УДК 621.391

В.В. Поповский[1], С.М. Бобрицкий[2], В.Е. Саваневич[1], В.Н. Ткачев[1]

[1]*Харьковский национальный университет радиоэлектроники, Харьков*
[2]*Харьковский институт судебных экспертиз им. засл. проф. Н.С. Бокариуса, Харьков*

# ОПРЕДЕЛЕНИЕ ЧАСТОТ ФОРМАНТ ГОЛОСОВОГО СИГНАЛА НА ВЫХОДЕ ГРЕБЕНКИ ПОЛОСОВЫХ ФИЛЬТРОВ

*Разработан метод определения частот формант голосового сигнала на выходе гребенки полосовых фильтров, позволяющий улучшить точность пофонемного кратковременного спектрального анализа речевых фрагментов, передаваемых в пакетных сетях с потерями. Для этого для получения соответствующей системы уравнений максимального правдоподобия использован математический аппарат группированных выборок, позволяющий учесть факт дискретного, в частотном диапазоне, представления энергии голосового сигнала при оценке средних частот формант речевого сигнала.*

***Ключевые слова:*** *форманты речевых сигналов, пофонемный кратковременный спектральный анализ речевых сигналов, группированные выборки, локальное математическое ожидание.*

## Введение

В современных информационных и телекоммуканикационных системах значительное место занимают речевые технологии [1]. В рамках последних значимая роль отведена таким задачам идентификации речи как идентификация диктора и его эмоционального состояния по фрагментам речи, представленным цифровыми средствами записи, в том числе переданным по пакетным сетям с потерями [2 – 4].

**Анализ публикаций.** Акустическое качество звуков определяется общей формой спектра и соотношением уровней сигнала в полосах спектра [5]. При этом форманты (максимумы в спектре речевого сигнала) являются доступным для речеобразующего аппарата способом достижения необходимых полосных соотношений и могут быть использованы как устойчивые идентификационные признаки. К числу таких устойчивых признаков относятся и формантные соотношения – $F_2/F_1$, $F_3/F_1$, $F_3/F_2$, являющиеся идентификационными признаками, более устойчивыми к изменению эмоциональных ситуаций (в том числе снижение темпа речи), чем средние значения формант.

Согласно общей методике формантный анализ проводится в рамках статистического спектрального анализа [6] для отрезков речевого сигнала с предварительным выделением микрофрагментов с четко выраженной форматной структурой спектра. Часто такими фрагментами являются гласные фонемы.

Для автоматического исследования речевых фрагментов используется также метод, основанный на построении гауссовых статистических моделей параметрического представления речевого сигнала [7]. Сравнение таких моделей позволяет определить тождество дикторов с точки зрения совпадения статистик представленности в их речи звуков разных типов (с точки зрения их спектральной реализации).

Данные методы позволяют приблизиться к достаточно устойчивой автоматической идентификации личности по голосу. Хуже дела обстоят при дополнительном условии передачи используемых голосовых фрагментов современными пакетными сетями с потерями и спектральными искажениями голоса низкоскоростными кодеками. Резко сокращающаяся избыточность данных, передающих голосовые фрагменты, приводит к необходимости минимизировать потери информации на этапе формирования идентификационных признаков. Это возможно, в том числе, путем учета особенности прохождения речевых сигналов через различные интегрирующие системы. К таким системам относятся АЦП и полосовые фильтры.

**Постановка задачи.** Накопленный специалистами научный и экспертный материал позволяют утверждать, что спектр слитной человеческой речи в любой момент времени и на любом локальном интервале в процессе произношения фонем может быть представлен совокупностью гауссоид, каждая из которых представляет собой отдельную форманту. Количество таких формант Q может составлять различную величину. Но в качестве идентификационного признака достаточно, обычно, четырех (Q = 4). Каждая гауссоида представляется математическим ожиданием $f_j$ и средние квадратические отклонения (СКО) $\sigma_j$. Шумы имеют равномерный спектр в исследуемом диапазоне. Следовательно, экспериментальный спектр на любом временном срезе представим смесью вероятностных распределений:

$$W(f) = p_0 + \sum_{j=1}^{Q} \frac{p_j}{\sqrt{2\pi}\sigma_j} \exp\left\{ -\frac{1}{2\sigma_j^2}(f - f_j)^2 \right\}, \quad (1)$$

где $p_0$ – относительный вес шумовой составляющей речевого сигнала; $p_j$ – относительный вес энергии речевого сигнала, соответствующего j-й форманте.

