Метод оценки формантных частот, основанный на полигармонической математической модели речевого сигнала

Голубинский А.Н.
Булгаков О.М.

Московская государственная юридическая академия имени О.Е. Кутафина.
Россия, 123995 Москва, Садовая-Кудринская, дом 9.
Тел. (499)244-85-24. Факс (499) 244-87-76. E-mail: galyashina@rambler.ru

Предложен метод оценки формантных частот для вокализованных участков речи. Вычислены точностные характеристики оценки формантных частот по методу, основанному на полигармонической модели речевого сигнала. Приведены результаты экспериментального расчёта оценок первых четырёх формантных частот предложенным методом.

Речевая наука и речевые технологии на сегодняшний день занимают уже достаточно ощутимое место в нашей жизни [1]. Для аутентификации (верификации и идентификации) личности по голосу необходимо осуществить параметризацию речевого сигнала. В качестве существенных параметров, отражающих индивидуальные особенности, уникальность голоса, часто используют формантные частоты (ФЧ). Под формантами понимают области глобальных спектральных максимумов речевого сигнала, характеризующие резонансные свойства речевого тракта как акустической системы [2,3]. Оценка ФЧ может проводиться с помощью кепстральных коэффициентов, на основе коэффициентов линейного предсказания, с использованием метода моментов спектра и другими способами. Все вышеперечисленные методы имеют определённые преимущества и недостатки, при этом для оценки ФЧ наиболее широкое распространение получил метод моментов и различные его модификации. К основным недостаткам метода моментов относят: несостоятельность оценок спектральных компонент (в этой связи выделяется целое направление обработки с помощью временных и спектральных окон), ошибки, связанные с эмпирическим подбором интервалов усреднения спектральных составляющих и эмпирическим определением компонентов вектора нормализующей функции. При этом, несмотря на некоторые успехи [4,5] в разработке алгоритмов расчета ФЧ, всё же остаётся ряд проблемных вопросов, связанных с неустойчивостью оценок ФЧ, которая появляется в зависимости от эмпирически подобранных интервалов усреднения спектральных составляющих и при различных априори заданных полосах частот для поиска глобального максимума среди множества локальных.

Заметим, что участки речи с большой степенью вокализации наиболее значимы для аутентификации личности по голосу. При этом импульсная модуляционная полигармоническая математическая модель речевого сигнала адекватно описывает вокализованные участки речи, в существенной мере определяющие индивидуальные особенности

голоса человека, и хорошо согласуется с физическими принципами акустической теории речеобразования [2]. Запишем модель речевого сигнала при отсутствии шумов в виде импульса АМ-колебания с несколькими несущими частотами для случая модуляции суммой гармоник, из практических соображений ограничившись конечным количеством гармоник ряда [6]:

$$u(t)=\sum_{k=0}^{K} M_k \cos(\,2\pi\, k F_0^{\text{МОД}}\, t+\Phi_k)\sum_{l=1}^{L} U_l \cos(\,2\pi\, l f_0\, t+\varphi_l)\,,\quad t\in[0;\tau_{\text{è}}]\,. \qquad (1)$$

Здесь M_k и $F_0^{\text{мод}}$ — соответственно глубина амплитудной модуляции k-й гармоники и наименьшая частота модулирующего колебания; U_l — амплитуда l-й гармоники несущего колебания; f_0 — частота основного тона (ЧОТ); $(K+1)$ и L — соответственно количество модулирующих и несущих гармоник. При этом математическая модель, записанная в виде (1) может применяться для параметрического описания как в рамках детерминированного (Φ_k и φ_l — некоторые константы), так и стохастического подходов (Φ_k и φ_l — случайные величины).

