АКУСТИЧЕСКИЙ ЖУРНАЛ, 2020, том 66, № 1, с. 71–85

ОБРАБОТКА АКУСТИЧЕСКИХ СИГНАЛОВ. КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ

УДК 612.85

ДЕТЕКТОРЫ АРТИКУЛЯТОРНЫХ СОБЫТИЙ

© 2020 г. В. Н. Сорокин*

Институт проблем передачи информации РАН, Б. Каретный пер. 19, Москва, 127994 Россия

**E-mail: vns@iitp.ru*

Поступила в редакцию 22.05.2019 г.
После доработки 03.09.2019 г.
Принята к публикации 05.09.2019 г.

Детекторы артикуляторных событий, т.е. детекторы перехода из одного артикуляторного состояния в другое, формируются на основе анализа спектрально-временных неоднородностей в речевом сигнале. Сегментация и распознавание триады типа */пауза–фрикативный–гласный/* выполняется в пространстве главных компонент спектра отклика детектора переходного процесса от паузы к фрикативному, спектра фрикативного в момент пика его энергии и спектра отклика детектора переходного процесса от фрикативного к гласному в момент пика этого детектора. Среднеквадратическая ошибка относительно ручной разметки для начала фрикативных составляет, в среднем, около 12 мс, а для момента перехода от фрикативного к гласному – около 5 мс. Ошибки распознавания триад с одним и тем же фрикативным и разными последующими гласными, а также ошибки распознавания триад, отличающихся только наличием или отсутствием голосового возбуждения, оказались порядка нескольких процентов.

Ключевые слова: распознавание речи, сегментация речевого сигнала, детекторы артикуляторных событий, детекторы начала речи

DOI: 10.31857/S0320791920010104

1. ВВЕДЕНИЕ

На заре эпохи автоматического распознавания речи наивно предполагалось, что речь состоит из последовательности фонем, так же как письменный текст представлен последовательностью букв. Технически задача формулировалась как создание автоматической пишущей машинки, на вход которой поступает речь, а выходом является текст. В то время казалось, что достаточно найти акустические параметры каждой фонемы, и проблема распознавания речи будет решена. В процессе исследований выяснилось, что акустические параметры речевого сегмента, который воспринимается как некоторая фонема и которому можно приписать соответствующий буквенный символ, чрезвычайно разнообразны. Это разнообразие связано с взаимодействием артикуляционных процессов, особенностями артикуляции дикторов и различными условиями внешней среды. Главный источник разнообразия состоит во взаимном влиянии артикуляторных процессов, так что акустические параметры фонемы зависят и от предыдущих, и от последующих звуков. Поэтому пришлось отказаться от поиска акустических инвариантов фонем, и проблема автоматического распознавания речи некоторое время находилась в идеологическом кризисе.

Выход из кризиса стали искать в формальном математическом подходе, который получил название “ignorance based approach”. Для сокращения объема информации и уменьшения влияния частоты основного тона на спектр речевого сигнала используется кепстральное преобразование, состоящее в обратном преобразовании Фурье логарифма спектра мощности. На фиксированном скользящем интервале длительностью 15–25 мс вычисляется примерно 20 коэффициентов кепстра кратковременного спектра речевого сигнала на этом интервале, а также первые и вторые разности этих коэффициентов по времени. Последовательность векторов этих параметров подвергается статистическому анализу. Наиболее успешным оказался подход, в котором предполагалось, что речевой сигнал состоит из последовательности некоторых абстрактных символов (не фонем), и вероятность перехода из одного состояние в другое можно описать с помощью скрытых Марковских моделей. Применение этого подхода позволило расширить объем распознаваемого словаря до сотен тысяч словоформ, и разработать системы распознавания, практически приемлемые в узком сегменте задач. Также популярны методы распознавания, основанные на использовании искусственных нейронных сетей. При этом основной источник информации о речи заключается

в моделях языка – лексических, синтаксических и прагматических. Именно прагматические ограничения на используемый словарь и структуру фраз обеспечивают определенную эффективность используемых в настоящее время коммерческих систем распознавания.

Такой подход обладает двумя принципиальными недостатками, определяющими практическую невозможность существенного улучшения эффективности распознавания речи в задачах с произвольной тематикой. Первый недостаток заключается в статистической природе метода, в результате которой вероятность правильного распознавания слов катастрофически падает в условиях эксплуатации, отличающихся от условий обучения. Различие в типах микрофона, расстоянии и направлении на него, наличие внешних шумов и реверберации помещений могут привести к такому ухудшению характеристик, что распознавание становится практически невозможным. Второй недостаток состоит в выборе единиц распознавания, которые лишь косвенно связаны с объективно существующими единицами восприятия речи.

Другое направление исследований исходит из представления о том, что при восприятии речи наиболее информативными являются не столько стационарные состояния артикуляторов, сколько переходы из одного состояния в другое [1]. Это представление было сформулировано в [2, 3] в виде так называемых “landmarks”, отмечающих моменты быстрого изменения спектра или смены дифференциальных признаков фонем в потоке речи. На основе этих представлений исследуются потенциальные возможности систем распознавания речи [4–8]. Обзор методов распознавания с использованием идеи landmarks представлен в [9]. Однако и при таком подходе принципиально важным является определение единицы восприятия речи и физически адекватное описание перехода от одной единицы к другой. В работах Stevens за единицу распознавания принимается фонема, а дифференциальные признаки фонем фактически являются инвариантами, несмотря на то, что опыт предыдущих исследований показывает бесперспективность поиска инвариантов.

При поиске единиц распознавания и методов сегментации речевого потока на эти единицы следует использовать сведения о свойствах слухового восприятия. Установлено, что переходные процессы и стационарные состояния звука сопровождаются активизацией разных отделов слуховой зоны коры головного мозга [10]. Некоторые музыкальные инструменты трудно различить по их звучанию на стационарных звуках, тогда как это различие определяется на переходах от одного звука к другому. С другой стороны, спектр изолированных стационарных фрикативных звуков речи позволяет оценить их фонетическое качество. Поэтому при восприятии звука важны как спектры стационарных состояний, так и переходные процессы. Необходимо оговориться, что в исследованиях этих переходных процессов используются жаргонные термины “амплитудная модуляция” и “частотная модуляция”, которые не соответствуют определению модуляций в технических науках, и просто подразумевают любое изменение амплитудных или частотных параметров речевого сигнала.

Хорошо известны так называемые on- и off-эффекты, сопровождающиеся всплеском активности слуховой системы при включении и выключении звукового стимула. Найдены нейроны, отвечающие за эти эффекты [11, 12]. Обнаружены также нейроны, реагирующие на амплитудные или частотные модуляции звука [13–18]. Некоторые нейроны обладают порогом по скорости изменения огибающей звукового сигнала [19] или избирательно реагируют на возрастание или уменьшение частоты [20, 21]. Восприятие амплитудно-спектральных модуляций в слуховой системе зависит также от эффектов временной маскировки [22–26]. Обнаружено также, что в слуховой системе человека выполняется сглаживание сигналов с различными постоянными времени: от 2 до 100 мс. Восприятие амплитудных модуляций улучшается, если кратковременной модуляции предшествует долговременная [27, 28].

Некоторые свойства слуховой системы относительно восприятия амплитудно-частотных модуляций речевого сигнала были реализованы в [29] в виде оператора, вычисляющего разность логарифмов амплитудных спектров речевого сигнала, сглаженных по частоте или времени с различными параметрами. Отсчеты по времени этих спектров могут выполняться с различными сдвигами – как с задержкой, так и с опережением, имитируя эффекты временной маскировки.

Система управления артикуляцией формирует переход от одной формы речевого тракта к другой, так что единицей управления является слог [30]. Представление о том, что слог служит единицей распознавания речи, было сформулировано в [31]. В исследованиях процессов восприятия речи рассматриваются сегменты типа слогов, состоящих из двух или трех артикуляторных состояний. Обсуждаются такие кандидаты, как закрытые слоги типа гласный-согласный (ГС), открытые слоги типа согласный гласный (СГ), а также различные сочетания из трех элементов СГС, СГГ, ССГ, ГСГ, ГСС. Количество таких слогов весьма велико. Поэтому отсутствие представительных баз речевых данных, малые объем памяти и скорость обработки данных компьютерами ранее не позволяли поставить задачу формирования эталонов слогов для конкретного языка и создать на этой

основе алгоритмы автоматического распознавания речи, в какой-то степени адекватные процессам субъективного распознавания. В настоящее время созданы обширные базы речевых данных, и производительность компьютеров приближается к необходимой для решения этой задачи. Теперь главная проблема состоит в разработке алгоритмов анализа динамики и квазистатических состояний артикуляторов путем исследования спектрально-временных неоднородностей речевого сигнала. В случае успеха можно рассчитывать на существенное повышение вероятности правильного распознавания элементов речи на уровне слогов, и, в конечном счете, на такое улучшение эффективности систем автоматического распознавания речи, которое сделает показатели технических систем сравнимыми со свойствами субъективного распознавания.

В данной статье описывается подход к формированию детекторов артикуляторных событий, т.е. перехода из одного артикуляторного состояния в другое, на примере сегментации и распознавания триад типа */пауза–фрикативный–гласный/*.

2. МОДЕЛИ АКУСТИЧЕСКИХ ПРОЦЕССОВ

2.1. Спектральный анализ

Исследования слухового анализатора человека привели к представлениям о том, что в его периферическом отделе выполняется спектральный анализ звука. В этот анализ вовлечены колебания базилярной мембраны, которые вызываются гидродинамическими процессами в каналах внутреннего уха, и отклики внутренних волосковых клеток на эти колебания. Четыре ряда внешних волосковых клеток участвуют в положительной механической обратной связи, усиливая колебания базилярной мембраны в области пучностей. В целом вся система спектрального анализа нелинейна, и ее математическое описание настолько сложно, что пока не получило применения в речевых технологиях.