При этом регистрации доступны значения энергий речевого сигнала на выходах полосовых фильтров (в частотных каналах) в заданном временном окне, которые легко можно привести к относительным значениям энергии речевого сигнала в k-м частотном канале t-го временного окна $\upsilon^*_{kt}$. Тем самым имеет место группированная выборка [9]. Теоретическим аналогом опытных относительных значений энергии речевого сигнала являются вероятности попадания речевого сигнала в рассматриваемые полосовые фильтры с частотными границами $f_{нк}$, $f_{кк}$ t-го временного окна:

$$\upsilon_{kt}(\Theta) = \int_{f_{нк}}^{f_{кк}} W(f)df .$$

Опытные энергии речевого сигнала в частотных каналах одного временного окна и в разных окнах независимы.

Итак, необходимо на основе совокупности значений $\upsilon^*_{kt}$ и принятой модели мгновенного спектра речевого сигнала синтезировать процедуру максимально правдоподобной оценки частот формант речевого сигнала $f_j$ ($j = \overline{1,Q}$ $Q = 3$ либо $Q = 4$). Также в множество оцениваемых параметров могут быть внесены СКО формант $\sigma_j$ ($j = \overline{1,Q}$).

## Основной материал

**Решающее правило оценки частот формант голосового сигнала по данным гребенки полосовых фильтров на t-м временном окне.** Для максимизации точности оценки частот формант исследуемого временного окна (исследуемой серии временных окон) необходимо учесть факт дискретности пространства наблюдения (оцениваются непрерывные величины – частоты формант, по дискретным – значениям энергии речевого сигнала на выходах полосовых фильтров).

Если предположить, что вероятности попадания речевого сигнала в рассматриваемые полосовые фильтры не равны нулю, либо, что полосовые фильтры с нулевыми значениями данных вероятностей исключены из рассмотрения, то общий вид уравнений, входящих в систему уравнений максимального правдоподобия принимает вид [10]:

$$\sum_{k,t}^{N_{чк}} \frac{\upsilon^*_{kt}}{\upsilon_{kt}(\Theta)} \frac{\partial \upsilon_{kt}(\Theta)}{\partial \theta_m} = 0 , \qquad (2)$$

где $\theta_m$ – m-й оцениваемый параметр.

Только одно слагаемое (1) зависит от частоты j-ой форманты голосового сигнала. Поэтому, помня о том, что $\exp'(x) = x'\exp(x)$, выражение для производной по частоте j-ой форманты заданного временного окна может быть записано следующим образом:

$$\frac{d\upsilon_{kt}(\Theta)}{df_j} = \frac{p_j}{\sqrt{2\pi}\sigma_j} \int_{f_{нк}}^{f_{кк}} \exp\left(-\frac{1}{2\sigma_j^2}(f-f_j)^2\right)\frac{f-f_j}{\sigma_j^2}df =$$

$$= p_j(N_{f_{кк}}(f_j;\sigma_j^2) - N_{f_{нк}}(f_j;\sigma_j^2)) , \qquad (3)$$

где $N_z(m_z;\sigma^2) = \frac{1}{\sqrt{2\pi\sigma}}\exp\left(-\frac{1}{2\sigma^2}(z-m_z)^2\right)$.

Для дальнейшего преобразования выражения (3) целесообразно воспользоваться понятием [11] локального математического ожидания случайной величины на интервале $[f_{нк}, f_{кк}]$:

$$m^{лок}_{f_k} = \frac{1}{F_{fk}(m_f;\sigma^2)} \int_{f_{нк}}^{f_{ki}} fN_f(m_f;\sigma^2)df = m_f +$$

$$+\frac{\sigma^2}{F_{fk}(m_f;\sigma^2)}(N_{f_{ki}}(m_f;\sigma^2) - N_{f_{нк}}(m_f;\sigma^2)) .$$

При этом выражение (4) примет вид:

$$\frac{d\upsilon_{kt}(\Theta)}{df_j} = p_j(N_{f_{кк}}(f_j;\sigma_j^2) - N_{f_{нк}}(f_j;\sigma_j^2)) =$$

$$p_j\frac{F_{fk}(f_j;\sigma_j^2)}{\sigma_j^2}(m^{лок}_{f_k} - f_j) , \qquad (4)$$

где $F_{fk}(m_f;\sigma^2) = \int_{f_{кк}}^{f_{нк}} N_f(m_f;\sigma^2)df$.

