метод имеет ряд преимуществ:

- выделение гармонической структуры сигнала позволяет выделять речь на фоне широкого класса шумов стационарного и нестационарного характера, обладающих сплошным спектром;
- известно [5], что основную информацию о свойствах голоса диктора несут вокализованные звуки речи, спектры которых обладают гармонической структурой;
- наличие гармонической структуры позволяет очищать речь от фоновых шумов, а затем синтезировать «чистую» речь и получать те её характеристики, которые предполагается использовать в процессе идентификации дикторов [3].

В этой работе будут развиты методы выделения гармонической структуры сигнала и идентификации дикторов, пригодные для использования в условиях телефонного канала. Основания этих методов были защищены патентами [6, 7].

Применить метод поиска гармонической структуры, описанный в [3], для обработки речи, переданной по телефонному каналу, не позволили следующие его особенности:

- использование окон анализа большой длительности (160 ms) и высокой частоты оцифровки сигнала (16 кГц);
- поиск гармонической структуры в спектре как в целом.

Второе замечание требует пояснения. Очевидно, что спектр вокализованного звука может включать как области, содержащие линейчатый спектр, так и области, характеризующиеся сплошным спектром. Таким образом, описание спектра как целого, обладающее свойством линейчатости во всех областях, несёт в себе ошибку, которой постараемся избежать в новом методе.

В дополнение отметим, что здесь будут приведены экспериментальные результаты, касающиеся некоторых свойств речи, которые удалось выявить в ходе применения разработанного метода.

## Модель выделения характеристик (Метод смешанных структур (MCC))

Итак, цель разрабатываемой модели — выделение гармонической структуры и определение на её основе характеристик голоса диктора, которые можно использовать для его идентификации.

В качестве модели вокализованного звука часто используется следующее представление [3, 8]:

$$x(t) = \sum_{i=1}^{m} A_i \cos(\omega_i t + \varphi_i), \quad (1)$$

где $A_i, \omega_i, \varphi_i$ — амплитуда, частота и фаза $i$-го обертона соответственно.

Использование представления (1) при синтезе речи приводит к неестественности её звучания, поэтому в качестве модели сигнала выбрана другая модель

$$x(t) = \int d\rho \sum_{i=1}^{m} K(\omega_i, \rho) \cos(\rho t + \varphi(\rho)), \quad (2)$$

где в качестве ядра модели выбрано выражение

$$K(\omega_i, \tau_i, \rho) = A_i \exp\left(-\frac{(\omega_i - \rho)^2}{\tau_i^2}\right). \quad (3)$$

Такая модель сигнала позволяет нам ввести понятие ширины $i$-го обертона линейчатого спектра $\tau_i$, которая всегда наблюдается при измерении спектра реального речевого сигнала. Причины возникновения ширины обертона могут быть различные: с одной стороны, это может быть изменение частоты обертона за время проведения измерений, а с другой стороны, длительность измерений может быть не кратна периоду обертона. Использование представления (2)–(3) для аппроксимации вокализованных звуков приводит к естественному звучанию при их синтезе.

Очевидно, что представление (2)–(3) при условии, что всё множество ширин обертонов стремится к нулю $\{\tau_i\} \to 0$, асимптотически сходится к представлению (1).

Для анализа спектра воспользуемся вейвлет-преобразованием в частотной области

$$L(\omega, \tau, \{\omega_i, \tau_i\}) = \sum_{i=1}^{m} \theta(\omega, \tau, \omega_i, \tau_i) = \sum_{i=1}^{m} \int_0^{\Omega} \psi(\rho, \omega, \tau) K(\rho, \omega_i, \tau_i) d\rho, \quad (4)$$

где $[0, \Omega]$ — диапазон спектра, $S(\rho, \{\omega_i\})$ — спектр Фурье вокализованного звука со своим множеством обертонов, $\psi(\rho, \omega, \tau)$ — вейвлет-функция, в качестве которой выбрана модифицированная «мексиканская шляпа» (см. **рисунок 1**).