В анализе речи наиболее распространены различные виды кратковременного преобразования Фурье. В этом преобразовании используются так называемые “окна”, которые свертываются с речевым сигналом на N отсчетах конечного интервала времени. Существует большое количество этих окон – прямоугольное, в котором сигнал не преобразуется, треугольное (окно Блекмана), множество окон Кайзера–Бесселя с разными параметрами, окна Хемминга, Хенинга и Лапласа. Параметры кратковременного преобразования Фурье оказывают существенное влияние на вид динамического спектра, и нет никаких теоретических соображений относительно их выбора при анализе речевого сигнала. Наиболее велика погрешность такого преобразования при малом числе отсчетов, например, при раздельном анализе речевого сигнала на интервалах открытой и закрытой голосовой щели внутри периода основного тона. Спектральные характеристики речевого сигнала можно определить и посредством алгоритмов линейного предсказания, wavelet-анализа или с помощью преобразования Уолша. Недостаток всех формальных методов спектрального анализа речевого сигнала состоит в отсутствии прямой связи со свойствами слухового анализа человека.

Вместо формальных методов спектрального анализа и чрезмерно сложных нелинейных моделей периферического слухового анализа – от барабанной перепонки до активности внутренних волосковых клеток, предпринимаются попытки феноменологического описания результатов субъективного спектрального анализа с помощью относительно простых математических средств [32]. Другой подход представлен в [33], где весовая функция каждого фильтра слухового анализатора описывается как

$$g_k(t) = t^{n-1}e^{-b_k t}\cos(\omega_k t + \varphi_k), \qquad (1)$$

где n – порядок функции, b_k определяет ширину полосы пропускания, которая пропорциональна центральной частоте фильтра, ω_k – центральная круговая частота, а φ_k – фазовая константа, которая обычно принимается равной нулю. Помимо простоты описания, этот подход обладает тем преимуществом, что число фильтров k не зависит от частоты дискретизации речевого сигнала, и может быть выбрано для достижения требуемой точности представления спектра. Этот подход исследовался в [34], где было найдено, что при $n = 3$, 4, 5 это выражение хорошо описывает поведение слухового фильтра. Впоследствии такой способ описания динамического спектра речевого сигнала получил название системы гамма-тона (gamma-tone) и использовался для распознавания речи в [35].

В экспериментах, описываемых ниже, спектральный анализ выполняется в системе гамматона для 512 фильтров, расположенной по шкале мел в диапазоне частот $f = 50...8000$ Гц с частотой дискретизации речевого сигнала 16 кГц. При вычислении спектра отклик каждого фильтра $r_k(t)$ на входной речевой сигнал $s(t)$,

$$r_k(t) = \int_{-\infty}^{t} g_k(t-\tau)s(\tau)\,\mathrm{d}\,\tau, \qquad (2)$$

преобразуется как $S_k(t) = |r_k(t)|$ с последующим сглаживанием. Поскольку задержка отклика гамма-фильтров различна, в данной работе было выполнено выравнивание откликов путем измерения задержки для тестового сигнала в виде дельта-функции.

2.2. Модель взаимодействия турбулентного источника возбуждения с речевым трактом

Фрикативные согласные генерируются с участием источника шума турбулизации воздушного потока в речевом тракте там, где вслед за сужением имеется расширение. В отличие от голосового источника, этот источник является источником давления. Теоретические модели взаимодействия источника шума с акустическими процессами в речевом тракте описывались в [3, 36, 37] методами длинной линии на основе электромеханических аналогий или аппарата передаточных функций. Спектральные свойства фрикативных согласных изучались на физических моделях [38, 39] и путем математического моделирования с использованием прямых MRI измерений формы речевого тракта [40].

При артикуляции глухих фрикативных в речевом тракте возникают два источника турбулентного шума. Один источник находится в месте наибольшего сужения, координата которого вдоль средней линии речевого тракта определяется типом артикулируемого фрикативного. Другой источник находится на выходе из голосовой щели. Согласно [3], площадь голосовой щели S_{vs} при артикуляции глухих фрикативных близка к 0.3 см^2. Его спектральные характеристики мало зависят от формы речевого тракта. Частота первого обертона турбулентного шума на выходе из голосовой щели, найденная в [41], близка к 700 Гц. Эта величина находится в диапазоне оценок, полученных в [38] на физической модели голосовой щели.

Характеристики шума турбулентного потока – спектр и интенсивность – изучаются в специфических экспериментах с обтеканием препятствий определенного вида [42, 43]. Установлено, что если число Рейнольдса Re превышает критическое значение $\mathrm{Re}_{\mathrm{cr}}$, то широкополосный спектр турбулентного шума обладает пиками энергии на частотах

$$f_n = n\mathrm{Sh}(\mathrm{Re})\,v/d, \qquad (3)$$

где $n = 1, 2, \ldots$, Sh – число Струхаля (для относительно гладких труб $\mathrm{Sh} \approx 0.2$), v – скорость воздушного потока в сужении, d – эквивалентный диаметр сужения. Амплитуда обертонов шума A_n быстро падает с ростом n. По оценке [44], для речевого тракта $\mathrm{Re}_{\mathrm{cr}} \approx 1600$–$1800$. В экспериментах на препаратах гортани собак были получены оценки критических чисел Рейнольдса $\mathrm{Re}_{\mathrm{cr}} \approx 1800$–$7000$ на выходе из голосовой щели в [45]. Моделирование турбулентных шумов в артикуляторном синтезаторе дает типичную оценку для фрикативных звуков $\mathrm{Re} \approx 3000$, так что частоты первого резонанса турбулентного шума оцениваются как 2050 Гц для /*с*/, 1460 Гц для /*ш*/, 985 Гц для /*х*/ и 1150 Гц для /*ф*/ [41].

Скорость воздушного потока в сужении речевого тракта, которая необходима для возникновения турбулентного шума, обеспечивается надлежащим перепадом давления ΔP под и над голосовой щелью. В экспериментах по непосредственному измерению подсвязочного давления в [46] было установлено, что для звонких фрикативных это давление выше, чем для глухих фрикативных. Это связано с тем, что максимальная площадь голосовой щели при автоколебаниях голосовых складок меньше, чем при артикуляции глухих фрикативных, и, соответственно, сопротивление потоку выше.

Оценки частот и относительных амплитуд пиков в спектре турбулентных шумов важны, но математическая модель спектра турбулентного шума при артикуляции фрикативных звуков все же неизвестна. Вместе с тем, можно предложить качественную теорию взаимодействия источника турбулентного шума с речевым трактом. Математическая модель акустических процессов в речевом тракте обычно описывается как волновое уравнение типа Вебстера, причем в полосе частот до 4 кГц справедливо предположение о существовании только плоских волн. Для этого уравнения известна зависимость парциального возбуждения $a_k(t)$ резонансных колебаний от источника $G(x,t)$, распределенного вдоль речевого тракта

$$a_k(t) = \frac{2\int_0^l G(x,t)S(x,t)\psi_k(x)dx}{l\int_0^l S(x,t)\psi_k^2(x)dx}, \qquad (4)$$

где k – номер временной моды акустических колебаний, $\psi_k(x)$ – k-я собственная функция волнового уравнения речевого тракта, $S(x,t)$ – площадь поперечного сечения тракта вдоль его средней линии с координатой x, t – время, l – длина речевого тракта.

Особенности взаимодействия турбулентного источника с акустическими колебаниями в речевом тракте можно качественно описать, предположив, что источник сконцентрирован лишь в одном месте с координатой x_0, и сам источник представлен как

$$G(x,t) = f(t)\delta(x - x_0), \qquad (5)$$

где δ – дельта-функция. Тогда парциальные возбуждения есть

$$a_k(t) = \frac{2}{l}S(x_0,t)\psi_k(x_0)G(x_0,t). \qquad (6)$$

Отсюда видно, что если источник возбуждения находится в узле собственной функции, где $\psi_k(x_0) = 0$, то и амплитуда возбуждения соответствующего резонанса равна нулю. Такой же эф-

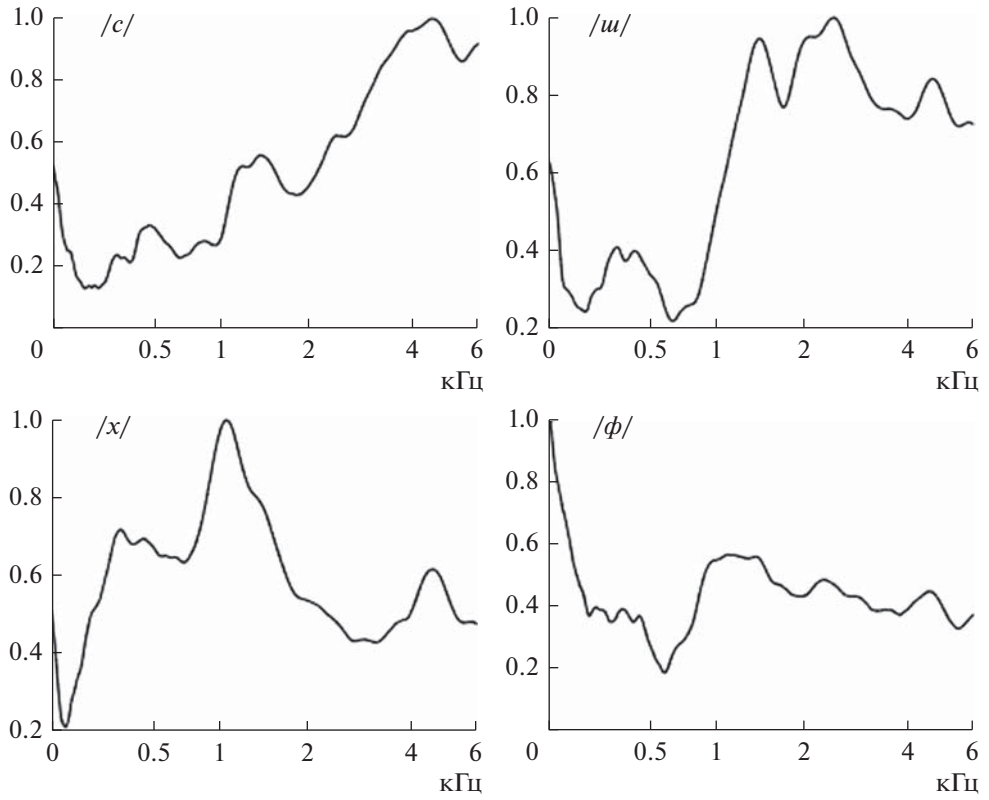

Рис. 1. Средние нормированные спектры изолированных фрикативных на фоне нейтральной артикуляции.

