УДК 004.934

М. Ф. Бондаренко, А. В. Работягов, С. В. Щепковский

# ПРОБЛЕМЫ РАСПОЗНАВАНИЯ РЕЧЕВЫХ ОБРАЗОВ СПЕКТРАЛЬНЫМИ МЕТОДАМИ В ОБЛАСТИ ПРИКЛАДНОЙ ЛИНГВИСТИКИ

## 1. Введение

Как известно, современное направление в области речевых технологий к решению основных задач — *кто говорит* (идентификация человека), *что говорит* (например, распознавание речевых команд искусственной технической системой, кодирование / декодирование речевого сообщения) и *как говорит* (установление эмоционального состояния человека) — характеризуется в целом как *спектральное*. Однако это направление имеет некоторые существенные проблемы.

## 2. Положения акустической теории речеобразования

Фундаментальные положения акустической *спектральной* теории речи были сформулированы в XIX ст. выдающимся немецким физиком Г. Гельмгольцем и до сегодняшнего дня остаются неизменными. В работах известных ученых-лингвистов, начиная от И. А. Бодуэна де Куртенэ, А. И. Томсона, Л. В. Щербы (XIX — начало XX вв.) и заканчивая Г. Фантом, Дж. Фланаганом, Л. Р. Зиндером, Л. А. Чистович, Л. В. Бондарко, М. А. Сапожковым, Р. К. Потаповой и др. (вторая половина XX — начало XXI вв.), мы видим, что для объяснения речевых процессов и количественного описания речевых сигналов используется их спектральное представление. На сегодняшний день главное положение акустической теории речеобразования[1] формулируется следующим образом. «Речевой сигнал возникает в результате воздействия одного или нескольких источников звука на систему резонаторов, образуемых воздушными полостями речевого тракта. Если обозначить через $S(f)$ амплитудно-частотный спектр колебаний, создаваемых источником звука, через $T(f)$ — передаточную функцию резонаторной системы речевого тракта, то амплитудно-частотный спектр результирующих звуковых колебаний $P(f)$ может быть представлен равенством:

$$P(f) = S(f) \cdot T(f).$$

Как показывает приведенная формула, частотная фильтрация состоит в том, что амплитуда каждой из частотных составляющих источника звука умножается на значение передаточной функции тракта на той же частоте. Свойства источников звука и резонаторной системы тракта не являются неизменными. Произнесение отдельных звуков и, тем более, звуковых последовательностей представляет собой сложный динамический процесс, характеристики которого меняются во времени. Поэтому в приведенную выше формулу включается параметр времени:

$$P(f, t) = S(f, t) \cdot T(f, t).$$

...В результате получается сложный периодический сигнал со спектром гармонической структуры, задаваемой голосовым источником» [1, с. 102–103].

Как известно, цель спектрального анализа состоит в «разложении» сложного суммарного колебания на отдельные составляющие элементы и измерении свойств этих элементов, представленных в виде количественных *спектральных* характеристик. Спектральные речевые характеристики получаются в результате разложения «речевой» функции в ряд Фурье. В первую очередь к спектральным характеристикам относятся мгновенный и текущие спектры, частоты основного тона, первой, второй и третьей формант, динамика частоты основного тона, которые определяются на участках языковых и речевых единиц: звуке, слоге, слове, словосочетании, предложении [2]. Используемые для целей обработки речи экспериментальные методы и математический аппарат (дискретное преобразование Фурье, теории вероятностей, статистики и др.) достаточно глубоко продуманы и детально изложены в научно-технической литературе.

В работах М. А. Сапожкова, Н. Г. Загоруйко и Р. К. Потаповой, Т. К. Винцюка [3–6] изложены общие подходы и методы анализа и описания речевых сигналов, имеющих в своей основе частотную структуру речевого сигнала, т. е. спектр[2].

Данные методы позволяют аппроксимировать речевой сигнал с различной степенью точности в зависимости от представления спектра. Они применяются

[1] «Задача акустической теории речеобразования состоит в том, чтобы выявить и количественно описать аэродинамические и акустические процессы, которые происходят в речевом тракте при артикуляции. Понимание этих процессов создает возможность обратных заключений: от акустики к артикуляционной картине» [1, с. 100].

[2] Спектр (лат. spectrum — «представление», «образ»). При спектральном представлении акустический сигнал представляется в виде «наложения» большого числа гармоник. Разложение сигнала в спектр обычно проводится с помощью быстрого преобразования Фурье (БПФ), реализованного в большинстве компьютерных звуковых редакторов и специальных программ обработки речи. Исследователи речи чаще всего используют представление речевого сигнала в виде трех- или двумерных сонограмм (спектрограмм). В первом случае по осям координат откладываются время, частота и спектральная плотность (энергия) частотной составляющей, а на двухмерной сонограмме ось энергии заменяется интенсивностью цвета в плоскости время–частота.

