# Отчёт об исследовании формантных синтезаторов: каскадный или параллельный?

Speech Communication 2 (1983) 251-273
North-Holland

RESEARCH REPORT
FORMANT SYNTHESIZERS: CASCADE OR PARALLEL?

J.N. HOLMES
Joint Speech Research Unit,
Received August 1983

**Аннотация**. В последние годы были широко распространены споры относительно достоинств каскадного и параллельного соединений формантных генераторов в синтезаторе речи. Этот доклад показывает, что теоретически менее привлекательное параллельное соединение способно производить более точное приближение к свойствам реальных сигналов речи, чем это вообще возможно для каскадного синтезатора, как для гласных, так и согласных. Однако, для достижения этого результата необходимо позаботиться о фазовых характеристиках формантных генераторов и надлежащим образом сформировать края кривых формантных откликов. Хотя для параллельного синтезатора необходима дополнительная информация об амплитуде, этот параметр акустической спецификации является в результате непосредственно связанным со свойствами человеческой речи и может быть легко измерен по изображению спектра.

**Ключевые слова**. Формантные синтезаторы, параллельное/каскадное соединение.

## 1. Введение

С тех пор, как управляемые динамически формантные синтезаторы впервые были использованы для синтеза речи [1,2], существуют противоположные точки зрения в пользу каскадного и параллельного соединения резонаторов. Недавно Клатт [3] в очень обстоятельном документе, по-видимому, соединил две стороны вместе описанием программного формантного синтезатора, который использует и каскадное, и параллельное соединение, выбирая одно или другое в зависимости от типа синтезируемого звука. Гласные и согласные, похожие на гласные, используют каскадное соединение, в то время как остальные согласные используют параллельное соединение. В этом докладе я буду доказывать, что при правильной реализации параллельная конфигурация на самом деле превосходна во всех существенных отношениях для гласных и согласных. Этот вывод прямо противоположен тому, который представлен Рабинером в 1968 г. [4].

По общему признанию сторонников каскадного соединения для гласных, чтобы иметь дело со многими согласными звуками должны быть обеспечены специальные условия [5,6]. Таким образом, полноценный синтезатор каскадного типа, как правило, гораздо более сложен, чем это необходимо только для гласных, и требуется значительное количество дополнительной управляющей информации для согласных. Если не уделяется особое внимание, использование отдельных резонаторных систем для различных типов звука может нарушить естественную непрерывность резонанса, как это происходит в человеческой речи на фонетических границах. Например, на переходе между гласной и фрикативным, таким как [s], энергия фрикативного начинает проявляться в высших формантах перед тем, как прекратится голосовое возбуждение, и это шипение постепенно сливается со спектральной структурой полностью

глухого [s]. Синтезатор Клатта даёт необходимую преемственность резонанса путём предоставления в нём двух резонаторных систем с одинаковыми параметрами резонанса (то есть каждый резонатор дублируется в каскадном и параллельном соединении).

Конечно, можно использовать каскадный синтез, чтобы сделать разумные приближения для спектров как гласных, так и согласных без отдельной системы для согласных, как это делается в линейном кодировании с предсказанием (linear predictive coding, LPC). В вокодерах LPC функция передачи синтезатора по своей сути содержит только полюсы в ***z*** области, используемые обычно для описания фильтров сэмплированных данных [7]. Её реализация по сути эквивалентна каскадному формантному синтезатору, получающему сэмплированные данные, в котором все формантные частоты и полосы могут быть выбраны с полной свободой. Полюсы такой системы могут рассматриваться как представляющие истинные формантные резонансы во время гласных, но некоторые из них, с гораздо большим затуханием, адаптированы к общей роли формирования огибающей спектра в других звуках для изменения интенсивности остальных значимых формант. Основным недостатком синтезаторов с LPC для исследования восприятия речи является то, что трудно связать требуемые параметры формант с информацией управления LPC. LPC синтезаторы не используются даже для применения в вокодерных приложениях настолько хорошо, как это могло бы быть, потому что математический критерий, применяемый в обычном анализе LPC для определения передаточной функции в синтезаторе, не очень хорошо подходит к потребностям человеческого слухового восприятия, и для улучшения этой ситуации Макхоулом и Коселлом [8] и Штрубе [9] были описаны более сложные методы анализа.

Для синтезаторов, которые работают с описанием формант, стоит изучать, какие характеристики могут быть достигнута с использованием только параллельной системы. Клатт цитирует две причины в пользу использования каскадного соединения для некоторых звуков. Первая: "относительные амплитуды формантных пиков для гласных сразу получаются правильными без необходимости иметь отдельные элементы управления амплитудой для каждой форманты", а другая в том, что эта конфигурация включает в себя "более точную модель функции передачи голосового тракта во время создания не назальных сонорных". В этой статье я покажу, что оба этих бесспорных теоретических преимуществ на самом деле не имеют значения на практике, и что простая конфигурация, которая может быть достигнута только с помощью параллельного соединения, фактически предлагает преимущество в характеристиках по сравнению с каскадным соединением, даже для гласных.