фект возникает и тогда, когда источник возбуждения не сосредоточен в одной точке, а распределен относительно нее так, что интеграл в числителе (4) равен нулю.

Этот эффект объясняет, почему спектры фрикативных согласных (кроме */ф/*) выглядят так, как если бы речевой сигнал был пропущен через полосовой фильтр (рис. 1).

Исследование перцептивных свойств фрикативных звуков на артикуляторном синтезаторе подтверждает эту модель взаимодействия турбулентного источника с волновыми процессами в речевом тракте [41]. На рис. 2 показана форма речевого тракта в средне-сагиттальной плоскости для русского фрикативного */х/* на фоне нейтрального гласного.

Предположим, что начало речевого тракта находится на выходе из голосовой щели. Тогда можно вычислить собственные функции волнового уравнения, задав граничные условия на голосовой щели и губах. Функция площади поперечного сечения и первые три собственные функции волнового уравнения для этого случая показаны на рис. 3.

Координата наибольшего сужения тракта равна 12.3 см, и источник турбулентного шума находится несколько дальше в сторону губ. В окрестности этого источника находится узел третьей собственной функции с частотой резонанса около 2035 Гц. Поэтому можно было бы ожидать, что

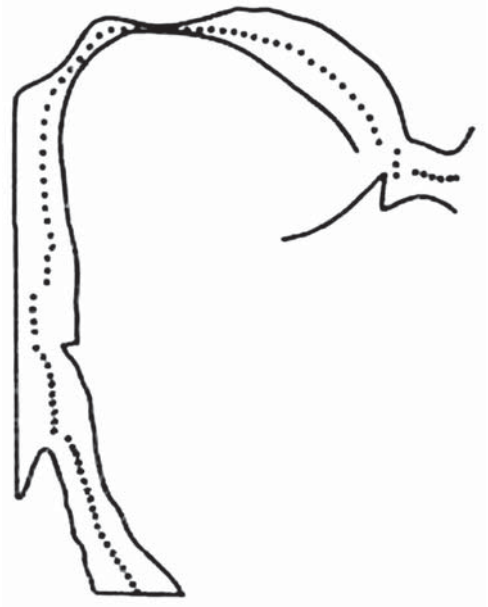

Рис. 2. Форма речевого тракта в средне-сагиттальной плоскости. (• • •) – средняя линия.

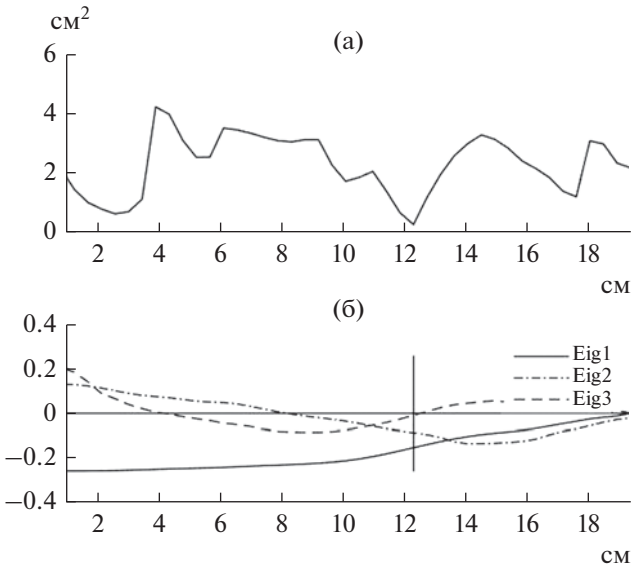

Рис. 3. (а) – Площадь поперечного сечения речевого тракта для фрикативного /*x*/, (б) – три собственные функции акустического давления, обозначенные как Eig1, Eig2, Eig3.

энергия в спектре речевого сигнала около этой частоты будет снижена по сравнению с энергией в области первого резонанса. Спектры фрикативных звуков на рис. 1 получены для того же диктора, для которого использовались измерения формы речевого тракта на рис. 2. Видно, что в спектре фрикативного /*x*/ в изолированном произнесении заметно подавлены частоты выше 2000 Гц, что соответствует модели взаимодействия турбулентного источника с речевым трактом.

Площадь голосовой щели при артикуляции фрикативных сопоставима с минимальной площадью речевого тракта, поэтому в действительности для фрикативных звуков нужно рассматривать речевой тракт не от голосовой щели до губ, а от легких до губ. Граничные условия со стороны легких очень сложны, и точные вычисления резонансов тракта затруднены. Однако качественно можно утверждать, что, поскольку длина речевого тракта для фрикативных увеличивается примерно вдвое, то и число резонансов в заданной полосе частот также увеличивается. Отсюда следует, что какова бы ни была форма спектра турбулентного источника, в спектре фрикативных должны присутствовать отклики довольно большого числа резонансов. Более того, форма спектра одного и того же фрикативного зависит от гласного, на фоне которого этот фрикативный формируется. Примеры такой зависимости будут приведены ниже. Это влияние обычно не принимается во внимание в известных сообщениях об измерениях спектра фрикативных, что приводит к искажению представлений о свойствах этих звуков.

3. МОДЕЛИ ДЕТЕКТОРОВ

3.1. Детекторы спектрально-временных неоднородностей

Скорость изменения формы речевого тракта зависит от степени участия разных артикуляторов, амплитудно-частотные характеристики которых различаются. Соответственно, и скорость переходных процессов от одного артикуляторного состояния к другому различна. Скорость изменения акустических параметров речевого сигнала нелинейно зависит от скорости артикуляторных движений. Так, скорость изменения резонансных частот зависит от площади минимального сужения в речевом тракте и может быть больше скорости артикуляторных движений [30]. Развитие турбулентного шума у фрикативных происходит значительно медленнее, чем у импульсного источника в момент взрыва смычных согласных. Кратковременная пауза длительностью в несколько миллисекунд между фрикативным и последующим гласным звуком возникает в процессе сведения голосовых складок, когда турбулентный шум уже прекратился, а колебания складок еще не начались. Отсюда следует, что детектор смены артикуляторных состояний должен учитывать разные длительности речевых сегментов и разные скорости переходных процессов. Иными словами, свойства этого детектора при переходе из состояния a_i в состояние a_j зависят, как минимум, от пары (i, j).

В [29] был предложен оператор, который позволяет моделировать неспецифические детекторы спектрально-временных неоднородностей в речевом сигнале, учитывая эффекты временной и частотной маскировки и разнообразие длительности переходных процессов. Эти детекторы неспецифичны в том смысле, что они реагируют на любые изменения в спектре речевого сигнала, и их отклик не связан однозначно с конкретным переходным процессом. В исходной формулировке этот оператор описывается как

$$D(\omega,t) = \lg\frac{S(\omega+\Delta\Omega,\theta_1,t\pm\Delta T_1,\tau_1)+C}{S(\omega-\Delta\Omega,\theta_2,t\mp\Delta T_2,\tau_2)+C}. \qquad (7)$$

Здесь S – динамический амплитудный спектр, ω – частота, t – время, C – некоторая константа. Параметры τ_1 и τ_2 есть постоянные времени сглаживающего фильтра, $\tau_2 \geq \tau_1$. Параметры θ_1 и θ_2 определяют ширину полосы частот, на интервале которой выполняется усреднение спектра в каждый момент времени, $\theta_2 \geq \theta_1$. Сдвиг отсчета времени относительно текущего момента t задается параметрами ΔT_1 и ΔT_2. Знаки при ΔT_1 и определяют опережение или отставание отсчетов спектра в числителе и знаменателе (7). Сдвиг отсчета значения спектра относительно текущего значения ω задается параметром $\Delta\Omega$.

Асимптотические свойства (7) устанавливаются при нулевых значениях некоторых параметров. При $\tau_1 = 0$, $\tau_2 = 0$, $\Delta\Omega = 0$, $\theta_1 = 0$, $\theta_2 = 0$, и $\Delta T_1 \to 0$, $\Delta T_2 \to 0$, $C = 0$, оператор (7) вычисляет логарифмическую производную по времени:

$$D(\omega,t) = \lg S(\omega,t+\delta t) - \lg S(\omega,t-\delta t) \underset{\delta t\to 0}{=} 2\delta t\left[\lg S(\omega,t)\right]' = \lg\frac{S'(\omega,t)}{S(\omega,t)}, \quad (8)$$

поскольку второй член в (8) есть центральная разность, сходящаяся к производной $\partial S(\omega,t)/\partial t$ при $\delta t \to 0$. Здесь штрих означает производную по времени. В этом случае оператор $D(\omega,t)$ оказывается инвариантен к умножению спектра речевого сигнала на произвольную амплитудно-частотную характеристику, постоянную во времени. В общем случае инвариантность к характеристике канала не достигается вследствие присутствия аддитивных шумов, но их влияние заметно ослабляется. Аналогично, при других асимптотических условиях $D(\omega,t)$ представляет собой логарифмическую производную по частоте.