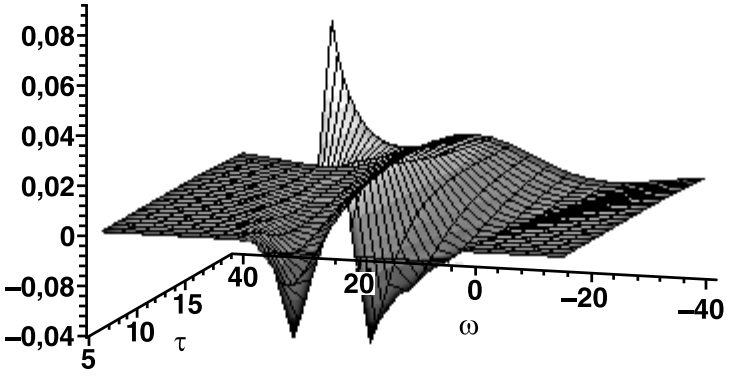


*Рис. 1. Модифицированная «мексиканская шляпа», рассчитанная при* $\rho = 0$

$$\psi(\rho,\omega,\tau)=\frac{1.031}{\sqrt{2}\tau^{3/2}}\exp\left(-\frac{(\omega-\rho)^2}{\tau^2}\right)\left(1-2\frac{(\omega-\rho)^2}{\tau^2}\right), \quad (5)$$

где $\tau$ — масштаб вейвлет-функции.

Численное интегрирование показывает, что если спектр состоит из одного обертона $\omega_1$, ширина которого $\tau_1$, то максимум интеграла

$$(\omega_1,\tau_1)=\max_{\omega,\tau}\theta(\omega,\tau,\omega_1,\tau_1),$$

соответствует положению и ширине обертона.

Этот вывод позволяет нам выполнить следующие операции:

в диапазоне частот спектра $[\Omega_1, \Omega_2]$ для каждого значения частоты и каждого значения ширины обертона вычислить значения интегралов $\theta(\omega,\tau)$ (диапазон значений ширин обертонов выбирается из следующих соображений: ширина обертона не может быть менее, чем четверть минимальной частоты основного тона, и не более, чем четверть максимальной частоты основного тона);

для каждого значения частоты основного тона $\omega_0$ рассчитать значение суммы

$$F(\omega_0)=\sum_{i=1}^{m}I(i)\theta(i\omega_0), \quad (6)$$

где индикаторная функция

$$I(i)=\begin{cases}1, \ if\ \ \theta(i\omega_0)>0\ \ and\ (\theta((i_0-1)\omega)>0\ \ or\ \ \theta((i+1)\omega_0)>0)\\ 0, \qquad else\end{cases},$$

применение которой позволяет избежать эффектов принятия за частоту основного тона единичных спектральных всплесков и найти величину:

$$\theta(i\omega_0)=\max_{\tau}\theta(i\omega_0,\tau);$$

разделить диапазон частот основного тона (90 Гц–450 Гц) на три непересекающиеся области (90 Гц–179 Гц, 180 Гц–359 Гц и 360 Гц–450 Гц) и для каждой области найти максимальное значение суммы (6) и частоту основного тона, которая доставила сумме этот максимум, т.е.:

$$F(\omega^*)=\max_{\omega_0}F(\omega_0),$$

$$\omega^*=\arg\max_{\omega_0}F(\omega_0).$$

Таким образом, результатом выполненных операций будет множество параметров $\{\omega_i^*, F(\omega_i^*),\ (\omega_i^*)\}_{i=1,\dots,3}$, вычисленное для каждого диапазона частоты основного тона в диапазоне спектра $[\Omega_1, \Omega_2]$.

Сделаем пояснения к изложенной последовательности операций. Известно, что методы выделения мгновенной частоты основного тона страдают ошибками её удвоения. Этот эффект возникает при условии, что амплитуды чётных обертонов многократно превосходят амплитуды нечётных обертонов основного тона [9]. Эффект можно ослабить, если разделить диапазон частот основного тона на три непересекающиеся области (границы которых были указаны выше) и проводить обработку речи в каждой из выделенных областей независимо. Поскольку эффект удвоения частоты неустойчив во времени, совместная обработка некоторой последовательности фреймов, каждый из которых обработан с учётом множества диапазонов частот основного тона, поможет снизить этот эффект. Области выбраны так, чтобы каждая из них не содержала удвоенных частот.