в различных модификациях в большинстве работ по анализу и распознаванию речи как один из основных уровней переработки сигнала для сегментации, формирования дифференциальных признаков (например, признаков фонем), определения места и способа образования звуков, переходных и временных характеристик. К этим методам относятся:

1) ортогональные методы, которые позволяют описывать мгновенный спектр речевого сигнала последовательностью коэффициентов Фурье;

2) корреляционные методы, использующие автокорреляционную функцию;

3) спектрально-полосные методы, которые проводят разложение сигнала в определенных полосах частот (по применяемому математическому аппарату относят к ортогональным методам);

4) формантные методы, где в качестве характеристик речевого сигнала рассматриваются параметры формант[3] и антиформант[4] — частота, амплитуда, ширина полосы, скорость изменения и др.;

5) метод ро-параметра, основанный на измерении плотности нулевых переходов речевого сигнала в соответствующей формантной области;

6) дискриминантный метод (является вариантом метода ро-параметра);

7) метод моментов спектра (по существу близок к методу нулевых переходов);

8) метод анализа через синтез (формантные параметры попарно сравниваются в компараторах с параметрами, выделяемыми основным анализатором);

9) метод выделения формантных частот путем последовательного исключения формант из речевого сигнала;

10) метод инверсных фильтров;

11) полюсно-нулевой метод;

12) кепстральный[5] метод (для определения периода основного тона);

13) статистические методы (варианты: метод линейного предсказания, автокорреляционный и ковариационный. Например, анализ речевого сигнала на основе линейного предсказания заключается в разложении его спектра на две составляющие: сглаженного спектра, представленного в виде модели спектра, характеристика которого содержит только полюсы, и спектра функции возбуждения, содержащего информацию о погрешности предсказания;

14) векторный метод;

15) метод кратковременных и быстрых преобразований Фурье;

16) метод преобразования Адамара (Уолша);

17) метод кодирования с предсказанием;

18) методы, базирующиеся на статистических характеристиках ритмики и темпа речи (количество фонетических слов, звуков в секунду, распределение длительности звуковых сегментов, речевых пауз);

19) отдельную группу образуют методы, в которых характеристику голосового источника применяют как компоненту свертки;

20) иерархические композиционные модели, основанные на теории оптимальных решений — динамическом программировании.

Например, на базе перечисленных методов в конкретных устройствах реализованы: бинарные (дихотомические) системы в двоичной метрике; системы на основе клиппированной речи и энергетического подхода, где определяются логарифмы энергии в различных полосах частот (т. н. спектр мощности: средний и мгновенный спектр, кросскорреляции спектральных компонентов); системы с пространственным представлением речи, в которых описание параметров происходит на базе нечетких множеств; динамические портреты речевого сигнала; параметры модели речеобразующего тракта; системы функциональной обработки, использующие параметры основного тона голоса — среднее значение, дисперсию, статистические моменты, изрезанность мелодического контура, микро- и макровариации; параметры модели линейного предсказания; гомоморфная обработка речевых сигналов.

В частности, в криминалистических подразделениях правоохранительных органов применяемые на практике компьютерные системы при проведении судебно-акустических (фоноскопических) экспертиз для идентификации человека по признакам речи в качестве основных детерминирующих признаков используют спектральные характеристики, которые получают, в основном, при помощи ортогональных, формантных и кепстральных методов.

### 3. Проблемы распознавания речевых образов

Следует указать на то важное обстоятельство, что эффективность экспериментально-фонетического исследования при спектральном отображении речевых сигналов зависит от соблюдения определенных требований, предъявляемых к речевому материалу и передаточным каналам записи. Эти требования относятся, главным образом, к сопоставимости (равнозначности):

1) объемов речевого материала;

2) фрагментов (образцов) речи для непосредственного исследования;

3) каналов записи (например, сопоставимость телефонных каналов или технических звукозаписывающих устройств);

4) эмоционального состояния говорящего;

5) манеры произношения говорящего.

[3] Форманта — частота максимума спектральной энергии, область концентрации энергии в спектре звука речи [3, с. 561].

[4] Антиформанта — частота минимального значения спектральной энергии в огибающей спектра звука речи [3, с. 561].

[5] Кепстр(ум) — «перевернутый» спектр, обратное дискретное преобразование Фурье (преобразование Фурье от логарифма модуля спектра) [3, с. 555].

Но каким образом проводить исследование, если такие требования не соблюдены? Как проводить сравнение образцов речи, если, например, на фоноскопическую экспертизу поступил речевой материал, представленный лишь одной фразой длительностью 3 секунды? На практике же при проведении фоноскопических экспертиз современными спектральными методами должны учитываться, в частности, следующие ограничения:

1) ограничение по длительности — минимальная длительность «чистой речи», как правило, должна составлять не менее 10 мин;

2) ограничение по спектру — речевой сигнал теряет часть полезной информации вследствие ограниченной частотной пропускной способности совокупного передаточного канала записи, предел которой в области верхних частот составляет 3...6 кГц для большинства диктофонов среднего класса (высококачественная звукозаписывающая аппаратура не учитывается, поскольку запись большинства фонограмм производится именно на диктофонах среднего класса) и 3 кГц — для телефонных линий. Это, естественно, отражается на спектре речевого сигнала и ограничивает «идентификационные» возможности спектральных методов.