Суть метода оценки ФЧ, основанного на полигармонической модели (ПГМ) речевого сигнала состоит в следующем. Положим, что известны значения оценок $\widehat{f}_0$ и $\widehat{U}_l$, при этом количество амплитуд гармоник принимают равным: $L_{\max}=\lfloor f_{\text{â}}/\widehat{f}_0\rfloor$, где $\lfloor\ \rfloor$ — целая часть числа; $f_{\text{â}}=f_d/2$ — верхняя частота; f_d — частота дискретизации. По известным значениям U_l определяют их глобальный максимум $U_{l\max 1}$, при этом за оценку ФЧ принимают аргумент, соответствующий найденному глобальному максимуму: $\widehat{F}_1=l_{\max 1}\cdot\widehat{f}_0$, т.е. оценка соответствует резонансно усиленной $l_{\max 1}$-й гармонике основного тона, или же $(l_{\max 1}-1)$-му обертону. Далее находят первый минимум $U_{l\min 1}$, после значения аргумента которого $l_{\min 1}$, определяют следующий первый максимум $U_{l\max 2}$; оценка второй ФЧ: $\widehat{F}_2=l_{\max 2}\cdot\widehat{f}_0$. Заметим, что при необходимости можно сузить интервал поиска максимумов, используя диапазоны наиболее вероятного нахождения соответствующих ФЧ [2,3]. Далее находят следующий первый минимум $U_{l\min 2}$, после значения аргумента которого $l_{\min 2}$, ищут первый максимум $U_{l\max 3}$; оценка третьей ФЧ: $\widehat{F}_3=l_{\max 3}\cdot\widehat{f}_0$, и т.д.

Таким образом, вычисление оценки ФЧ и оценок амплитуд несущих гармоник на базе метода наименьших квадратов [6], должно опираться на достаточно точную оценку ЧОТ. Оценку ЧОТ предлагается получать на основе оптимальной обработки при использовании метода максимального правдоподобия (МП), который обладает высокой потенциальной и реальной точностью. Пусть детерминированный сигнал $u(t,f_0)$ принимается на фоне шума $n(t)$, при этом требуется оценить значение существенного параметра f_0, заключённого в сигнале $u(t,f_0)$, обрабатывая принятую реализацию случайного сигнала $\xi(t,f_0)$:

$$\xi(t,f_0)=u(t,f_0)+n(t)\,, \qquad (2)$$

где $u(t,f_0)$ — детерминированная компонента (1); $n(t)$ — шумовая компонента в виде модели гауссовского δ-коррелированного случайного процесса с нулевым средним значением и функцией корреляции: $R(t_1,t_2)=N_0\delta(t_1-t_2)/2$; N_0 — односторонняя спектральная плотность мощности шума.

При наблюдении сигнала (2) на фоне гауссовского шума $n(t)$ полученный логарифм функционала отношения правдоподобия (ЛФОП) $M(f)$ позволяет

синтезировать приемник МП, а оценка МП параметра f определяется как значение аргумента, при котором наблюдается глобальный максимум ЛФОП [7]:

$$\hat{f}_0 = \arg\sup M(f). \quad (3)$$

В ходе исследований было выявлено, что полигармоническое модулирующее колебание, входящее в $u(t, f_0)$, практически не оказывает влияния на точность оценки ЧОТ. Таким образом, для оценки ЧОТ будем использовать ПГМ без учёта модуляции:

$$u(t, f_0) = \sum_{l=1}^{L} U_l \cos(2\pi l f_0 t + \varphi_l) = \sum_{l=1}^{L} \{x_l \cos(2\pi l f_0 t) + y_l \sin(2\pi l f_0 t), \quad (4)$$

где $x_l = U_l \cos(\theta_l)$; $y_l = U_l \sin(\theta_l)$; $\theta_l = -\varphi_l$. Однако, при оценке ЧОТ, как правило, трудно получить априорную информацию о распределении амплитуд U_l и начальных фаз φ_l гармоник, образующих сложный полигармонический сигнал (4). В этой связи для оценки ЧОТ рассмотрим модель речевого сигнала, где неизвестны U_l и φ_l всех гармоник, т.е. неизвестны x_l и y_l в модели (4). Осуществляя максимизацию ЛФОП M(*f*) по неизвестным несущественным параметрам x_l и y_l можно показать [8], что в итоге ЛФОП трансформируется, принимая для разрешаемых источников следующий вид:

$$M(f) = \frac{N_0}{2T}\left[\sum_{l=1}^{L}\left(\frac{2}{N_0}\int_0^T \xi(t, f_0)\cos(2\pi l f t)\,\mathrm{d}t\right)^2 + \sum_{l=1}^{L}\left(\frac{2}{N_0}\int_0^T \xi(t, f_0)\sin(2\pi l f t)\,\mathrm{d}t\right)^2\right]. \quad (5)$$