В настоящей работе свойства оператора (7) исследуются с использованием гребенки гамма-тон фильтров при $\Delta\Omega = 0$, $\theta_1 = 0$, $\theta_2 = 0$, $\Delta T_1 = 0$ и различных сочетаниях параметров τ_1, τ_2 и ΔT_2:

$$D(f_k,t) = \lg\frac{S(f_k,t_{\text{1}}) + C}{S(f_k,t \pm \Delta T_2,\tau_2) + C}, \quad (9)$$

где f_k – центральная частота k-го фильтра. В таком виде $D(f_k,t)$ в каждый момент времени фактически сравнивает спектры, сглаженные с малой и большой постоянной времени τ_1 или τ_2, причем отсчет спектра в знаменателе (9) производится либо с отставанием от текущего момента времени t, если $\Delta T_2 < 0$, либо с опережением при $\Delta T_2 > 0$. Сглаживание выполняется в виде решения обыкновенного дифференциального уравнения первого порядка, в котором сглаживание определяется параметрами τ_1 и τ_2. Параметры τ_1, τ_2 и ΔT_2 зависят от характеристик перехода между соседними сегментами речевого сигнала. Для быстрых переходов типа взрыва смычки τ_1 должно быть малым, а для длительных переходов, например, к фрикативному, достаточно большим, с тем, чтобы подавить быстрые флуктуации. Постоянная времени τ_2 обычно должна быть в несколько раз больше, чем τ_1. Сдвиг по времени $\Delta T_2 > (3...6)\tau_1$, если $\Delta T_2 < 0$, и $\Delta T_2 > (3...6)\tau_2$, если $\Delta T_2 > 0$. При $\Delta T_2 < 0$ детектор реагирует на возрастание амплитуды в каждой частотной полосе, а при $\Delta T_2 > 0$ – на ее спад. Предварительные эксперименты с триадами типа */пауза–фрикативный–гласный/* привели к следующей оценке параметров: $\tau_1 = 2$, 5 или 15 мс, $\tau_2 = 15$ или 25 мс, $\Delta T_2 = 0$, 10, 25 или 45 мс.

Значения $D(f_k,t)$ могут быть как положительными, так и отрицательными. Поэтому для каждого набора параметров (τ_1, τ_2, ΔT_2) формируются три типа первичных детекторов, реагирующих только на возрастание амплитуды, только на спад амплитуды и на любое изменение амплитуды:

$$D_{\text{up}}(f_k,t) = 0, \quad D(f_k,t) \le 0, \qquad D_{\text{up}}(f_k,t) = D(f_k,t), \quad D(f_k,t) > 0, \quad (10)$$

$$D_{\text{down}}(f_k,t) = 0, \quad D(f_k,t) \ge 0, \qquad D_{\text{down}}(f_k,t) = D(f_k,t), \quad D(f_k,t) < 0, \quad (11)$$

$$D_{\text{change}}(f_k,t) = D(f_k,t). \quad (12)$$

Этим детекторам соответствуют амплитудные детекторы, оценивающие суммарное по всем частотам изменение амплитуды

$$A_{\text{up}}(t) = \sum_k D_{\text{up}}(f_k,t), \quad A_{\text{down}}(t) = \sum_k D_{\text{down}}(f_k,t), \quad A_{\text{change}}(t) = \sum_k \left|D_{\text{change}}(f_k,t)\right|. \quad (13)$$

Рис. 4 иллюстрирует некоторые свойства такой системы детекторов. Здесь для слова */семь/*, произнесенного мужским голосом, показана реакция детекторов $A_{\text{up}}(t)$ для оператора $D_{\text{up}}(f_k,t)$ с параметрами $\tau_1 = 15$ мс, $\tau_2 = 25$ мс и $\Delta T_2 = -45$ мс и $A_{\text{change}}(t)$ для оператора $D_{\text{change}}(f_k,t)$ с параметрами $\tau_1 = 2$ мс, $\tau_2 = 15$ мс и $\Delta T_2 = -15$ мс. На этом рисунке вертикальные линии обозначают начало перехода от паузы к фрикативному */c'/* и начало перехода от фрикативного к гласному */e/*.

Множество амплитудных детекторов $A_{\text{up}}(t)$, $A_{\text{down}}(t)$, $A_{\text{change}}(t)$ и динамических спектров (детектограмм) $D_{\text{up}}(f_k,t)$, $D_{\text{down}}(f_k,t)$, $D_{\text{change}}(f_k,t)$ составляет основу для формирования специфических детекторов, реагирующих на переход из одного артикуляторного состояния в другое, т.е. детекторов артикуляторных событий.

3.2. Детекторы артикуляторных событий

В настоящей работе исследование свойств неспецифических детекторов и детекторов артикуляторных событий было выполнено на примере задачи сегментации начальных глухих и звонких фрикативных с последующим гласным.

Для того чтобы отличить начальные фрикативные речевого сигнала от шумовых компонент канала, необходимо построить временные и спектральные модели каждого фрикативного в данном языке во всех возможных фонетических окружениях. Свойства фрикативных исследовались в [3, 47–50]. Рассматривались такие параметры, как центр тяжести спектра, моменты спектра, частота максимума спектра, наклон спектра, длительность, средняя амплитуда и ди-

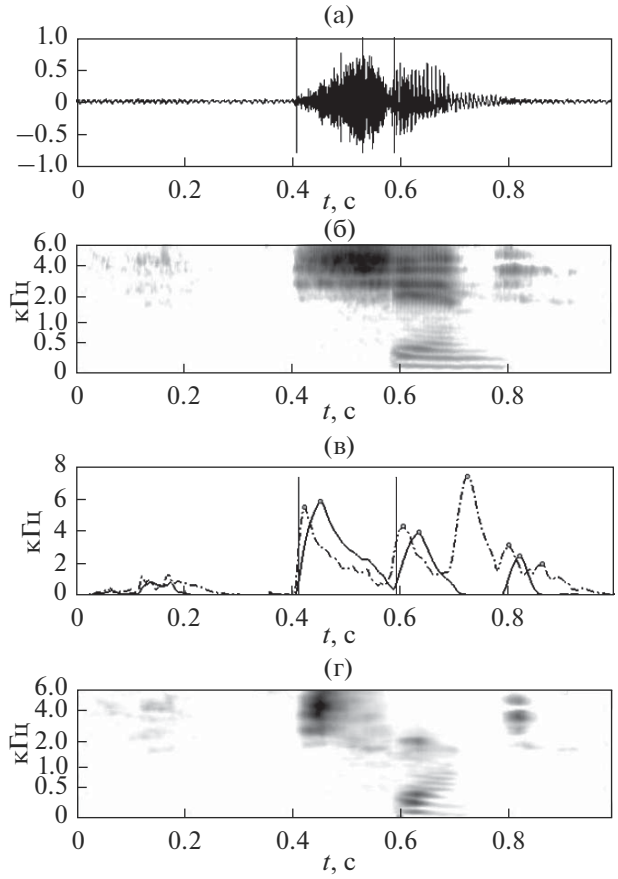

Рис. 4. Слово */семь/*, мужской голос. (а) – Осциллограмма звукового давления, нормированная к 1; (б) – логарифмическая гамма-тон спектрограмма; (в) – детекторы $A_{\text{up}}(t)$ (–) и $A_{\text{change}}(t)$ (–•–); (г) – гамма-тон спектрограмма отклика детектора $D_{\text{up}}(f_k,t)$ (“детектограмма”).

намика огибающей, частота второй форманты в момент возникновения фрикативного, а также уравнение локусов (предельных значений формантных частот на границе фрикативного) [51]. В [52] сообщается, что распознавание фрикативных по кепстральным коэффициентам обеспечивает меньшую ошибку, чем при использовании популярных признаков. Средний спектр сегмента фрикативного позволяет решить обратную задачу относительно формы речевого тракта с малой погрешностью [53].

Процесс турбулизации воздушного потока при артикуляции фрикативных развивается постепенно. На рис. 5 видно, что максимальная энергия в спектре фрикативного */ш/* в слове */шесть/* достигается лишь через 60 мс после начала турбулизации. В этот момент первый пик амплитуд спектра фрикативного */ш/* близок к частоте второй форманты последующего гласного, что определяется влиянием последующего гласного. Пик

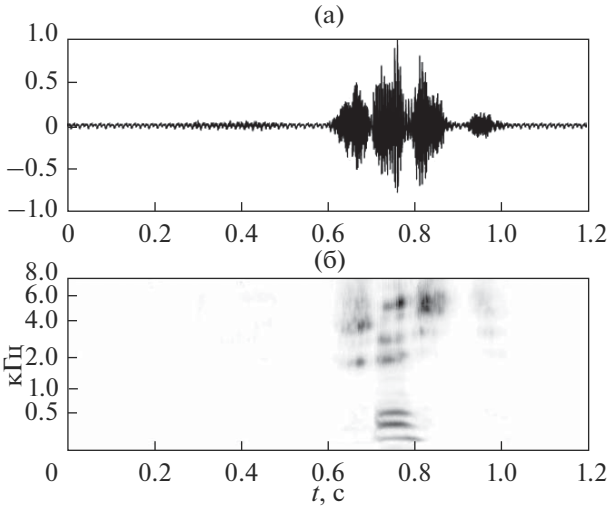

Рис. 5. (а) – Осциллограмма звукового давления и (б) – спектр слова */шесть/* в линейном масштабе амплитуд.

амплитуд в начале фрикативного */с'/* близок к частоте четвертой форманты предыдущего гласного, и это демонстрирует влияние предыдущего гласного на начало последующего фрикативного. Видны также следы влияния второй и третьей формант предыдущего гласного в виде пиков в спектре */с'/*.

На рис. 6 показаны нормированные средние спектры фрикативных */с, ш, ф, х/*, произнесенных одним диктором в слогах перед различными гласными. Отсчет спектра взят в момент наибольшей энергии фрикативного. Амплитуды спектра представлены в линейном масштабе.