Разрешить указанное противоречие возможно лишь при помощи иных, не использующих спектральную идеологию методов исследования речи. Это во-первых.

Во-вторых, спектральное отображение является всего лишь одним из возможных вариантов описаний речевых сигналов. Как отмечают специалисты в области науковедения, правила и методы эмпирической интерпретации теории могут развиваться бесконечно, так как любая теория не имеет предела в своем развитии. С позиции гносеологии, формы описания объекта познания бесконечны, поскольку сам объективный мир изменяется и развивается. Исходя из этого обстоятельства, правомерно сделать вывод: как существующая акустическая теория речеобразования, так и разработанные на ее основе спектральные методы исследования речи не могут претендовать на исчерпывающее и тем более абсолютное описание.

В-третьих, чрезмерно широкое применение математики, статистики (например, статистического анализа временных рядов) так или иначе «служит утонченным способом маскировки недостатка знаний о предмете исследования». Профессор У. Сиберт утверждает, что «применение статистики оправдано лишь после того, как решена основная задача классификации или идентификации параметров» [7, с. 133].

Действительно, никто не будет отрицать тот факт, что реальный речевой сигнал на выходе резонаторной системы речевого тракта — это сложный акустический сигнал. Данное обстоятельство логически оправдывает применение надежного, хорошо проверенного математического аппарата для дальнейшего исследования речевых сигналов. В качестве такого аппарата применяется преобразование Фурье для анализа временных рядов[6] (речевой сигнал как раз и представляется в таком виде). С его помощью находят значения частот и амплитуд гармонических составляющих сигнала. И решение, как кажется на первый взгляд, найдено. Однако, как отмечает известный специалист в области распознавания речи Дж. Р. Доддингтон, на этом пути возникают «определенные трудности в использовании столь привлекательных спектрографических различий (в частности, для распознавания человека по параметрам речи). ... Действительная трудность заключается в разном звучании и спектрографическом представлении одного и того же голоса в разное время» [8].

В отношении указанных выше спектральных методов следует отметить, что все они используют так называемые глобальные свойства объекта, к которым, несомненно, относится спектр. Как отмечается в книге «Распознавание образов» [7], глобальные операторы (признаки) оказались фактически бесполезными для решения некоторых задач распознавания образов. «Глобальные операторы «рассматривают» не деревья, а лес в целом; они, возможно, способны отличить лиственный лес от хвойного, однако не смогут обнаружить яблоню, затерявшуюся в березовой роще» [7, с. 269]. В противовес глобальным признакам в книге отдается предпочтение локальным признакам как наиболее эффективным для решения прикладных задач.

К тому же, исследование нестационарных сигналов при помощи преобразования Фурье обладает рядом существенных недостатков. Крупный специалист в области вычислительной математики (теории сплайнов, теории приближения функций, теории вэйвлетов[7]) и теории обработки сигналов, профессор Чарльз К. Чуи (США) отмечает: «Формула

$$f(\omega) = \int_{-\infty}^{\infty} e^{-it\omega} f(t)dt$$

преобразования Фурье в таком виде неудобна для практических задач» [9, с. 28]. «Неудобство» формулы состоит в следующем.

---

[6] Физически преобразование Фурье $S(f)$ представляет собой распределение интенсивности сигнала по частоте, т.е. является функцией плотности.

[7] Теория *вэйвлетов* (англ. wavelet — «малые волны»), или как их иногда называют в русской литературе *всплесков*, имеет истоки в таких классических областях математики, как теория функций вещественного переменного, теория ортогональных рядов, преобразование Фурье и др. интегральные преобразования, теория функций комплексного переменного, функциональный анализ. Теория вэйвлетов неразрывно связана с развитием прикладных областей современной науки: цифровой обработки сигналов и изображений, теории фильтрации и кодирования, теории сплайнов, дискретных и быстрых преобразований. Наиболее бурное развитие теории вэйвлетов приходится на 80–90-е годы XX века [9].

1) Чтобы извлечь спектральную информацию $f(\omega)$ об аналоговом сигнале $f(t)$ по этой формуле, «следует использовать бесконечные интегралы времени: иметь информацию о прошлом и будущем сигнала, чтобы вычислить спектр для одной частоты $\omega$».