Пусть найдены множества параметров для последовательности из M фреймов. Введём меру того, что частота основного тона $\omega^*_{it}$, полученная в $i$-ом диапазоне $t$-ого фрейма, имеет своё продолжение в $j$-ом диапазоне $(t+1)$-ого фрейма:

$$\mu_{ijt} = \ln(F(\omega^*_{jt+1})) - \frac{(\omega^*_{it} - \omega^*_{jt+1})^2}{\sigma^2}, \quad (7)$$

где σ — допустимое изменение частоты основного тона от фрейма к фрейму (параметр модели). Тогда мера того, что за M фреймов частота основного тона совершила маршрут по диапазонам с известными номерами $\{i_1,..,i_t,..,i_M\}$ и будет равна

$$\beta\left(i_1,..,i_t,..,i_M\right) = \sum_{i=1}^{M-1} \mu_{i_t i_{t+1} t}. \quad (8)$$

Задача состоит в том, чтобы среди всех возможных траекторий найти траекторию с максимальной мерой (8). Для решения этой задачи можно использовать метод динамического программирования, подробно описанный в [10]. Заметим, что последовательность номеров диапазонов основного тона, которые определены методом динамического программирования для максимума меры (8) при условии, что значение этого максимума превышает некоторый порог $Q$, т.е.

$$\max_{i_1,..,i_t,..,i_M} \beta\left(i_1,..,i_t,..,i_M\right) > Q \quad (9)$$

определяет саму частоту основного тона и множество амплитуд её обертонов $\{\omega, \theta\}$. Если же максимум меньше порога, то принимается решение, что основного тона в данной последовательности фреймов нет.

Заметим, что метод (4)–(9) был развит для диапазона частот спектра, ограниченного значениями $[\Omega_1, \Omega_2]$. В зависимости от значений границ $[\Omega_1, \Omega_2]$ об одном и том же спектре могут быть приняты различные решения о частоте основного тона и различные решения о том, присутствует ли основной тон вообще в последовательности исследуемых фреймов. Это связано с тем, что некоторые вокализованные звуки содержат шум, обладающий сплошным спектром. Такое обстоятельство наводит на мысль о необходимости разбиения спектра на интервалы, в каждом из которых необходимо проводить независимые исследования на предмет линейчатости спектра.

Для разбиения полного спектрального диапазона телефонного канала [300 Гц, 3400 Гц] на три пересекающихся поддиапазона, используем понятие mel-шкалы

$$mel(f) = 1125 \ln\left(1 + \frac{\omega}{700}\right),$$

и mel-диапазона частот, который разбивается на $n$ равных интервалов, как это делается при MFCC-преобразовании [11] ($n$ соответствует количеству фильтров). Поддиапазоны выбираются так, чтобы в них входило целое количество интервалов. В данной работе $n$=26 и поддиапазоны имеют следующие границы: [300 Гц, 1489 Гц], [1378 Гц, 2296 Гц] и [2143 Гц, 3400 Гц].

Если в каждом из поддиапазонов провести операции (4)–(9) и для каждого принять решение, обладает он линейчатым спектром или сплошным, то в общей сложности можно получить $2^3$ различных типов спектров. Если в поддиапазоне спектр линейчатый, то его будет характеризовать частота основного тона и амплитуды её обертонов, а если спектр сплошной, то — значения свёрток спектра с банком фильтров, как это выполняется при вычислении MFCC [11]. Спектр как целое можно описывать следующим набором параметров: $\delta$ — бинарный вектор, $\delta \in R^3$, каждая из компонент которого соответствует решению о линейчатом или сплошном типе спектра в данном спектральном поддиапазоне (1 — спектр линейчатый, 0 — спектр сплошной); $\omega_0$ — частота основного тона; $c$ — вектор компонент дискретного mel-косинусного преобразования вида

$$c_n = \sum_{i=1}^{m} g_i \cos\left(\frac{\pi n}{mel(\Omega)}(mel(\omega_i)+0.5)\right), \qquad (10)$$

где $\Omega$ — правая граница спектрального диапазона телефонного канала, $g_i$ — амплитуда $i$-го обертона частоты основного тона с частотой $i\omega_0$ в случае, если для данной области принято решение $\delta = 1$, и компонента вектора MFCC с частотой максимума полосы пропускания фильтра $\omega_1$, если решение $\delta = 0$.

На **рисунке 2** показан спектр вокализованного звука, серые продольные линии задают размах выбранных поддиапазонов спектра, под линиями указаны