Как видно из (5) оптимальная обработка полигармонического сигнала (4), с целью оценки его ЧОТ, сводится к формированию билинейной формы из квадратурных компонент корреляционного интеграла. При этом для разрешаемых источников должно выполняться соотношение [8]:

$$\Psi_{ij}(f_0, f_0) < 1; \quad i \neq j; \quad i, j = \overline{1, L}, \quad (6)$$

где модуль нормированной взаимной функции неопределенности каждой пары источников сигнала (гармоник, образующих полигармонический сигнал) рассчитывается для модели вида (4) как:

$$\Psi_{ij}(f_1, f_2) = \frac{\sin[\pi(j f_2 - i f_1)T]}{\pi(j f_2 - i f_1)T} = \mathrm{sinc}[\pi(j f_2 - i f_1)T], \quad (7)$$

а в точке истинного значения принимает вид:

$$\Psi_{ij}(f_0, f_0) = \mathrm{sinc}[\pi(j-i) f_0 T]. \quad (8)$$

Так как период основного тона наиболее вероятно находится в диапазоне: $T_0 \in [3;14;3]$мс [3], то, выбирая интервал наблюдения T>38,5мс всегда можно обеспечить выполнение условия разрешения гармоник (6) для сигнала $u(t, f_0)$ вида (4).

Заметим, что при наблюдении дискретного сигнала $u_i(f_0) \equiv u(i\Delta, f_0)$ интегралы в компонентах ЛФОП (5) заменяются на соответствующие суммы, а в опорных сигналах непрерывное время t заменяется на дискретное $t_i = i\Delta$, где $\Delta = 1/f_d$ — интервал дискретизации.

Выражение для условной дисперсии при неизвестном априорном распределении f_0 и U_l, при условии разрешения гармоник в случае высокой апостериорной точности [8], с учетом (7) имеет вид:

$$D(\hat{f}_0 \mid f_0) = \frac{1}{z^2}\left[\sum_{i=1}^{L} U_{0i}^2 \left(\sum_{i=1}^{L} U_{0i}^2\right)^{-1} \frac{\partial^2 \Psi_{ii}(f_1, f_2)}{\partial f_1 \partial f_2}\bigg|_{f_1 = f_2 = f_0}\right]^{-1} = \frac{3}{\pi^2 z^2 T^2}\sum_{l=1}^{L} U_l^2 \left[\sum_{l=1}^{L} U_l^2 l^2\right]^{-1}, \quad (9)$$

где U_{0i} — истинные значения амплитуд; Z^2 — отношение сигнал-шум по мощности.

Теперь вычислим оценки амплитуд гармоник математической модели (1). Пусть задано J существенных отсчетов автокорреляционной функции (АКФ) K_j, вычисленных по эк-

сперимента́льным данным речевого сигнала. Также известна АКФ $Ka(\tau)$ математической модели речевого сигнала, которая задана в следующем виде:

$$K_a(\tau) = \sum_{l=1}^{L} U_l^2 S_l(\tau), \quad (10)$$

где U_l — амплитуды несущих гармоник в математической модели; $S_l(\tau)$ — функция, которая зависит от номера несущей гармоники l, интервала времени τ, а также от параметров математической модели (таких как f_0, $F_0^{\text{мод}}$, $\tau_{\text{и}}$ и др. [6]). Выражение (10) в матричной форме:

$$\mathbf{Ka} = \mathbf{S}\,\mathbf{V}, \quad (11)$$

где $\mathbf{Ka}$ — матрица-столбец размером J x 1 с элементами $Ka_j = Ka(j\Delta)$; $\mathbf{S}$ — прямоугольная матрица J x L с элементами $Ka_j = Ka(j\Delta)$; $\mathbf{S}$ — матрица-столбец размером J x L с элементами $S_{jl} = S_l(j\Delta)$. Ошибку модели (1) относительно экспериментальных данных в матричной форме определим как:

$$\boldsymbol{\varepsilon}(\mathbf{V}) = (\mathbf{Ka} - \mathbf{K})^{\mathrm{T}}(\mathbf{Ka} - \mathbf{K}) = (\mathbf{K} - \mathbf{SV})^{\mathrm{T}}(\mathbf{K} - \mathbf{SV}), \quad (12)$$