Как видно из этого рисунка, форма спектра и положение максимального пика спектра отличаются большим разнообразием и зависят от последующего гласного. Это особенно заметно у фрикативного */х/*, где максимальный пик находится в диапазоне от 800 до 4000 Гц. Это явление еще раз указывает на то, что не существует акустических инвариантов для звука речи, которые в фонетической транскрипции и письменном тексте обозначаются одним и тем же символом. Физической основой этого явления служит различие в условиях турбулизации воздушного потока на фоне последующего (или предыдущего) гласного, а само место наибольшего сужения в речевом тракте (“место артикуляции”) может заметно сдвинуться под влиянием этих гласных. Так, место возникновения турбулентного шума мягкого */х'/* на фоне таких гласных, как */и, е, я, ю/* смещается в сторону губ, а его спектр становится высокочастотным. В этом случае невозможно различить */х/* и */с/* только по положению пика спектра. Необходимо знать, на фоне какого гласного генерируется фрикативный.

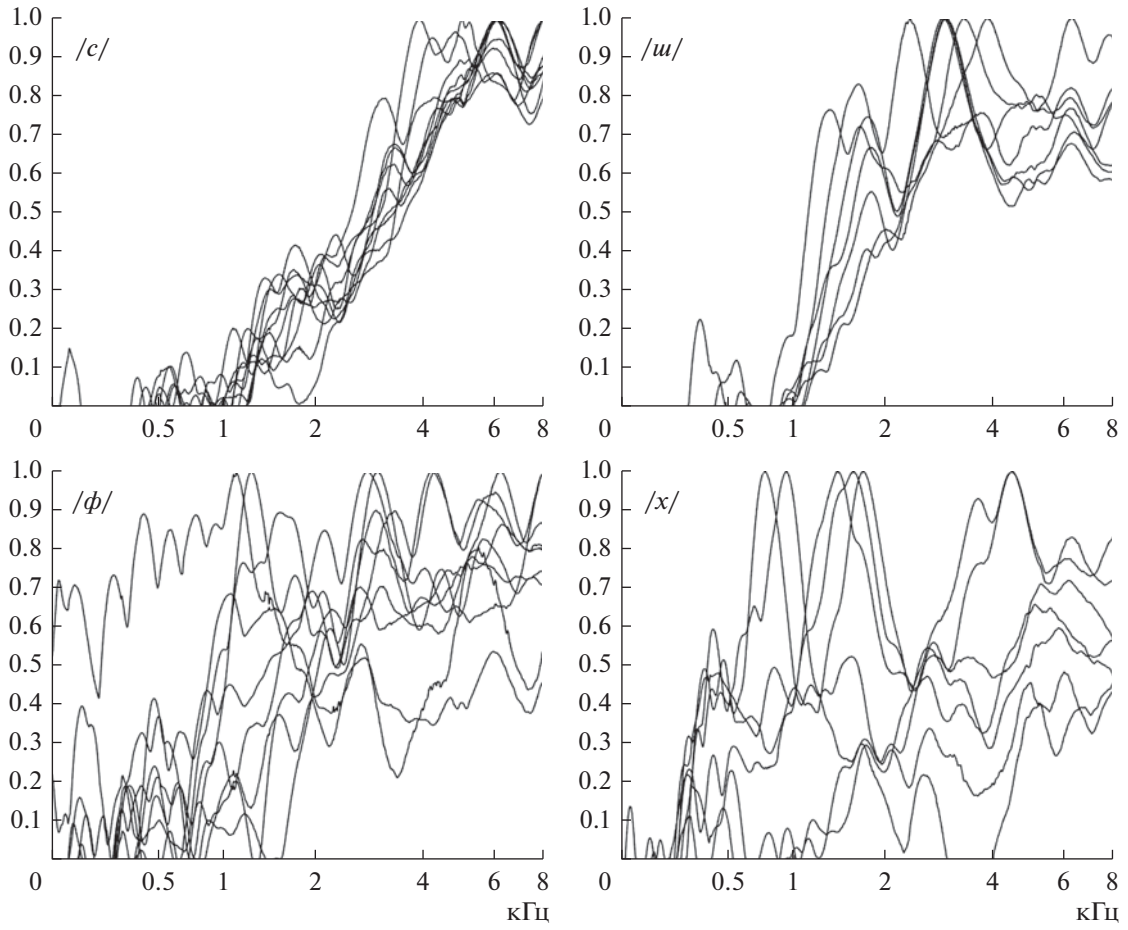

Рис. 6. Нормированные спектры фрикативных */с, ш, ф, х/* перед различными гласными.

Вследствие этих явлений описание спектра фрикативных различного рода функционалами представляется недостаточно эффективным. В [50] предлагалось использовать какую-либо меру сходства между вычисленным спектром фрикативного и, например, средним спектром, определенным на множестве реализаций фрикативного. Однако разнообразие спектров столь велико, что вместо среднего спектра целесообразно использовать некоторое множество характерных для данного фрикативного спектров. Эти характерные спектры можно найти различными способами, в том числе и с использованием иерархического метода *к*-средних (*k-means*) [54]. В этом алгоритме минимизируется сумма расстояний между элементами кластера и его центроидом. Обычно используется евклидова метрика L_2. Количество кластеров в нашей работе определяется итеративно, увеличивая их до тех пор, пока число элементов в каком-либо кластере не станет меньше заданного порога, например, 3% от общего числа реализаций. Центроиды кластеров принимаются за главные компоненты $\mu_m(f_k)$. Число найденных главных компонент m обычно равно 3 или 4. В качестве квазистатического описания спектра фрикативных используется сглаженный с постоянной времени 25 мс спектр в момент времени, когда достигается наибольшая энергия. Это предполагает, что к данному моменту процесс турбулизации воздушного потока полностью развился, и спектр шума наиболее характерен для данного фрикативного.

На рис. 7 показаны 3 компоненты для спектра сегмента */ш/* в слове */шесть/* и 4 компоненты для */с'/* в слове */семь/*, найденные, соответственно, по 12181 и 13552 реализациям этих фрикативных для 216 мужских голосов.

В экспериментах с женскими голосами принимали участие 177 женщин, 10619 реализаций слова */шесть/* и 6011 реализаций слова */семь/*. Спектры женских голосов и, соответственно, главные компоненты этих спектров отличаются от спектров мужских голосов и их главных компонент.

Мгновенный спектр переходного процесса от фрикативного к гласному $D_{\text{change}}(f_k, t)$ отсчитывается в момент максимума $A_{\text{change}}(t)$ с параметрами $\tau_1 = 2$ мс, $\tau_2 = 15$ мс, $\Delta T_2 = -15$ мс. Этот спектр зна-

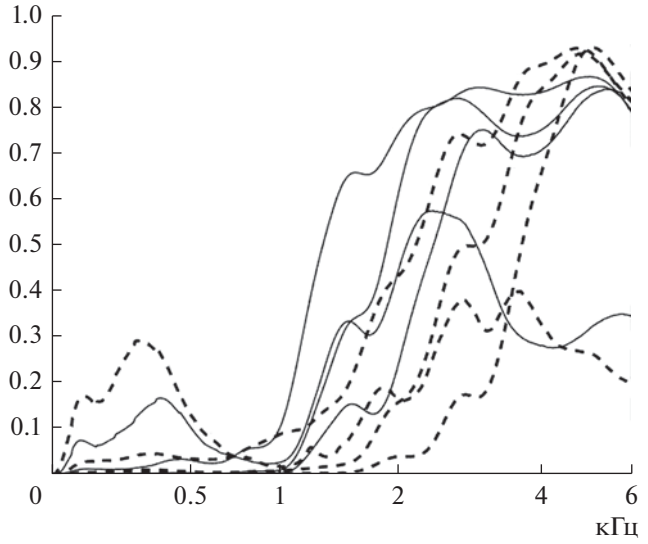

Рис. 7. Главные компоненты для спектра сегмента */ш/* в слове */шесть/* (–) и */с'/* в слове */семь/* (—).

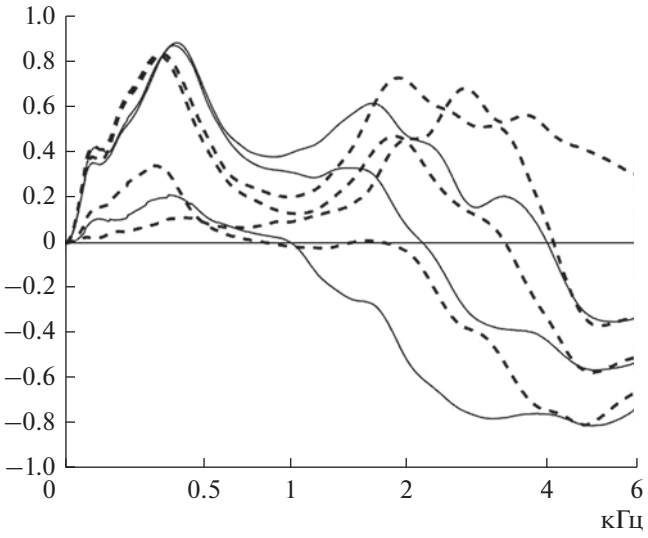

Рис. 8. Главные компоненты для спектра детектора перехода */шэ/* в слове */шесть/* (–) и */с'е/* в слове */семь/* (—).

копеременный. Его главные компоненты для перехода */шэ/* в слове */шесть/* и */с'е/* в слове */семь/* показаны на рис. 8.

Аналогичным образом находятся главные компоненты для спектра и детектограммы начальных сегментов слов английского языка */four, five, six, seven/*. База данных английских слов содержит голоса 22 мужчин и 26 женщин. Представительность исследованных сегментов значительно меньше, чем для числительных русского языка – около 200 реализаций для каждого слова.

4. СЕГМЕНТАЦИЯ И РАСПОЗНАВАНИЕ НАЧАЛЬНЫХ ФРИКАТИВНЫХ

Обучение детектора триады */пауза–фрикативный–гласный /* выполняется в два этапа. На первом этапе подбираются параметры неспецифических детекторов, наилучшим образом определяющих начало и конец фрикативного с использованием разметки слов по [55], и находятся главные компоненты спектра фрикативного в момент его максимального значения $\mu_{m1}^{\text{fr}}(f_k)$ между метками начала и конца фрикативного, а также главные компоненты $\mu_{m2}^{\text{pause/fr}}(f_k)$ в момент пика детектора перехода */пауза–фрикативный/* $A_{\text{up}}(t)$, и $\mu_{m3}^{\text{fr/vow}}(f_k)$ в момент пика детектора перехода */фрикативный–гласный/* $A_{\text{change}}(t)$. Эти пики выбираются среди множества пиков как ближайшие к моментам переходов по данным разметки. На втором этапе выполняется статистический анализ и автоматическое распознавание без поддержки разметки, которая иногда содержит грубые ошибки. В большинстве случаев эти ошибки удается ликвидировать на втором этапе обучения.