В ходе обсуждения ниже предполагается, как это обычно с терминальными аналоговыми синтезаторами, что целью является как можно ближе приблизиться к тем элементам речевых сигналов, которые значимы для восприятия, без собственной важности, связанной с человеческой речью, создаваемой механизмом. Кажется, принято считать, что для достижения этой цели достаточно воспроизвести кратковременный спектр речи, определяемый с помощью разрешения по частоте и времени, аналогичного человеческой слуховой системе.

В стационарном состоянии при произнесении человеком слов кратковременный спектр сигнала является результатом четырёх отдельных факторов:

(i) передаточной функции голосового тракта;
(ii) влияния излучения на губах и ноздрях;
(iii) спектра одного импульса объёмного потока в голосовой щели;
(iv) структурой спектральной линии, обусловленной периодичностью возбуждения.

Свойства возбуждения фрикативных и взрывных для непериодических источников звука

замещают факторы (iii) и (iv), а функция передачи голосового тракта зависит от положения точки возбуждения.

Основные проблемы в синтезе речи связаны с факторами (i) и (iii). Влияние излучения может быть хорошо представлено на большей части частотного диапазона речи с помощью простого дифференцирования [10], которое может быть выполнено напрямую, или его действие может быть объединено с другой функцией синтезатора. В терминальных аналоговых синтезаторах, по сути, вполне приемлемо объединить некоторые аспекты вышеприведённых факторов (ii) и (iii) в системе фильтров, формирующих спектр. Для нормального качества голоса и в каскадном, и в параллельном синтезаторах речи для получения источника голосового возбуждения используется периодический сигнал, а требование иметь не периодическое возбуждение для особых параметров голоса одинаково влияет и на каскадные, и на параллельные синтезаторы; по этой причине фактор (iv) не требует дальнейшего обсуждения в этом докладе.

## 2. Каскадное соединение в качестве модели неразветвлённого голосового тракта

Поскольку достоинства каскадных синтезаторов применяются специально для моделирования речевого тракта в течение не-назальных сонорных, обсуждение в данном разделе ограничено каскадным моделированием неразветвленного голосового тракта, возбуждаемого голосовой щелью. Другие факторы, которые влияют на синтетический спектр речи, рассматриваются в последующих разделах.

В настоящее время хорошо установлено [10], что если голосовой тракт рассматривается как неоднородная неразветвлённая акустическая труба, возбуждаемая исключительно на конце голосовой щелью, и излучает звук только со стороны рта, и если требуется такой диапазон частот, что передача звука в трубе может полностью рассматриваться в виде плоских волн, то передаточная функция имеет вид

$$H(s) = \prod_{n=1}^{\infty} \frac{S_n S_n^*}{(S - S_n)(S - S_n^*)} \qquad (1)$$

то есть она содержит только полюсы, соответствующие различным резонансным режимам модели голосового тракта. Среднее расстояние между полюсами в частотной области задаётся

$$d = c/2L \qquad (2)$$

где ***d*** является расстоянием между полюсами, ***L*** является длиной вокального тракта и ***c*** - скорость звука в тракте. Для голоса типичного взрослого мужчины это среднее расстояние составляет около 1 кГц. Хотя число полюсов бесконечно, существует несколько причин, почему резонансы выше 5 кГц имеют очень малое прямое влияние на выходной сигнал:

(i) затухание этих высших резонансов очень велико, в основном потому, что потери на излучение во рту гораздо больше для длин волн меньших, чем размер ротового отверстия;
(ii) спектральная плотность мощности гортанного источника звука выше 5 кГц очень мала;
(iii) и чувствительность, и разборчивость по частоте человеческого слухового восприятия выше 5 кГц значительно сокращается.

Вышеприведённые причины (i) и (ii) приводят к очень небольшой голосовой энергии, производимой выше 5 кГц, а причина (iii) придаёт спектральной структуре любого сигнала в этой области очень малое значение.