где $\mathbf{K}$ — матрица-столбец размером J x 1 с элементами K_j; T — знак транспонирования. Необходимое условие обращения (12) в минимум:

$$\partial\boldsymbol{\varepsilon}(\mathbf{V})/\partial\,\mathbf{V} = 0. \quad (13)$$

Положим, что для аутентификации личности по голосу $Ka_0 = \text{const}$, при данном допущении систему нелинейных уравнений (13), состоящую из полиномов четвертой степени, можно свести к линейной, а её решение в виде вектора параметров V_l имеет вид:

$$\mathbf{V} = (\mathbf{S}^{\mathrm{T}}\mathbf{S})^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{K}. \quad (14)$$

Таким образом, решение системы (13) относительно оценок параметров математической модели $\hat{U}_l$:

$$\hat{U}_l = \sqrt{V_l},\ l = \overline{1,L}. \quad (15)$$

Определим характеристики оценок ФЧ, полученных по методу, основанному на ПГМ. Заметим, что грубый потенциальный промах для оценки первой ФЧ $\hat{F}_1$ на практике весьма маловероятен (ввиду относительно большого различия между соседними U_l в этой области), поэтому относительная ошибка оценки частоты первой форманты:

$$|\delta\hat{F}_1^{\text{ош}}| = |l_{\max 1}\cdot\hat{f}_0 - l_{\max 1}\cdot f_0|/(l_{\max 1}\cdot f_0)\cdot 100\% \approx 3\cdot|\delta f_0|, \quad (16)$$

где $|\delta f_0| = \sigma_{f_0}/\hat{f}_0\cdot 100\%$ — относительная ошибка оценки ЧОТ; $\sigma_{f_0} = \sqrt{D(\hat{f}_0 \mid f_0)}$ — среднеквадратическое отклонение; границы доверительного интервала (для доверительной вероятности P = 99,7%) соответствуют величине $\pm 3\sigma_{f_0}$. В качестве характеристики оценки второй ФЧ и выше будем использовать усреднённую относительную ошибку, соответствующую грубым потенциальным промахам при принятии за оценку ФЧ $\hat{F}_q = l_{\max q}\cdot\hat{f}_0$ соседнего правого или левого обертона, которая в итоге имеет вид:

$$|\delta\hat{F}_q^{\text{пр}}| = |l_{\max q}(1\pm 3\cdot|\delta f_0|)|/(l_{\max q}^2 - 1)\cdot 100\%,\ q = 2,3,\ldots \quad (17)$$

В таблице 1 приведены значения ФЧ для гласных звуков, полученные методом, основанным на ПГМ $\hat{F}_1^{ПГМ}$ и методом моментов $\hat{F}_1^{ММ}$, также здесь указаны относительные рассогласования оценок между этими методами: $|\delta\hat{F}_q^{3}|$ =

$= (|\widehat{F}_q^{ПГМ} - \widehat{F}_q^{ММ}| / \widehat{F}_q^{ММ}) \cdot 100\%$. Экспериментальный расчёт оценок ФЧ на основе математической модели (1) проводился при параметрах: $f_d = 8$кГц; $J = 300$; количество отсчётов звука $N = 3500$ ($\tau_и = N\Delta$); $K = 1$; $F_0^{МОД} = 10$ Гц; $M_0 = = 1$; $\Phi_0 = 0$; $M_1 = 1$; $\Phi_0 = \pi$; $L_{max} = 26 \div 27$, при этом для оценки ЧОТ: $L = 3$; $T = \tau_и$; $Z^2 = 10$.

Как видно из таблицы 1, оценки ФЧ, полученные двумя методами, характеризуются близкими значениями — относительные рассогласования $|\delta\widehat{F}_q^i|$, за исключением $\widehat{F}_1$ звука /о/ (рассогласование 11,564% обусловлено примерным равенством амплитуд соседних обертонов в окрестности F_3) не превышали 5,3%. Это даёт основание полагать, что метод оценки ФЧ при использовании ПГМ имеет удовлетворительные точностные характеристики. При этом разработанный метод даёт конструктивный подход к вычислению оценок ФЧ в рамках математической модели, записанной в явном виде, а также лишён ряда недостатков, которые присущи методу моментов и его модификациям.