2) «Формула не отражает эволюцию частот со временем. Что действительно необходимо — это определить интервалы времени, которые дают спектральную информацию о любой нужной частотной области (или диапазоне частот). Кроме того, так как частота сигнала обратно пропорциональна длительности его периода, то в случае высокочастотной спектральной информации временной интервал может быть взят относительно малым для обеспечения нужной точности, а в случае низкочастотной спектральной информации такой временной интервал должен быть взят относительно большим. Другими словами, важно иметь гибкое частотно-временное окно, которое автоматически сжимается в окрестностях высоких частотных центров и расширяется у низких частотных центров» [9, с. 28].

3) «Для изучения спектрального поведения аналогового сигнала необходимо полное знание сигнала во временной области, включая и будущую информацию. Вдобавок, если сигнал меняется в малой окрестности некоторого момента времени, то это влияет на весь спектр. Действительно, в крайнем случае преобразование Фурье от дельта-распределения $\delta(t-t_0)$ с носителем в единственной точке $t_0$ есть $e^{-it_0\omega}$, которое покрывает всю частотную область. Следовательно, во многих приложениях, таких как анализ нестационарных сигналов и обработка сигнала в реальном времени, применение одной формулы преобразования Фурье весьма неадекватно» [9, с. 91].

В этом аспекте также показателен вывод специалиста в области распознавания речи В. Н. Сорокина (Институт проблем передачи информации РАН) в отношении существующих методов, которые применяются непосредственно на этапах обработки первичных данных. В. Н. Сорокин отмечает, что системы распознавания речи, основанные на использовании метода скрытых марковских моделей достигли предела своих возможностей и, как следствие, все еще (!) не удовлетворяют большинству практических применений [10].

В-четвертых, оправдание и «весомые» аргументы в пользу применения спектральных характеристик звукового сигнала многие ученые пытаются найти в *спектральном* объяснении некоторых нейрофизиологических процессов, происходящих в так называемом *слуховом пути*. Выделим, в частности, следующее объяснение процессов передачи полезной информации на нейрональном уровне. «Центральный отдел слуховой системы имеет сложное нейроанатомическое строение. Образующие его нейроны организованы в несколько уровней, постепенно приближающихся к слуховой коре мозга, которая является конечной точкой так называемого классического *слухового пути*. Связь между уровнями в основном последовательная: нейроны данного уровня, получая информацию от предыдущего, передают ее после определенных преобразований следующему уровню. На всем пути сохраняются узкая настройка нервных клеток на определенные частоты, свойственная нейронам слухового нерва, и связь этих частот с пространственным расположением нейронов и их группировок в пределах своего уровня. Это значит, что каждая группа нейронов обрабатывает информацию в ограниченной частотной полосе, а анализ спектральной информации на всем пути ее прохождения к мозгу происходит в отдельных «сквозных» частотных каналах, в совокупности представляющих частотный диапазон человеческого слуха» [1, с. 226]. «Частотные» свойства периферической слуховой системы интересны как для специалистов, изучающих слух, так и для специалистов, исследующих восприятие речи, поскольку «в общем виде известно, что полезная информация, обеспечивающая восприятие речевого сигнала, заключена в его динамическом спектре» [11, с. 171]. Это подтверждается тем фактом, что реакция импульсной активности отдельных нейронов на воздействие какого-либо тонального сигнала действительно имеет место [12]. Однако является ли это действительно веским основанием для однозначного утверждения, что «нейроны обрабатывают именно спектральную информацию» и «анализируя спектр, можно объяснить (насколько это возможно) механизмы, например, восприятия речи? По мнению специалистов в области нейрофизиологии — нет. И этому есть серьезные обоснования. Механизм различения качества в слуховом органе действует «совсем по-иному» [12]. Общеизвестно, что различимых по высоте тонов больше тысячи. «Вдоль основной мембраны улитки лежат тысячи концевых органов, и на первый взгляд кажется, как и считал Гельмгольц, что каждый из них может представлять узкую полосу спектра слышимых звуков. Тем не менее оказалось, что элементы слухового нерва вовсе не обладают узкой настройкой, а что каждый рецептор обладает довольно широкой полосой реактивности, которая находит на полосу его соседа и занимает большой участок слышимых частот. ...Таким образом, надо отказаться от попыток объяснить всякое различение качества однородной теоретической схемой... В нервном коде один и тот же элемент не играет разных ролей в зависимости от всей констелляции. Вместо этого используются элементы разного рода, и каждый имеет свою ограниченную область в спектре воспринимаемых качеств раздражителя. ...Всякий стимул... может активировать много афферентных элементов

с разными рецептивными характеристиками. И полное описание значения стимула для организма осуществляется относительным распределением возбуждения... Такого рода организация получила название *частотный профиль популяции*, или *структура распределения по волокнам*» [12, с. 47].