При оценке СКО формант голосового сигнала необходимо записать выражение для соответствующей производной:

$$\frac{d\upsilon_{kt}(\Theta)}{d\sigma_j} = -\frac{p_j}{\sigma_j}F_{fk}(f_j;\sigma_j^2) +$$

$$+\frac{p_j}{\sqrt{2\pi}\sigma_j}\int_{f_{нк}}^{f_{кк}} \exp(-\frac{1}{2\sigma_j^2}(f-f_j)^2)\frac{(f-f_j)^2}{\sigma_j^3}df = \qquad (5)$$

$$= \frac{p_j}{\sigma_j}\left[(f_{нк}-f_j)N_{f_{нк}}(f_j;\sigma_j^2) + (f_j - f_{кк\,j})N_{f_{кк}}(f_j;\sigma_j^2)\right] .$$

С учетом выражений для производных (4) и (5) система уравнений максимального правдоподобия (2) состоит из 2Q уравнений вида:

$$\sum_{k,t}^{N_{чк}} \upsilon^*_{kt}\lambda_{jkt}(m^{лок}_{f_k} - f_j) = 0 ; \qquad (6)$$

$$\sum_{k,t}^{N_{чк}} \upsilon^*_{kt}\lambda_{jkt}\left[(f_{нк}-f_j)N_{f_{нк}}(f_j;\sigma_j^2) + (f_j - f_{кк\,j})N_{f_{кк}}(f_j;\sigma_j^2)\right] /$$

$$/F_{fk}(f_j;\sigma_j^2) = 0 , \qquad (7)$$

где $\lambda_{jkt} = p_j F_{fk}(f_j;\sigma_j^2) / \upsilon_{kt}(\Theta)$. (8)

С учетом принятых обозначений, выражение для теоретической вероятности попадания энергии речевого сигнала в рассматриваемые полосовые фильтры принимает вид:

$$\upsilon_{kt}(\Theta) = p_0(f_{kk} - f_{нк}) + \sum_{j=1}^{Q} p_j F_{fk}(f_j; \sigma_j^2).$$

Система уравнений максимального правдоподобия (6) – (7) является трансцендентной. Для ее решения был выбран метод последовательных приближений [12], при реализации которого система уравнений была представлена в виде Q пар уравнений вида:

$$\hat{f}_j = \frac{\sum_{k,t}^{N_{чк}} \upsilon_{kt}^* \lambda_{jkt} m_{f_k}^{лок}}{\sum_{k,t}^{N_{чк}} \upsilon_{kt}^* \lambda_{jkt}}; \qquad (9)$$

$$\hat{\sigma}_j = \frac{\sum_{k,t}^{N_{чк}} \upsilon_{kt}^* \lambda_{jkt} (m_{f_k}^{лок} - \hat{f}_j)^2}{\sum_{k,t}^{N_{чк}} \upsilon_{kt}^* \lambda_{jkt}}. \qquad (10)$$

Выражения для оценки относительных весов имеют вид [10]:

$$\hat{p}_j = \frac{1}{N_{чк}} \sum_{k,t}^{N_{чк}} \lambda_{jkt}. \qquad (11)$$

Выражение (11) является стандартной оценкой весов дискретного смешивающего распределения смеси вероятностных распределений [10].

**Сущность метода оценки частот и разброса Q формант речевого сигнала на совокупности временных окон, сформированных гребенкой полосовых фильтров.** При получении оценок частот Q формант поочередно выполняются две операции. Первая – расщепление статистик гребенки полосовых фильтров исследуемой совокупности временных окон на статистики (Q + 1)-й гребенки. В результате данной операции в соответствии с начальными приближениями по значению параметров $\Theta$, полученными на предыдущей итерации (на первой итерации – начальные приближения, поступившие на вход процедуры), составляющие энергии голосового сигнала, попавшие в каждый полосовой фильтр гребенки, распределяются на Q независимых процедур оценки частоты форманты. При этом отделяются, в соответствии с принятой моделью речевого сигнала и шумов, шумовые составляющие, источником которых не является голосовой сигнал, параметры которого подлежат оценке. Результатом первой операции является совокупность коэффициентов расщепления $\lambda_{jkt}$.