Таблица 1

Гласные	$\widehat{f}_0$, Гц	$\widehat{F}_1^{ПГМ}$, Гц	$\widehat{F}_1^{ММ}$, Гц	$\lvert\delta\widehat{F}_1\rvert$, %	$\lvert\delta\widehat{F}_1^{оц}\rvert$, %	$\widehat{F}_2^{ПГМ}$, Гц	$\widehat{F}_2^{ММ}$, Гц	$\lvert\delta\widehat{F}_2\rvert$, %	$\lvert\delta\widehat{F}_2^{пр}\rvert$, %
о	152,3	456,9	458,1	0,24	0,363	913,8	912,2	0,197	17,163
у	149,8	149,8	149,2	0,537	0,369	449,4	450,9	0,355	37,510
а	144,6	723,0	725,4	0,276	0,383	1156,8	1160,8	0,362	12,704
э	144,3	577,2	581,9	0,825	0,384	1731,6	1732,8	0,081	8,404
и	153,6	153,6	154,9	0,903	0,360	2304,0	2189,4	5,254	6,707
ы	152,6	305,2	307,8	1,548	0,363	1526,0	1550,7	1,612	10,155

Гласные	$\widehat{F}_3^{ПГМ}$, Гц	$\widehat{F}_3^{ММ}$, Гц	$\lvert\delta\widehat{F}_3\rvert$, %	$\lvert\delta\widehat{F}_3^{пр}\rvert$, %	$\widehat{F}_4^{ПГМ}$, Гц	$\widehat{F}_4^{ММ}$, Гц	$\lvert\delta\widehat{F}_2\rvert$, %	$\lvert\delta\widehat{F}_2^{пр}\rvert$, %
о	2132,2	2410,8	11,564	7,188	3198,3	3173,4	0,797	4,778
у	2546,6	2588,7	1,638	5,904	3295,6	3224,2	2,221	4,556
а	2458,2	2468,1	0,397	5,905	3325,8	3323,0	0,084	4,358
э	2164,5	2169,1	0,207	6,706	3030,3	3040,1	0,319	4,780
и	3072,0	2971,2	3,400	5,020	3532,8	3594,7	1,730	4,363
ы	2289,0	2304,9	0,694	6,706	3509,8	3391,5	3,503	4,362

Полученные оценки ФЧ могут быть использованы в качестве параметров для аутентификации личности по голосу [5,9].

ЛИТЕРАТУРА

1. *Сорокин В.Н.* Фундаментальные исследования речи и прикладные задачи речевых технологий // Речевые технологии. 2008. № 1. С. 18–48.

2. *Фант Г.* Акустическая теория речеобразования. М.: Наука, 1964. 284 с.

3. *Сапожков М.А.* Речевой сигнал в кибернетике и связи. М.: Связьиздат, 1963. 452 с.

4. *Рабинер Л.Р., Шафер Р.В.* Цифровая обработка речевых сигналов. М.: Радио и связь, 1981. 496 с.

5. *Лабутин П.В., Раев А.Н., Коваль С.Л.* Патент на изобретение № 2230375 РФ: МПК 7 G10L15/00, G10L17/00. Метод распознавания диктора и устройство для его осуществления. № 2002123509/09; заявл. 03.09.02; опубл. 10.06.04.

6. *Голубинский А.Н.* Обработка речевого сигнала на основе модели в виде импульса АМ-колебания с несколькими несущими частотами // Телекоммуникации. 2008. № 12. С. 13–17.

7. *Куликов Е.И., Трифонов А.П.* Оценка параметров сигналов на фоне помех. М.: Сов. радио, 1978. 296 с.

8. *Лукин А.Н.* Радиофизические методы измерения параметров сложных источников излучения: дис. докт. физ.-мат. наук: 01.04.03. Воронеж, 1998. 415 с.

9. *Голубинский А.Н., Булгаков О.М.* Аутентификация личности по вокализованным участкам речи на основе частоты основного тона и амплитуд кратных гармоник в области первых двух формант // Системы управления и информационные технологии. 2009. № 4.1. С. 134–139.