К сказанному добавим следующее. Проведенное нами небольшое исследование показало, что слуховое восприятие двух речевых сигналов, имеющих одинаковый спектр, но разную пространственную ориентацию, совершенно различно. В исследовании диктор произносил слово *вокзал*. Речевой сигнал записывался в цифровом виде в компьютер. Этот сигнал назначался нами как исходный. Из исходного сигнала путем компьютерной «инверсии» нами формировался второй сигнал. В результате второй сигнал имел вид «перевернутого наоборот»[8] первого исходного сигнала, т. е. конец первого исходного сигнала становился началом второго сигнала. Оба сигнала хранились в цифровом виде в памяти компьютера. При прослушивании этих двух сигналов их аудитивное восприятие было совершенно различным: исходный сигнал звучал как *вокзал*, а смысл второго речевого сигнала был полностью разрушен. Однако в процессе определения акустических характеристик двух сигналов выяснилось, что их спектры одинаковы. Одним из объяснений этому явлению может стать представление речевого сигнала как некоторого события [12], определяемого не наличием некоторой совокупности частотных составляющих, а последовательностью некоторых акустических элементарных микрособытий (микроколебаний), где на первое место выдвигаются пространственно-временные и структурные представления [13].

Аналогичных представлений придерживается А. В. Бару. В своей работе «Слуховые центры и опознание звуковых сигналов» он пишет: «Предполагается, что человек имеет специализированные механизмы обработки акустического речевого сигнала, отличающиеся от процессов, при помощи которых обрабатываются перечевые акустические сигналы. ... В слуховом анализаторе должны быть некоторые наборы детекторов формы и скорости измерения сигнала» [14].

В-пятых, одним из основных препятствий, стоящих на пути распознавания речевых образов, является несостоятельность существующих в настоящее время акустико-фонетических признаков. Например, в современных компьютерных системах автоматического распознавания и понимания слитной речи, сориентированных на дикторонезависимый интерфейс (например, в системе Via Voice), все еще наиболее уязвимым является отсутствие базы параметрических данных в области просодии[9] того или иного языка [3, с. 528].

Как известно, универсальная акустическая классификация (*известная также как теория различительных признаков* [15]) является, по существу, попыткой описания звуковых контрастов на основе акустических характеристик. Ее в 1952 г. предложили, а затем дополняли и уточняли Р. Якобсон, Г. Фант и М. Халле. В этой классификации выделены основные акустические признаки, на основе которых формируются важнейшие звуковые противопоставления[10]. Признаковая база акустической классификации включает 12 бинарных различительных признаков, набор которых считается достаточным для описания смыслоразличительных звуковых контрастов, возможных в конкретных языках. Каждый признак представляет собой противопоставление между двумя относительными проявлениями одного и того же акустического свойства, т. е. своего рода шкалу выраженности определенного акустического качества. Звуки, у которых степень проявления акустического свойства больше некоторого порога, имеют положительное значение соответствующего признака-классификатора, в противном случае — отрицательное. Отличительная особенность этой системы признаков состоит в том, что для описания согласных и гласных предлагается один и тот же набор признаков, в соответствии с которым звук речи классифицируется [1, с. 296–303; 15] как:

1) вокальный / невокальный (вокальные звуки, в отличие от невокальных, акустически характеризуются четко выраженной формантной структурой);

2) консонантный / неконсонантный (консонантные звуки, в отличие от неконсонантных, имеют более низкий общий уровень интенсивности);

[8] В исследованиях И. П. Дукельского такой способ называется *инверсальным* или *обратным* прослушиванием. Инверсальное прослушивание, т. е. прослушивание в обратном порядке ранее записанных на пленке слов или слогов, следует отличать от *инверсии*, т. е. произнесения в обратном порядке слов или слогов. При инверсальном прослушивании оказывается нарушенной последовательность расположения отрезков речевого потока, в то время как инверсия предполагает автоматическую замену звуков и, соответственно, положений. В отличие от обычного, нормального восприятия речевого потока, при инверсальном прослушивании «большинство звуков теряет в большей или меньшей степени естественный характер звучания, а именно — свою ясность, однородность, слитность, громкость, нормальную длительность» [16, стр. 48]. При этом необходимо выделить психологические особенности восприятия. «По сравнению с нормальным прослушиванием при инверсальном прослушивании период времени, необходимый для восприятия слога, значительно возрастает, что свидетельствует о функциональном нарушении *протекания нервного процесса*» [16, стр. 52].

[9] Просодия — «система фонетических средств (высотных, силовых, временных), реализующихся в речи на всех языковых уровнях. Вышеуказанные фонетические средства соответствуют основным физическим характеристикам: частоте основного тона, интенсивности и длительности» [3, с. 558].

[10] Например, спектрограммы, иллюстрирующие большинство звуковых контрастов, полно представлены в «Общей фонетике» С. В. Кодзасова и О. Ф. Кривновой [1].