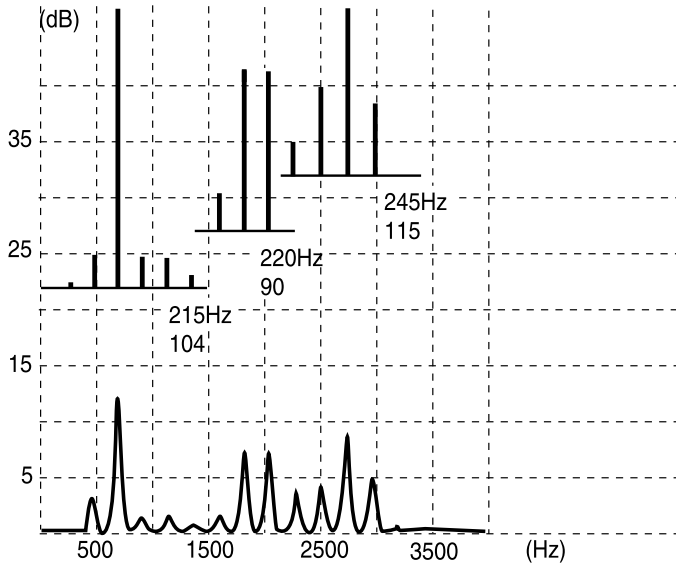


*Рис. 2. Спектр вокализованного звука и пример результата его предварительной обработки*

значения частот основного тона и значения интегралов (6), вычисленных для поддиапазонов. Если значение интеграла более 30, то принимается, что в поддиапазоне присутствует линейчатый спектр.

В ходе исследований было замечено, что частоты основного тона, определённые в различных поддиапазонах спектра, могут существенно различаться. Этот эффект был проверен как в телефонном канале, так и на микрофонных записях. Если в первом и втором поддиапазонах обнаружены частоты основного тона $\omega_1$ $\omega_2$ , то относительное смещение частоты основного тона

$$\lambda = \sum_{i=1}^{N} \frac{2(\omega_{2i} - \omega_{1i})}{\omega_{2i} + \omega_{1i}} ,$$

где $N$ — количество измерений, для дикторов, средние частоты основного тона которых находились в интервале от 100 до 150 Гц, составило 11,41%, в интервале от 150 до 200 Гц — 9,20%, в интервале от 200 до 250 Гц — 6,68%. Для каждого диапазона средних частот основного тона были отобраны 10 дикторов. Идеологически полученный эффект схож с эффектом смещения обертонов струн, который был описан с помощью уравнения маятника четвёртого порядка [12].

## Метод обучения и идентификации дикторов

Как было сказано ранее, всё множество векторов $C = \{c_t\}_T$ разделено на 8 различных типов с помощью соответствующих бинарных векторов $\delta_t$. Заметим, что вектор $\delta_t = \{0, 0, 0\}$ определяет не интересующий нас вектор $c_t$. Таким образом, остаётся семь различных типов данных. Построим статистическую модель для каждого типа в отдельности, полагая, что он описывается смесью плотностей нормальных распределений [13]:

$$p(c_t \mid \Theta) = \sum_{j=1}^{M} \beta_j N_j(c_t \mid \theta_j), \qquad (11)$$

где $Nj(c_t \mid \theta_j)$ — $j$-ая плотность нормального распределения, $\beta_j$ — априорная вероятность $j$-ой плотности, причём $\sum_{j=1}^{M} \beta_j = 1$. Параметры статистической модели каждого типа данных и параметры распределения частоты основного тона могут быть вычислены с помощью известного EM-алгоритма [13]. Априорная вероятность того или иного типа данных может быть оценена по формуле:

$$p(\delta) = \frac{W(\delta)}{T} ,$$

где $W(\delta)$ — частота встречаемости данных типа $\delta$.

В процессе идентификации для каждого диктора D и входных данных $X = \{\omega_t, \delta_t, c_t\}$ необходимо вычислить правдоподобие:

$$P(X \mid D) = \prod_{t=1}^{T} p(\omega_t \mid D) p(\delta_t \mid D) \sum_{i=1}^{M} \beta_j^{(\delta_t)} N(c_t \mid \theta_j^{(\delta_t)}) . \qquad (12)$$

Затем следует найти аргумент максимума (12) по всему списку дикторов, для которых вычислены параметры распределений.

## Эксперименты

Эксперименты были проведены на множестве фонограмм, записанных в компании «Стэл — Компьютерные Системы». Записи сделаны с помощью звонков с мобильного телефона в стандарте GSM на стационарный городской телефон, которые вводились в компьютер с помощью аппаратного комплекса «Ольха-9». Фонограммы содержали обучающий материал 58 дикторов (целевые дикторы), в дополнение к которым использовались 2-е фонограммы, содержащие смеси голосов 16 дикторов-женщин и 16 дикторов-мужчин, не включённых во множество целевых дикторов. Эти фонограммы смесей использовались для создания моделей неизвестных дикторов, таким образом решалась задача открытой идентификации. Длительность обучающего материала для целевых дикторов составляла от 30 до 40 секунд. Длительность для обучения моделей неизвестного диктора — около 160 секунд. Отношение сигнал/шум в фонограммах — от 12 до 15 дБ. Для тестирования использовались записи длительностью 30 секунд.