Операция, следующая за расщеплением, – взаимно независимая оценка параметров каждой из Q формант. При этом используется Q процедур оценки, на каждую из которых поступает подвыборка – один из Q + 1 результатов расщепления гребенки полосовых фильтров исследуемой совокупности временных окон, полученных на итерации. Каждая процедура оценки на n–й итерации по детерминированному правилу (9) – (10) формирует оценку параметров соответствующей форманты $\hat{f}_j$.

Сформированные оценки поступают в качестве начального приближения на процедуру расщепления и так далее. Итерационный процесс продолжается до тех пор, пока либо все коэффициенты расщепления $\lambda_{jkt}$ на n–м и (n – 1)–м шаге не станут практически попарно равны, либо оценки $\hat{\Theta}_{Q(n-1)}$ и $\hat{\Theta}_{Qn}$ практически совпадут попарно: $\left|\hat{\theta}_{js(n-1)} - \hat{\theta}_{jsn}\right| < \varepsilon_s$, где $\varepsilon_s$ – константа, определяющая требуемую точность вычислений.

Таким образом, выражения (9) – (10) сводят задачу определения параметров Q формант к совокупности из Q независимых задач определения параметров одной форманты с предшествующим расщеплением энергий сигналов на выходе гребенки полосовых фильтров путем вычисления коэффициентов $\lambda_{jkt}$ (8).

**Результаты проведения эксперимента.** В рамках кратковременного (пофонемного) спектрального анализа на эталонных сигналах было проведено исследование точности определения формант речевых фрагментов. Для этого в пакете MATLAB была разработана соответствующая вычислительная процедура. Данная процедура была исследована на большем количестве дикторов при пребывании их в различных эмоциональных состояниях. Для этого небольшие по длительности участки речи указанных дикторов были предварительно записаны в wav-файлы. Было установлено, что в худшем случае СКО определения частоты форманты не превышает нескольких десятков Гц при исследовании гребенки из 21 цифрового полосового фильтра с периодом съема данных 5,7 мс. При этом была установлена устойчивость метода к импульсным воздействиям.

## Выводы

В статье разработан итерационный метод оценки частот формант исследуемого временного окна (исследуемой серии временных окон) голосового сигнала на выходе гребенки полосовых фильтров. Для максимизации точности оценки частот формант метод учитывает сам факт оценки непрерывных параметров по дискретному пространству наблюдений (оцениваются непрерывные величины – частоты формант, по дискретным – значениям энергии речевого сигнала выходах счетного числа полосовых

фильтров). Данный учет стал возможен благодаря использованию математического аппарата группированных выборок для описания энергии речевого сигнала на выходах полосовых фильтров, соответствующей как голосовому сигналу так и сигналам и помехам. При этом энергия форманты, соответствующая полосовому фильтру вместо того, чтобы приписываться средней частоте полосы фильтра, согласно полученных аналитических выражений системы уравнений максимального правдоподобия, стала приписываться локальному математическому ожиданию частоты энергии форманты в диапазоне соответствующего полосового фильтра.

Путем экспериментальных исследований на натурных данных подтверждена работоспособность метода, определены его предварительные точностные характеристики. Практическая значимость метода заключается в возможности его использования в системах автоматической идентификации личности и ее эмоционального состояния по голосу. Дальнейшие исследования целесообразно сконцентрировать на разработке методов объединения различных идентификационных признаков с целью создания указанной системы с приемлемыми показателями качества автоматической идентификации.