3) прерванный / непрерванный (к прерванным относятся согласные, у которых акустически выражен интервал отсутствия или сильного ослабления звуковой энергии в полосе частот выше основного тона, после которого следует «взрыв» или резкое изменение формантной картины);

4) глоттализованный / неглоттализованный (глоттализованные согласные характеризуются резким включением интенсивного источника шума);

5) резкий / нерезкий (в основе признака лежат относительные различия в интенсивности, длительности и степени упорядоченности фрикативного шума, т. е. своего рода шкала шумности. Резкие согласные, в отличие от нерезких звуков, имеют интенсивный и длительный шум);

6) звонкий / глухой (звонкие звуки, в отличие от глухих, произносятся с участием голосового источника);

7) компактный / диффузный (компактные звуки, в отличие от диффузных, характеризуются большей концентрацией энергии в относительно узкой срединной (у диффузных — краевой) части частотного диапазона спектра, большей интенсивностью и длительностью);

8) низкий / высокий (к низким (низкотональным) относятся звуки, у которых энергия сосредоточена в более низких частотах, чем у высоких (высокотональных));

9) бемольный / простой (у бемольных звуков, в отличие от простых, частоты всех формант понижаются и энергия высокочастотной части спектра ослабляется);

10) диезный / простой (у диезных вокальных звуков увеличиваются частоты нижних формант и возрастает интенсивность по сравнению с простыми (недиезными) звуками);

11) напряженный / ненапряженный (напряженные звуки, в отличие от ненапряженных, характеризуются большей длительностью, большей интенсивностью, более отчетливым и богатым спектром);

12) носовой / ртовый (у носовых звуков, образуемых с участием носового резонатора, в отличие от ртовых, в спектре появляются устойчивые, мало изменяющиеся форманты назализации (200...300 Гц)).

В-шестых, на сегодняшний день среди специалистов в области лингвистики нет единого, однозначного мнения по поводу того, какую же структурную единицу (участок, элемент) речевого сигнала на акустико-фонетическом (физическом) уровне принимать за минимально-информационный *сегмент* для надежного распознавания.

Изучение любого речевого сигнала «предполагает предварительную сегментацию этого сообщения с выделением ключевых сегментов и их признаков, отношения между ними, их структурной организации с целью дальнейшего распознавания и понимания последнего. При этом встает вопрос определения самих единиц сегментации речевого высказывания, методики, критериев их обнаружения, специфики и формы их взаимодействия» [3, с. 269].

Для подтверждения к сказанному достаточно перечислить такие структурные единицы, как фонемы, аллофоны, транземы (или дифоны), слоги, звуки и т. п. Эти единицы принимают за минимальный участок речевого сигнала и обозначают такими понятиями, как «сегмент», «минимальный сегмент». Большинство исследователей относит понятие «минимальный сегмент» либо к слогу, либо к звуку [3], либо к фонеме [1], либо к некоторой области речевого сигнала, длительностью порядка 10...30 мс, либо к некоторому гибкому частотно-временному окну [9]. В работе [13] в качестве «минимального сегмента» принят так называемый «элементарный сегмент гласного звука речи».

Так, например, Р. К. Потапова к элементам микросегментации относит сегменты следующего порядка: интразвуковые (межзвуковые переходные процессы, смычка, фрикация, эксплозия и т. д.), звуковые, интерзвуковые (сочетания двух соседних звуков) и слоговые. Она пишет: «Для решения проблемы сегментации звучащей речи большое значение имеет обращение к слогу... Сегментация может проводиться в два этапа: на слоги, а затем на звуки, их составляющие, в результате чего уточняются границы между слогами» [3, с. 300]. Л. В. Бондарко определяет слог как «минимальный сегмент речевой цепи». С. В. Кодзасов и О. Ф. Кривнова указывают, что «звуковая информация об означающем языкового знака не может быть сведена только к его фонемному составу, поскольку существуют звуковые явления, сферой реализации которых служат фонемные цепочки: слоги или слова. Такие явления относят к *супрасегментным* (от лат. supra — «над») звуковым средствам языка, в отличие от фонем — *сегментных* единиц» [1, с. 26–27]. Н. И. Дукельский в работе «Принципы сегментации речевого потока» отмечает, что «единственным критерием для отделения в речевом потоке кратчайшего функционально значимого отрезка... является... изменение вида источника речеобразования. Такого рода кратчайшие функционально значимые отрезки речевого потока будем называть сегментами» [16, с. 21]. На представленных в работе Н. И. Дукельского осциллограммах речевого потока «сегменты» выглядят в виде некоторых речевых участков, соответствующих определенному звуку речи, как согласному, так и гласному. Границами, определяющими расположение «сегмента», являются границы между двумя соседними звуками.