Момент перехода от паузы к фрикативному и момент перехода от фрикативного к гласному в словах */шесть, семь/* вычисляется как момент пересечения некоторой функцией $g(t)$ порога $\delta = 0.1$

$$g(t) = \frac{|h(t) - cA_{\max}|}{A_{\max}}, \qquad (14)$$

где $A_{\max}$ – амплитуда наибольшего пика детектора $A_{\text{up}}(t)$ или $A_{\text{change}}(t)$, функция $h(t)$ определена на интервале времени $[t_{\text{peak}} - \Delta, t_{\text{peak}}]$, t_{peak} – положение пика во времени. Для перехода */пауза–фрикативный/* $h(t) = A_{\text{up}}(t)$, и установлены параметры $c = 0.2$; $\Delta = 200$ мс, а для перехода */фрикативный–гласный/* $h(t) = A_{\text{change}}(t)$, и $c = 0.8$; $\Delta = 40$ мс.

На втором этапе обучения для сегментации используются собственные функции, вычисленные по разметке. Анализ начинается с определения момента перехода от фрикативного к гласному $T_{\text{fr/vow}}$, поскольку здесь можно использовать дополнительную информацию в виде перепада коэффициента автокорреляции от малых величин на фрикативном до больших на гласном. Выбирается такой пик $A_{\text{change}}(t_{f/v})$, что средний коэффициент автокорреляции K_{ac} на интервале $[t_{f/v}, t_{f/v} + 40$ мс$]$ $K_{\text{ac}} > 0.3$, а мера сходства между $D_{\text{change}}(f_k, t_{f/v})$ и множеством главных компонент для этого перехода $\mu_{m3}^{\text{fr/vow}}(f_k)$, найденных на первом этапе обучения детектора, больше 0.5. Для глухих фрикативных коэффициент автокорреляции K_{ac} вычисляется в полосе частот 70...500 Гц, а для звонких – в полосе 500...6000 Гц. При поиске момента начала

фрикативного после паузы $T_{\text{pause/fr}}$ используются только главные компоненты $\mu_{m2}^{\text{pause/fr}}(f_k)$. Для оценки меры сходства между спектром S фрикативного, спектром переходного процесса между паузой и фрикативным или спектром переходного процесса между фрикативным и гласным и соответствующими главными компонентами используется дискретная форма коэффициента Коши–Буняковского:

$$K_{\text{cb}}^{(m)} = \frac{\sum_{k=1}^{N} S(f_k)\mu_m(f_k)}{\sqrt{\sum_{k=1}^{N} S(f_k)^2 \sum_{k=1}^{N} \mu_m(f_k)^2}}, \qquad (15)$$

где $S(f_k)$ – спектр в соответствующий момент времени, $\mu_m(f_k)$ – m-я главная компонента, k – отсчет на частоте f_k, N – число гамма-фильтров. Для новых положений границ сегментов вновь вычисляются собственные функции.

На втором этапе обучения примерно в 20% случаев не определяется момент начала фрикативного из-за медленного развития процесса турбулизации, вследствие чего амплитуда пика детектора $A_{\text{up}}(t)$ оказывается ниже порога. Этот порог устанавливается отдельно для каждого фрикативного для минимизации ложных срабатываний на шумах во время паузы. Поэтому момент достижения максимальной энергии в спектре фрикативного определяется на интервале $[T_{\text{fr/vow}}-100$ мс, $T_{\text{fr/vow}}]$.

Контроль качества оценки моментов начала и конца фрикативного осуществлялся с использованием другой базы данных, в которой была выполнена ручная разметка речевых сигналов. В этой базе представлены записи голосов 24 мужчин и 15 женщин, произносивших разнообразные сочетания числительных русского языка. При этом использовались параметры триад, вычисленные только по исходной базе данных. Среднеквадратическая ошибка относительно ручной разметки для начала фрикативных составляет, в среднем, около 12 мс, а для момента перехода от фрикативного к гласному – около 5 мс. Учитывая тот факт, что ручная разметка выполнялась с погрешностью не менее 5 мс, можно признать точность автоматической разметки вполне удовлетворительной.

Часть ошибок приходится на такие варианты произнесения, что между глухим фрикативным и последующим гласным образуется пауза, длительность которой может доходить до 20 мс и более. Этот эффект появляется в результате сближения голосовых складок от положения, оптимального для обеспечения условий турбулизации потока в речевом тракте, в положение, необходимое для начала автоколебаний голосовых складок. В определенный момент скорость воздушного потока уже недостаточна для возникновения турбулентного шума, тогда как колебания голосовых складок еще не начались. В результате искажаются характеристики детектора в процессе обучения и появляются ошибки распознавания. В таких случаях вместо триады */пауза–фрикативный–гласный/* следует рассматривать последовательность */пауза–фрикативный–эпентетик–гласный/*, где термин "*эпентетик*" обозначает именно такие короткие паузы.

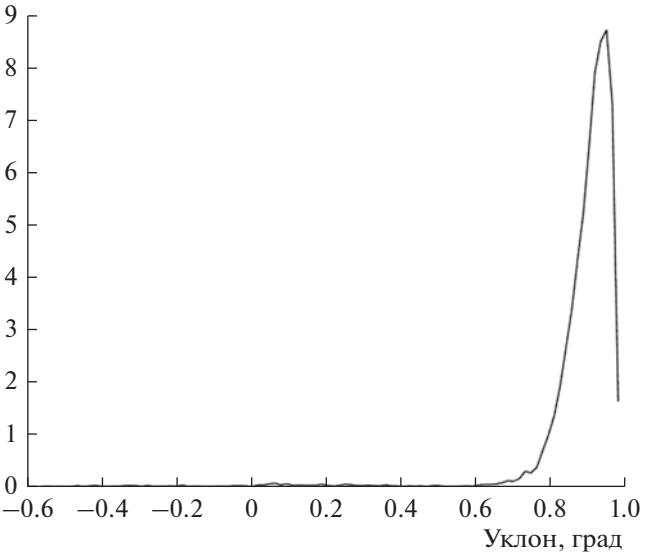

Рис. 9. Нормированная гистограмма коэффициентов Коши–Буняковского $K_{\max}$ для спектра детектора перехода */шэ/* в слове */шесть/*.

Мера сходства между спектрами фрикативного или перехода от паузы к фрикативному или от фрикативного к гласному и соответствующими главными компонентами определяется как максимальное значение $K_{\text{cb}}^{(m)}$ по всем m, $K_{\max} = \max(K_{\text{cb}}^{(m)})$. На рис. 9 показана типичная функция меры сходства для переходов от фрикативного */ш/* к гласному */э/*, для которой доля реализаций с максимальным значением коэффициента Коши–Буняковского $K_{\max}$, меньшего 0.5, составляет менее 1%.

Мера сходства неизвестной триады речи с каждой триадой типа i, для которого известны главные компоненты $\mu_{m1}^{\text{fr}}(f_k)$, $\mu_{m2}^{\text{pause/fr}}(f_k)$ и $\mu_{m3}^{\text{fr/vow}}(f_k)$, определяется как отношение правдоподобия между максимальными значениями $K_{\max}$ для переходов и спектра фрикативного, или по плотностям распределения $p(K_{\max})$:

$$L_K^{(i)} = \max\left(\frac{K_{\max}^{(i)\text{fr}}}{K_{\max}^{\text{fr}}}, \frac{K_{\max}^{(i)\text{pause/fr}}}{K_{\max}^{\text{pause/fr}}}, \frac{K_{\max}^{(i)\text{fr/vow}}}{K_{\max}^{\text{fr/vow}}}\right), \qquad (16)$$

Таблица 1. Вероятность распознавания (%), мужские голоса, английский язык. Критерий L_K

Мужчины	*si /six/*	*se /seven/*	*fo /four/*	*fa /five/*	θ */three/*
si /six/	100	6.3	0	0	0
se /seven/	3	100	0.4	1.3	0.4
fo /four/	0	0.5	100	1.9	0
fa /five/	0	0	1.5	100	0
θ */three/*	3.8	1.5	0.8	0	100

Таблица 2. Вероятность распознавания (%), женские голоса, английский язык. Критерий L_K

Женщины	*si /six/*	*se /seven/*	*fo /four/*	*fa /five/*	θ */three/*
si /six/	100	6.3	0	0.8	0.9
se /seven/	3	100	0.4	1.3	0.4
fo /four/	0	0.5	100	1.9	0
fa /five/	0	0	1.5	100	0
θ */three/*	0.9	1.9	0	0.9	100

$$L_p^{(i)} = \max\left(\frac{p(K_{\max}^{(i)\mathrm{fr}})}{p(K_{\max}^{\mathrm{fr}})}, \frac{p(K_{\max}^{(i)\mathrm{pause/fr}})}{p(K_{\max}^{\mathrm{pause/fr}})}, \frac{p(K_{\max}^{(i)\mathrm{fr/vow}})}{p(K_{\max}^{\mathrm{fr/vow}})}\right). \quad (17)$$

Например, при сравнении триад */пауза–ш–э/* и */пауза–с'–е/*, отношение правдоподобия по коэффициентам Коши–Буняковского есть

$$L_K^{(\text{ш э})} = \max\left(\frac{K_{\max}^{(\text{ш э})\mathrm{fr}}}{K_{\max}^{(c'e)\mathrm{fr}}}, \frac{K_{\max}^{(\text{ш э})\mathrm{pause/fr}}}{K_{\max}^{(c'e)\mathrm{pause/fr}}}, \frac{K_{\max}^{(\text{ш э})\mathrm{fr/vow}}}{K_{\max}^{(c'e)\mathrm{fr/vow}}}\right). \quad (18)$$

Если $L_K^{(i)} > 1$ или $L_p^{(i)} > 1$, то неизвестная триада принимается как принадлежащая множеству триад типа *i*. Если для сравнения предъявлена триада, которая на самом деле принадлежит типу *i*, но $L_K^{(i)} \le 1$ или $L_p^{(i)} \le 1$, то регистрируется ошибка распознавания. Ошибки распознавания между исследованными триадами для разных мер сходства оказались близки.