## Список литературы

*1. Олифер В.Г. Компьютерные сети. Принципы, технологии, протоколы. / В.Г. Олифер Н.А. Олифер. – М.: Питер, 2007. – 520 с.*

*2. Актуальные вопросы идентификации личности // Материалы научно-практической конференции 17 декабря 1998 г. – СПб., 1999. – С. 39-42, 98-115.*

*3. Коваль С.Л. Анализ динамики психоэмоционального состояния по акустическим характеристикам речи / С.Л. Коваль, А.С.Белан // Труды Общества Независимых расследователей Авиационных происшествий. – М.: Полиграф, 2001. – Вып. 12а. – С. 316-337.*

*4. ГОСТ Р 50840 – 95. Передача речи по трактам связи. Методы оценки качества, разборчивости и узнаваемости. – М.: Госстандарт, 1995.*

*5. Галунов В.И. Акустическая теория речеобразования и система фонетических признаков / В.И. Галунов, В.И. Гарбарук // 100 лет экспериментальной фонетике в России: материалы международной конференции. – СПб.: Филологический факультет Санкт-Петербургского университета, 2001. – С. 58-62.*

*6. Рабинер Л. Цифровая обработка речевых сигналов / Л. Рабинер, Р. Шафер. – М.: Радио и связь, 1981. – 496 с.*

*7. Семенов В.Ю. Новый метод вычисления линейных спектральных частот речевых сигналов, основанный на универсальном алгоритме решения трансцендентных уравнений / В.Ю. Семенов // Акуст. Вісн. – 2002. – 5, N 4. – С. 38-50.*

*8. Патент РФ 2230375. Метод распознавания диктора и устройство для его осуществления / С.Л. Коваль, П.В. Лабутин, Л.И. Раев.– опубл. 10.06.2004.*

*9. Бодин Н.А. Оценка параметров распределения по группированным выборкам / Н.А. Бодин // Теоретические задачи математической статистики: тр. Института им. Стеклова. – 1970. – № 3. – С. 110-150.*

*10. Прикладная статистика: Классификация и снижение размерности / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин. – М.: Финансы и статистика, 1989. – 607 с.*

*11. Савaневич В.Е. Определение координат статистически зависимых объектов на дискретном изображении / В.Е. Савaневич // Радиоэлектроника и информатика. – 1999. – № 1. – С. 4-8.*

*12. Брандт З. Анализ данных: Статистические и вычислительные методы для научных работников и инженеров: пер. с анг. / З. Брандт. – М.: Мир, АСТ, 2003. – 686 с.*

*Поступила в редколлегию 20.05.2010*

**Рецензент:** д-р техн. наук, проф. В.А. Лошаков, Харьковский национальный университет радиоэлектроники, Харьков.

### ВИЗНАЧЕННЯ ЧАСТОТ ФОРМАНТ ГОЛОСОВОГО СИГНАЛУ НА ВИХОДІ ГРЕБІНКИ СМУГОВИХ ФІЛЬТРІВ

В.В. Поповський, С.М. Бобрицький, В.Є. Саваневич, В.М. Ткачов

*Розроблено метод визначення частот формант голосового сигналу на виході гребінки смугових фільтрів, що дозволяє поліпшити точність пофонемного короткочасного спектрального аналізу мовних фрагментів, що передаються в пакетних мережах з втратами. Для цього для отримання відповідної системи рівнянь максимальної правдоподібності використано математичний апарат групованих вибірок, що дозволяє врахувати факт дискретного, в частотному діапазоні, представлення енергії голосового сигналу при оцінці середніх частот формант мовного сигналу.*

***Ключові слова:*** *форманти мовних сигналів, пофонемний короткочасний спектральний аналіз мовних сигналів, груповані вибірки, локальне математичне очікування.*

### DETERMINATION OF FREQUENCIES OF FORMANT OF VOCAL SIGNAL ON OUTPUT OF COMB OF BANDPASS FILTERS

V.V. Popovskiy, S.M. Bobritskiy, V.E. Savanevich, V.N. Tkachev

*The method of determination of frequencies of formant of vocal signal is developed on the output of comb of polosovых filters, allowing to improve exactness of phonematic of brief spectral analysis of vocal fragments, passed in the networks of packages with losses. For this purpose for the receipt of the proper system of equalizations of maximal verisimilitude the mathematical vehicle of the grouped selections, allowing to take into account a fact discrete, is used, in a frequency range, presentations of energy of vocal signal at the estimation of middle frequencies of formant of vocal signal.*

***Keywords:*** *formants of vocal signals, phonematic brief spectral analysis of vocal signals, grouped selections, local expected value.*