Однако, по нашему мнению, назначая объект своего исследования в соответствии с указанными выше представлениями, многие ученые сталкиваются

с трудностями принципиального характера. Они заключаются в том, что первичные признаки, которые определяются на рассмотренных выше минимальных сегментах, теряют свою информативную способность. А ведь, как известно, от правильного выбора минимального сегмента и, следовательно, первичных признаков зависит успех распознавания в целом. Решение любой прикладной задачи в области речевых технологий неразрывно связано с построением формальной системы в базисе теории распознавания образов, где одним из основных этапов процесса распознавания является этап формирования первичных *информативных* признаков.

Поэтому, если рассматривать в качестве минимального сегмента:

1) слог, звук или «сегмент» Дукельского, то это не отвечает объективной (физической) реальности, так как сами слог, звук или «сегмент» Дукельского «построены» из еще более простых составных структурных элементов (как будет показано ниже);

2) фонему[11], то данная лингвистическая категория не может рассматриваться на акустическом уровне, поскольку это идеальные семиотические элементы, идеальные семиотические ценности, не заключающие в себе ничего физического, это «абстрактные единицы» [3, с. 85];

3) некоторую автоматически выделяемую область (области) речевого сигнала фиксированной или переменной (так называемые временные окна[12]) длительности, что вообще характерно для исследований речи спектральными методами, то в этом случае разрушается целостность объекта (как физическая, так и функциональная) и, как следствие, его информативность.

Как видим, вследствие указанных причин первичные признаки, формируемые в границах приведенных вариантов минимального сегмента речи, лишены своей качественной информативной определенности.

Следует сказать, что особенно остро проблема определения материальных (физических) границ, например между звуками, «внутри» самого звука, встает каждый раз перед теми, кто пытается описать речевые явления в терминах точных наук. Об отсутствии четких акустических признаков границ звука известный шведский фонетист Г. Фант пишет: «В результате подобной чисто акустической сегментации может быть получено некоторое число минимальных звуковых единиц, имеющих размер, равный размеру звука речи или меньший... Число таких последовательных во времени звуковых единиц, как правило, больше числа символов фонетической или фонематической транскрипции» [17, с. 35].

Целесообразно привести примеры акустических признаков сегментации речевого сигнала, которые так же, как и в случае идентификации человека при помощи современных методов, носят, в основном, спектральный характер. К ним относятся [3, с. 271]: наличие — отсутствие частоты основного тона; скачкообразное повышение частоты основного тона на переходе от согласного к гласному; скачкообразное понижение частоты основного тона на переходе от гласного к согласному; наличие / отсутствие шума; локализация полос шума на шкале частот; интенсивность полос шума; крутизна нарастания шума; длительность шума; наличие низко-, высокочастотной энергии; длительность сегмента.

В-седьмых, на сегодняшний день недостаточно глубоко, по нашему мнению, продуман вопрос, связанный с методами последующей обработки первичных акустических признаков.

Например, методы фоноскопической идентификации человека, в основном, используют так называемую «кодирующую» концепцию сравнения образцов речи. К этой группе относятся методы, формирующие для цели идентификации: 1) «кодовую страницу», 2) «параметрические коды»[13], 3) «отпечатки голосов»[14] и 4) «эталоны» дикторов[15]. Несмотря на различия в названиях, суть указанных методов одинакова. Обычно полная процедура идентификации, в которой диктору присваива-

[11] «Основными конкретными единицами фонетического уровня являются звуки, а основными абстрактными единицами — фонемы и морфонемы» [3, с. 122]. В устной речи фонема представлена в виде своих оттенков, вариантов, аллофонов, одни из которых меньше зависят от окружающих звуков, другие — больше. Первые называются *основными* аллофонами, вторые — *комбинаторными* и *позиционными* [3, с. 88].

[12] Оценка спектральных параметров речи производится с использованием окон частотного взвешивания. Существует множество различных временных окон, наиболее известны из них: окно Хеммингa, косинусное окно, окно Гаусса, прямоугольное окно, окно Блэкмана, «функция-окно» Габора, гибкое частотно-временное окно интегрального вэйвлет-преобразования и др. Наибольший интерес в исследовании речи спектральными методами представляет: 1) окно Гаусса, т. к. оно определено на всей числовой оси, содержит только один максимум в амплитудном спектре и не содержит ни одного артефакта, а также легко применимо при разного рода аналитических вычислений; 2) «функция-окно» Габора, параметры которого используются для перемещения окна с целью покрытия всей временной области для получения локальной информации о преобразовании Фурье сигнала [9]; 3) гибкое частотно-временное окно интегрального вэйвлет-преобразования, которое автоматически сжимается в окрестностях высоких частотных центров и расширяется у низких частотных центров [9].

[13] *Галяшина Е.* Речь под микроскопом. (Интернет-статья), 1999 г.

[14] Т. н. Система оперативной верификации и идентификации голоса (СОВИГ); разработчики — Д. Н. Коновалов, д.ф.-м.н., и А. Г. Бояров; система была представлена на прошедшей 2 ноября 2000 г. конференции «Информационная безопасность компьютерных систем».