Начало фрикативных типа */ф, х/* по пикам детекторов $A_{\mathrm{up}}(t)$ или $A_{\mathrm{change}}(t)$ иногда не определяется вследствие малого перепада энергии при переходе от паузы. В этом случае отношение правдоподобия вычисляется только по двум факторам – спектру фрикативного и спектру переходного процесса между фрикативным и гласным. Хотя момент начала фрикативного при этом не определяется, присутствие фрикативного уверенно детектируется.

В соответствии с общепринятой терминологией, обозначим показатель числа отказов как FRR (False Reject Rate), а FRR6 и FRR7, соответственно, как долю числа отказов (неправильного распознавания) триад из слов */шесть/* и */семь/*. Аналогично, обозначим показатель числа ошибок как FAR (False Accept Rate), а FAR67 и FAR76, соответственно, как долю числа ошибок (неправильного распознавания) триад из слов */шесть/* как */семь/*, и наоборот. Для мужских голосов из русской базы данных FRR6 = 0.03%; FRR7 = 0.4% и FAR67 = 2.2%, FAR76 = 1.8%, а для женских голосов – FRR6 = 0.025%; FRR7 = 0.0% и FAR67 = 1.4%, FAR76 = 1.9%.

Если в качестве критерия используется не отношение коэффициентов корреляции, а отношение плотности вероятностей, то для мужских голосов из русской базы данных FRR6 = 0.04%; FRR7 = 0.4% и FAR67 = 0.3%, FAR76 = 3.1%, а для женских голосов – FRR6 = 0.025%; FRR7 = 0.0% и FAR67 = 2.4%, FAR76 = 0.3%.

Матрицы оценок вероятности распознавания триад английских слов представлены в табл. 1, 2 с округлением до первого знака после запятой. Оценка меры сходства выполнялась по коэффициентам корреляции для главных компонент спектра фрикативного и спектра $D_{\mathrm{change}}(f_k, t)$ в моменты перехода от паузы к фрикативному и от фрикативного к гласному. Левый столбец в табл. 1 обозначает предъявленные элементы речи, а верхняя строка – распознанные.

Представляет интерес оценка ошибки распознавания триад с глухими и звонкими фрикативными с одним и тем же местом артикуляции. С этой целью была создана база данных для одного диктора (мужчины), который по 60 раз произнес слова */жесть, жертва, женщина, жэк, жерех, зенкер, зелень, зернь, зев, зеркало/*. Всего было произнесено 600 слов. В половине случаев запись речевого сигнала выполнялась через высококачественный направленный микрофон со встроенной звуковой картой, а в другой половине

Таблица 3. Вероятность распознавания (%), мужские голоса, русский язык. Критерий L_K

Мужчины	*шесть*	*семь*	*жэ*	*зе*
шесть	100	1.9	0.8	0.4
семь	1.2	100	0.3	0.1
жэ	2	0.8	100	0
зе	0	0	0	100

случаев – через микрофон ноутбука. Матрица вероятностей распознавания триад */пауза–фрикативный–гласный/* по критерию отношения максимальных коэффициентов корреляции для главных компонент спектра фрикативного и спектра $D_{\text{change}}(f_k, t)$ в моменты перехода от паузы к фрикативному и от фрикативного к гласному приведена в табл. 3.

Ошибки распознавания, представленные в табл. 1–3, получены на тех же произнесениях, на которых были определены главные компоненты, т.е. фактически на тренировочной базе данных. Обычно считается, что в этом случае ошибки занижены. Это безусловно справедливо для ограниченного объема данных, хотя само понятие ограниченности не определено. В нашем случае количество числительных */шесть, семь/* для мужских голосов более 12000, а для женских – более 10000 и 6000. Поэтому объем тренировочной базы весьма велик, что позволяет рассматривать полученные оценки ошибок как достаточно близкие к действительным. Тем не менее, для контроля полученных оценок были выполнены эксперименты по распознаванию начальных триад в словах */шесть, семь/* на базе данных для 49 дикторов. Выше для этой базы были приведены оценки ошибок определения моментов начала и конца фрикативных. Количество каждого слова в этой базе порядка 400. Средняя вероятность принять триаду */пауза–ш–э/* за триаду */пауза–с'–е/* составляет около 0.9%, а вероятность принять триаду */пауза–с'–е/* за триаду */пауза–ш–э/* – менее 0.5%. Таким образом, ошибки распознавания на контрольной выборке оказались даже меньше, чем на тренировочной базе.

5. ОБСУЖДЕНИЕ

Реализация представлений о наличии в слуховой системе человека детекторов спектрально-временных неоднородностей для автоматического распознавания элементов речи оказывается весьма продуктивной. При таком подходе на исследованных триадах типа */пауза–фрикативный–гласный/* не требуется определения формантных частот, сопровождающегося значительными ошибками. Фонетическая транскрипция триады формируется уже после распознавания этой триады. Даже в случае неопределенности решения при близких значениях отношения правдоподобия для разных типов фрикативных и гласных, когда фонетическая транскрипция невозможна, на одном из уровней речевого кода сохраняется последовательность */пауза–фрикативный–гласный/* [30].

Согласно опубликованным данным, вероятность ошибки распознавания глухих фрикативных по спектральным признакам находится в области 20% [50]. Полученные в данной работе оценки вероятностей распознавания триад */пауза–фрикативный–гласный/* оказываются на порядок меньше, что свидетельствует о предпочтительности совместного использования статических и динамических параметров сегментов речи. Эффективность описанного метода проявляется в очень малых ошибках распознавания триад с одним и тем же фрикативным, артикулируемым на фоне разных гласных, а также триад, различающихся только участием голосового источника на сегменте фрикативного. Чувствительность системы детекторов артикуляторных событий к особенностям элементов речи приводит к тому, что объединение статистик для мужских и женских голосов увеличивает ошибки распознавания. Это означает, что обучение детекторов следует производить отдельно для голосов дикторов разного пола. При распознавании речи можно либо сравнивать отклики детекторов для разного пола, либо сначала определить пол диктора.

Некоторые начальные глухие фрикативные, такие, как */ф, х/*, часто имеют низкий уровень. При использовании детекторов начала речи, основанных на перепаде энергии, это может привести либо к пропуску начального сегмента, либо к ложному срабатыванию на шумах канала. Детекторы артикуляторных событий позволяют решить проблему начала речевого сигнала даже для фрикативных с малой энергией. Неспецифические детекторы типа $A_{\text{up}}(t)$ и $A_{\text{change}}(t)$, с одной стороны, чувствительны к малым перепадам энергии, а с другой стороны, довольно помехоустойчивы. Для подавления шумов при поиске начала слабого фрикативного нужно использовать дополнительные признаки, такие как распределение амплитуд детекторов $A_{\text{up}}(t)$ и $A_{\text{change}}(t)$, интервалы времени между максимумами детекторов $A_{\text{up}}(t)$ и $A_{\text{change}}(t)$, и длительность фрикативного.

Взаимное влияние процессов артикуляции (так называемая коартикуляция) приводит к тому, что акустические характеристики элементов речевого потока зависят как предыдущего, так и от последующего элементов. Как было показано в данной работе, это явление оказывается решающим для формирования детекторов последовательности */пауза–фрикативный–гласный/*. Поэтому единицей принятия решения о составе речевого сигнала должна быть последовательность из трех элементов (триада). В [50] число элементов речи на акустическом уровне было оценено величиной примерно в 127 единиц. Это, однако, не означает, что необходимо сформировать множество детекторов триад, равное перестановкам из 127, поскольку далеко не все последовательности элементов физически реализуемы, а многие редко встречаются в речи дикторов. И все же число детекторов оказывается довольно большим. Некоторое время назад этот фактор был бы решающим из-за недостаточной мощности компьютеров. К счастью, современные вычислительные средства вполне позволяют одновременно обрабатывать тысячи потоков, каждый из которых соответствует реакции детектора артикуляторных событий.

Основная трудность в создании такой системы детекторов состоит в необходимости использования предварительной разметки речевого сигнала на первом этапе обучения для достаточно большого числа дикторов. Ручная разметка очень трудоемка, а полностью автоматические средства сегментации речевого сигнала страдают заметными ошибками. Выход из этой ситуации состоит в разработке метода формирования детекторов не для каждой конкретной триады, а для однородного класса триад. Метод, описанный в данной статье, пригоден для формирования множества детекторов для класса триад */пауза–фрикативный–гласный/* с произвольным сочетанием фрикативных и гласных без использования ручной разметки. Аналогично, для другого класса триад достаточно создать алгоритм на основе анализа свойств небольшого числа представителей этого класса и ограниченной выборки дикторов с использованием ручной или грубой автоматической разметки. Критерием для выбора типа неспецифических детекторов на этом этапе служит погрешность автоматической разметки относительно ручной разметки.

6. ЗАКЛЮЧЕНИЕ

Оправдывается предположение о том, что использование совокупности детекторов амплитудных и спектрально-временных модуляций в речевом сигнале перспективно для создания принципиально нового описания структуры речевого потока. Ошибки распознавания близких по акустическим свойствам речевых сегментов могут быть чрезвычайно малы. Конкретный алгоритм, описанный в работе, пригоден для распознавания класса последовательностей типа */пауза–фрикативный–гласный/*, а сам метод может быть распространен на описание произвольных элементов речи, состоящих из последовательности трех артикуляторных состояний.