Классический метод, состоящий из последовательного применения MFCC и EM-алгоритма для модели гауссовых смесей (GMM) из 16 гауссоид, привёл к равной ошибке первого и второго рода 8.55%. Метод, состоящий из последовательного применения MCC и EM для GMM каждого из восьми состояний, которая включала в себя смесь из 4 гауссоид, привёл к равной ошибке первого и второго рода 6.11%.

## Заключение

Классические методы предварительной обработки речи (MFCC, PLP), связанные с интегрированием мощности сигнала по определённой области спектра, игнорируют понятие структуры сигнала, применяя одну и ту же форму обработки сигнала, независимо от его природы. Здесь под природой сигнала мы имеем в виду источник, порождающий этот сигнал (голосовой, шумовой и т. д.). С точки зрения авторов работы, единственный способ повышения точности систем обработки речи состоит в том, чтобы научиться анализировать структуру сигнала.

В рамках настоящей работы мы сделали попытку применить для задачи идентификации диктора один из способов анализа такой структуры и достигли пусть незначительного, но успеха.

Другой важный результат проделанной работы — регистрация смещения гармоник частоты основного тона в высокочастотную область. Этот результат может быть важен как с точки зрения анализа структуры сигнала, о которой речь шла выше, так и с точки зрения других направлений обработки речи, в частности её синтеза.

## Литература

1. *Sandipan Chakroborty* and Goutam Saha* Improved Text-Independent Speaker Identification using Fused MFCC & IMFCC Feature Sets based on Gaussian Filter//International Journal of Signal Processing 5;1 Winter 2009.

2. *Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani Md. Saifur Rahman* Speaker Identification Using Mel Frequency Cepstral Coefficients//*3rdICECE 2004, 28–30 December 2004, Dhaka, Bangladesh.*

3. *Hiromasa Fujihara, Tetsuro Kitahara, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno* Speaker Identification under Noisy Environments by Using Harmonic Structure Extraction and Reliable FrameWeighting//*INTERSPEECH 2006 — ICSLP* September 17–21, Pittsburgh, Pennsylvania, pp 1459–1462.

4. *Masataka Goto* A real-time music-scene-description system: predominant-F0 estimation for detecting melody and bass lines in real-world audio signals//Speech Communication 43 (2004) 311–329.

5. *Рамишвили Г.С.* Автоматическое опознавание говорящего по голосу. М.: Радио и связь, 1981.

6. *Котов М.А., Леднов Д.А. и др.* Способ определения параметров линейчатых спектров вокализованных звуков и система для его реализации № 2007148606/09 (053252) от 27.12.2007.

7. *Зыков А.П., Леднов Д.А., Меркулов М.Н.* Система голосовой идентификации диктора//Патент на полезную модель № 85445 от 10.08.2009.

8. *Moorer, J.A.*, Signal processing aspects of computer music: A survey//In Proceedings of the IEEE, Vol. 65, no. 8, 1108–1137, 1977.

9. *Babkin A.A.* LPC Speech Coder AT 1000–1200 BPS//In Proceedings of DSPA-2000.

10. *Моттль В.В., Мучник И.Б.* Скрытые Марковские модели в структурном анализе сигналов. М.: ФИЗМАТЛИТ, 1999.

11. *Steve Young, Gunnar Evermann, and Others.* The HTK Book. Cambridge University Engineering Department, HTK version 3.4 edition, December 2006.

12. N. H. Fletcher Inharmonicity, nonlinearity, and music//The Physicist v.37, n.5, September 2000.

13. *Jeff A. Bilmes* A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models//International Computer Science Institute Berkeley CA, April 1998.

***Леднов Дмитрий Анатольевич –***
*кандидат технических наук, старший научный сотрудник, руководитель отдела речевых технологий ООО «Стел – Компьютерные Системы», окончил Казахский государственный университет, г. Алма-Ата, защитил кандидатскую диссертацию в Ростовском государственном университете, г.Ростов-на-Дону.*

***Хацкевич Андрей Валентинович –***
*сотрудник отдела речевых технологий ООО «Стел – Компьютерные Системы», программист, окончил Ростовский государственный университет, г.Ростов-на-Дону.*

***Широкова Анна Михайловна –***
*сотрудник отдела речевых технологий ООО «Стел – Компьютерные Системы», лингвист, окончила Московский государственный университет*