[15] *Галунов В. И.* Верификация и идентификация говорящего. (Интернет-статья), 2000 г.

ется определенного рода код (кодовая комбинация), такова. Анализ речевого сигнала начинается с перевода его в цифровую форму. Производится сегментация сигнала на отдельные элементы. Затем цифровой сигнал обрабатывается с помощью определенных алгоритмов (спектрального анализа, линейного предсказания, кепстральной обработки и др.). В результате получается параметрическое описание сегментов речевого сигнала в виде вектора первичных параметров. Следующий этап — это сравнение с имеющимися эталонными описаниями зарегистрированного числа дикторов в базе данных компьютера. Механизм сравнения реализуется при помощи метода динамического программирования, скрытых марковских моделей (в основном, для распознавания слитной речи), искусственных нейронных сетей или комбинаций указанных методов. По существу же задача компьютера — распознать человека по параметрам речевого сигнала — не отличается от задачи, решаемой экспертом-криминалистом. Задача компьютера заключается в том, чтобы сравнить некоторый код неизвестного диктора с эталонным кодом заявленного диктора (при верификации) или с эталонным параметрическим «кодовым» описанием каждого из конечного числа зарегистрированных дикторов (при идентификации).

Если сравнение при верификации показывает приемлемую «близость», которая вычисляется по определенным критериям, то система считает диктора «своим», а если значение близости превышает некий порог, то диктор объявляется «чужим». При идентификации компьютер, сравнивая «спектральные коды», выбирает наиболее «близкий» код диктора из числа кодов, имеющихся в его памяти. При этом следует подчеркнуть, что попытки применения на практике рассмотренной «кодовой» концепции привели к ряду критических замечаний (Е. Галящина, В. И. Галунов и др.).

## 4. Заключение

Таким образом, анализ современного состояния исследований в области распознавания речевых образов показал ряд проблем существующей речевой теории. Одно из решений этих проблем мы видим в развитии структурного направления в области прикладной лингвистики [13].

**Список литературы:** 1. *Кодзасов С. В., Кривнова О. Ф.* Общая фонетика: Учебник.– М.: Рос. гос. гуманит. ун-т, 2001.– 592 с. 2. *Головин Б. Н.* Введение в языкознание: Учебн. пособие.– М.: Высшая школа, 1983.– 232 с. 3. *Потапова Р. К.* Речь: коммуникация, информация, кибернетика: Учебн. пособие.– М.: Едиториал УРСС, 2003.– 568 с. 4. Распознавание слуховых образов / Под ред. Н. Г. Загоруйко.– Новосибирск: Наука, Сибирск. отд-ние, 1970.– 388 с. 5. *Сапожков М. А.* Речевой сигнал в кибернетике и связи.– М.: Связьиздат, 1963.– 452 с. 6. *Винцюк Т. К.* Анализ, распознавание и интерпретация речевых сигналов.– Киев: Наук. думка, 1987.– 262 с. 7. Распознавание образов. Исследование живых и автоматических распознающих систем: Пер. с англ.– М.: Мир, 1970.– 288 с. 8. *Доддингтон Дж. Р.* Распознавание дикторов: Идентификация людей по голосу // ТИИЭР.– 1985.– Т. 73, № 11.– С. 129–137. 9. *Чуи Ч.* Введение в вэйвлеты: Пер. с англ.– М.: Мир, 2001.– 412 с. 10. Современные речевые технологии: Сб. трудов IX сессии Российского акустического общества.– М.: ГЕОС, 1999.– 166 с. 11. *Чистович Л. А., Венцов А. В., Гранстрем М. Б. и др.* Физиология речи. Восприятие речи человеком.– Л.: Наука, Ленингр. отд-ние, 1976.– 386 с. 12. *Сомьен Дж.* Кодирование сенсорной информации: Пер. с англ.– М.: Мир, 1975.– 414 с. 13. *Бондаренко М. Ф., Дрюченко А. Я., Шабанов-Кушниренко Ю. П.* Гласные звуки в теории и эксперименте.– Харьков: ХНУРЭ, 2002.– 348 с. 14. *Бару А. В.* Слуховые центры и опознание звуковых сигналов.– Л.: Наука, Ленингр. отд-ние, 1978.– 192 с. 15. *Якобсон Р., Фант Г., Халле М.* Введение в анализ речи. Различительные признаки и их корреляты // Новое в лингвистике. Вып. II.– М.: Изд-во АН СССР, 1962.– С. 57–68. 16. *Дукельский Н. И.* Принципы сегментации речевого потока.– Л.: Ленингр. отд-ние изд-ва АН СССР, 1962.– 140 с. 17. *Фант Г.* Акустическая теория речеобразования: Пер. с англ. / Под ред. В. С. Григорьева.– М.: Наука, 1964.– 284 с.

*Поступила в редколлегию 18.02.2006*