СПИСОК ЛИТЕРАТУРЫ

1. *Furui S.* On the role of spectral transition for speech perception // J. Acoust. Soc. Am. 1986. V. 80. P. 1016–1025.
2. *Stevens K.N.* Evidence for the role of acoustic boundaries in the perception of speech sounds // In: Phonetic Linguistics: Essays in Honor of Peter Ladefoged, edited by *Fromkin V.A.* (Academic, Cambridge, MA). 1985. P. 243–255.
3. *Stevens K.N.* Acoustic Phonetics // MIT Press, Cambridge, MA. 2000.
4. *Liu S.A.* Landmark detection for distinctive feature-based speech recognition // J. Acoust. Soc. Am. 1996. V. 100. P. 3417–3430.
5. *Kirchhoff K., Finkard G.A., and Sagerer G.* Combining acoustic and articulatory feature information for robust speech recognition // Speech Commun. 2002. V. 37. P. 303–319.
6. *Hasegawa-Johnson M., Baker J., Borys S., Chen K., Coogan E., Greenberg S., Juneja A., Kirchhoff K., Livescu K., Mohan S., Muller J., Sonmez K., and Wang T.* Landmark-based speech recognition: Report of the 2004 Johns Hopkins summer workshop // In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP). 2005. V. 1. P. 1213.
7. *Juneja A. and Espy-Wilson C.* A probabilistic framework for landmark detection based on phonetic features for automatic speech recognition // J. Acoust. Soc. Am. 2008. V. 123(2). P. 1154–1168.
8. *Jansen A., Niyogi P.* Modeling the temporal dynamics of distinctive feature landmark detectors for speech recognition // J. Acoust. Soc. Am. 2008. V. 124(3). P. 1739–1758.
9. *He D., Lim B.P., Yang X., Hasegawa-Johnson M., Chen D.* Acoustic landmarks contain more information about the phone string than other frames for automatic speech recognition with deep neural network acoustic model // J. Acoust. Soc. Am. 2018. V. 143(6). P. 3207–3218.
10. *Seifritz E., Esposito F., Hennel F., Mustofi C.H., Neuhoff J.G., Bilecen D., Tedeschi G., Scheffler K., Di Salle F.* Spatiotemporal pattern of neural processing in the human auditory cortex // Science. 2002. V. 297. № 5587. P. 1706–1708.
11. *Delgutte B., Kiang N.Y.S.* Speech coding in the auditory nerve: I. Vowel-like sounds // J. Acoust. Soc. Am. 1984. V. 75. № 3. P. 866–878.
12. *Sinex D.G.* Auditory nerve fiber representation of cues to voicing in syllable-final stop consonants // J. Acoust. Soc. Am. 1993. V. 94. № 3. P. 1351–1362.
13. *Бибиков Н.Г.* Описание признаков звука нейронами слуховой системы наземных позвоночных. М: Наука, 1987.
14. *Joris P.X., Yin T.C.* Responses to amplitude-modulated tones in the auditory nerve of the cat // J. Acoust. Soc. Am. 1992. V. 91. P. 215–232.

15. *Rhode W., Greenberg S.* Encoding of amplitude modulation in the cochlear nucleus of the cat // J. Neurophysiology. 1994. V. 71. P. 1797–1825.
16. *Wang K., Shamma S.A.* Spectral shape analysis in the central auditory system // IEEE Trans. Speech Audio Proc. 1995. V. 3. № 5. P. 382–394.
17. *Bibikov N.G., Nizamov S.V.* Temporal coding of low-frequency amplitude modulation in the semicircularis of the grass frog // Hearing Research. 1996. V. 101. P. 23–44.
18. *Moore B.C.J.* Auditory processing of temporal fine structure: Effects of age and hearing loss // World Scientific, Singapore. 2014. P. 1–182.
19. *Suga N.* Responses of inferior collicular neurons of bats to tone bursts with different rise time // J. Physiol. 1971. V. 217. P. 159–177.
20. *Shamma S.A., Fleshman J.W., Wiser P.R., Versnel H.* Organization of response areas in ferret primary auditory cortex // J. Neurophysiol. 1993. V. 69. P. 367–383.
21. *Kowalski N., Versnel Y., Shamma S.A.* Comparison of responses in the anterior and primary auditory fields of the ferret cortex // J. Neurophysiol. 1995. V. 73. P. 1513–1523.
22. *Kowalski N., Versnel Y., Raab D.H.* Forward and backward masking between acoustic clicks // J. Acoust. Soc. Am. 1961. V. 33. P. 137–139.
23. *Raab D.H.* Forward and backward masking between acoustic clicks // J. Acoust. Soc. Am. 1961. V. 33. P. 137–139.
24. *Elliot L.L.* Backward and forward masking of probe tones of different frequencies // J. Acoust. Soc. Am. 1962. V. 34. P. 1116–1117.
25. *Babkoff H., Sutton S.* Monaural temporal masking of transients // J. Acoust. Soc. Am. 1968. V. 44. P. 1373–1378.
26. *Wojtczak M., and Viemeister N.* Forward masking of amplitude modulation: Basic characteristics // J. Acoust. Soc. Am. 2005. V. 118. P. 3198–3210.
27. *Roverud E. and Strickland E.A.* The effects of ipsilateral, contralateral, and bilateral broadband noise on the mid-level hump in intensity discrimination // J. Acoust. Soc. Am. 2015. V. 138. P. 3245–3261.
28. *Jennings S.G., Chen J., Fultz S.E., Ahlstrom J.B., Dubno J.R.* Amplitude modulation detection with a short-duration carrier: Effects of a precursor and hearing loss // J. Acoust. Soc. Am. 2018. V. 143(4). P. 2232–2243.
29. *Sorokin V.N., Chepelev D.N.* Initial analysis of speech signals // Acoust. Phys. 2005. V. 51. № 4. P. 536–542.
30. *Сорокин В.Н.* Теория речеобразования. М.: Радио и связь, 1985.
31. *Чистович Л.А., Кожевников В.А. и др.* Речь. Артикуляция и восприятие. М.: Наука, 1965.
32. *Чудновский Л.С., Агеев В.М.* Расчет избирательных фильтров первичного анализа речевых сигналов // Акуст. журн. 2014. Т. 60. № 4. С. 407–412.
33. *Moore B.C.J., Glasberg B.R.* Suggested formulae for calculating auditory-filter bandwidths and excitation patterns // J. Acoust. Soc. Am. 1983. V. 74. P. 750–753.
34. *Patterson R.D., Holdsworth J.* A functional model of neural activity patterns and auditory images // Advances in Speech, Hearing and Language Processing. 1996. V. 3. P. 547–563.
35. *Yin H., Hohmann V., Nadeu C.* Acoustic features for speech recognition based on Gammatone filterbank and instantaneous frequency // Speech Communication. 2011. V. 53. P. 707–715.
36. *Fant G.* Acoustic Theory of Speech Production. Hague, The Netherlands: Mouton, 1960.
37. *Flanagan J.L.* Speech Analysis, Synthesis, and Perception. New York: Springer-Verlag, 1972.
38. *Shadle C.H.* The aerodynamics of speech // In: The Handbook of Phonetic Sciences, edited by *Hardcastle W.J. and Laver J.* (Blackwell Publishers Ltd., Malden, MA). 1997. V. 2. P. 33–64.
39. *Ohala J.J., Solé M.-J.* Turbulence and phonology // In: Turbulent sounds: An interdisciplinary guide, edited by *Fuchs S., Toda M., and Żygis M.* (De Gruyter Mouton, Berlin, Germany). 2010. V. 2. P. 37–102.
40. *Narayanan S. and Alwan A.* Noise source models for fricative consonants // IEEE Trans. Speech Audio Process. 2000. V. 8(3). P. 328–344.
41. *Сорокин В.Н.* Синтез речи. М.: Наука, 1992.
42. *Лойцянский Л.Г.* Механика жидкости и газа. М.: Наука, 1978.
43. *Блохинцев Д.И.* Акустика неоднородной движущейся среды. М.: Наука, 1981.
44. *Titze I.* Non-linear source-filter coupling in phonation: Theory // J. Acoust. Soc. Am. 2008. V. 123(5). P. 2733–2749.
45. *Alipour F., Schere R., Patel V.* An experimental study of pulsatile flow in canine languages // J. Fluids Engineering. 1995. V. 117. P. 577–581.
46. *Signorello R., Hassid S., Demolin D.* Toward an aerodynamic model of fricative consonants // J. Acoust. Soc. Am. 2018. V. 143. EL386.
47. *Chan C., Ng K.* Separation of fricatives from aspirated plosives by means of temporal spectral variation // IEEE Trans. Acoust., Speech Signal Process. 1985. V. 33(5). P. 1130–1137.
48. *Jongman A., Wayland R., Wong S.* Acoustic characteristics of English fricatives // J. Acoust. Soc. Am. 2000. V. 108(3). P. 1252–1263.
49. *Ali A.M.A., der Spiegel J.V.* Acoustic-phonetic features for the automatic classification of fricatives // J. Acoust. Soc. Am. 2001. V. 109(5). P. 2217–2235.
50. *Сорокин В.Н.* Речевые процессы. М.: Народное образование, 2012.
51. *McMurray B., Jongman A.* What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations // Psych. Review. 2011. V. 118. P. 219–246.
52. *Spinu L., Kochetov A., Lilley J.* Acoustic classification of Russian plain and palatalized sibilant fricatives: Spectral vs. cepstral measures // Speech Communication. 2018. V. 100. P. 41–45.
53. *Sorokin V.N.* Inverse problem for fricatives // Speech Communication. 1994. V. 14. № 2. P. 249–262.
54. *Seber G.A.F.* Multivariate Observations. New York, Wiley, 1984.
55. *Цыплихин А.И., Сорокин В.Н.* Сегментация речи на кардинальные элементы // Информационные процессы. 2006. Т. 6. № 3. С. 177–207.