Однако, бесконечное число высокочастотных полюсов в уравнении (1) имеет весьма значительный кумулятивный эффект в области нескольких нижайших формант и в аналоговых каскадных синтезаторах обычно приближаются к этой цели используя лишь небольшое число явных резонаторов (обычно 4 или 5), а также используя схему "коррекции высшего полюса" [11], чтобы дать этим явным формантам необходимое усиление в области высокой частоты, которое обычно происходит в человеческой речи. Для этого обычно выбирают коррекцию верхнего полюса, который подходит для равномерной трубы той же длины, что и речевой тракт (например, для идеального нейтрального гласного). Если для модели 17 см голосового тракта используется пять явных формант, эта коррекция высшего полюса равна амплитуде отклика бесконечного множества резонаторов, частоты которых отстоят на 1 кГц друг от друга, начиная с 5.5 кГц. На высших частотах полосы речи величина этой коррекции очень велика (около 57 дБ на 5 кГц). В дискретных каскадных синтезаторах происходящее преобразование в ***z***-области в дискретных фильтрах обеспечивает бесконечный ряд полюсов в ***s*** плоскости, так что необходимо только обеспечить явные полюса в пространстве справа до половины частоты дискретизации (***Fs/2***) для достижения автоматической коррекции высшего полюса [4]. Существует, однако, небольшое теоретическое различие между аналоговым и дискретным методами, обеспечивающими получение характеристик коррекции высшего полюса, что может быть заметно, если пользователь синтезатора имеет контроль частот всех явных полюсов в передаточной функции. В дискретном случае любое движение частоты высшего из этих полюсов к или от ***Fs/2*** имеет соответствующий эффект, отображаемый выше ***Fs/2***, и, таким образом, приводит к изменению эффективной коррекции высшего полюса, вызывая соответственно подъём или падение уровня вблизи ***Fs/2***. Очевидно, что для такого изменения не может быть никакого теоретического обоснования. Существующая структура полюсов, порождаемая выше ***Fs/2***, связана с тем, что в эквивалентной модели акустической трубы дискретной всеполюсной сети [12] распространение плоской волны подразумевается на всех частотах, и эта труба имеет функцию области поперечного сечения (Рис. 1). Ни одно из этих предположений не является реалистичным; более того, при наиболее часто используемой частоте дискретизации 10 кГц в этой трубе есть только 10 секций. Рис. 2 показывает некоторые примеры частотной характеристики 5-ти формантного аналогового и дискретного синтезатора, использующего частоту сэмплирования 10 кГц, для иллюстрации различий, которые могут возникнуть между этими двумя формами коррекции высшего полюса для двух возможных конфигураций высших формант. На Рис. 2b видно, что на 5 кГц отклик дискретного на несколько децибел ниже характеристики аналогового, в то время как на Рис. 2d верно обратное. Конечно, при желании возможно близко приблизиться к 5-ти формантной аналоговой характеристике с помощью дискретной реализации, работающей на более высокой частоте дискретизации, например, 20 кГц, и имеющей форманты выше, чем F5, расположенные на 5.5 кГц, 6.5 кГц и так далее.


Рис. 1. Типичная функция поперечного сечения акустической трубы, применяемая в каскадном дискретном синтезаторе.

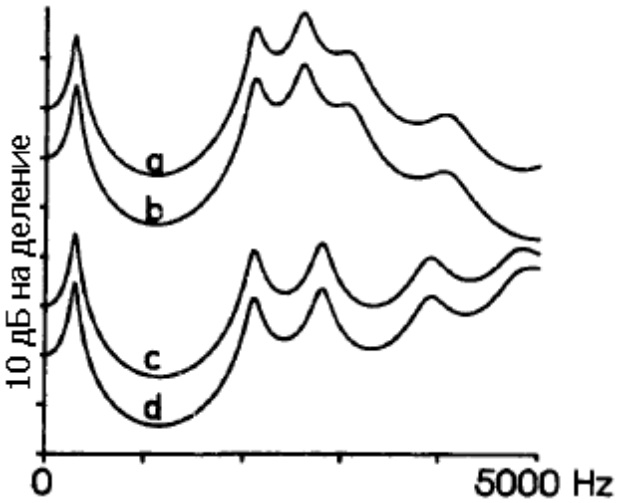



Рис. 2. (a) и (c) Типичные характеристики гласной для каскадного аналогового синтезатора с коррекцией высшего полюса, использующего разные частоты для F4 и F5;
(b) и (d), соответствующие характеристики для каскадного дискретного синтезатора, использующего частоту сэмплирования 10 кГц , показывают, как отклонение от аналогой характеристики зависит от частот F4 и F5.

Хотя Рис. 2 иллюстрирует возможные большие различия между аналоговой и дискретной коррекцией высшего полюса, нет никаких причин, почему каждый из них должен близко приблизиться к влиянию высших полюсов реального речевого тракта, особенно в экстремальных артикуляционных конфигурациях. Не учитываются ни фактические позиции частот этих высших полюсов, ни возможно большие изменения в их затухании. Разница в длине голосового тракта между артикуляциями с округлёнными и открытыми губами обычно также игнорируется. Сочетание всех этих эффектов может легко привести к ошибкам во много децибел в области около 3 - 4 кГц, а для тех гласных, у которых уровень в этой области обычно высок, последствия этого были бы восприняты как очень значительные.

Существует ещё одна причина неопределенности моделирования высокочастотных компонентов гласных. Форма сечения голосового тракта может быть довольно сложной во многих местах, с сечениями не менее 6 см в наиболее широкой части для некоторых гласных. Половина длины звуковой волны в голосовом тракте на 3 кГц составляет всего около 5 - 6 см, и, таким образом, следует ожидать, что отклонение от распространения плоской волны в точках, где сечение голосового тракта становится большим, будет заметно влиять на отклик на

3 кГц и выше. К сожалению, как только отклонение от распространения плоской волны должно быть принято во внимание, отклик тракта становится чрезвычайно трудно анализировать теоретически. Однако, некоторое представление вероятного порядка величины эффекта было продемонстрировано путём измерения в чрезвычайно простой акустической модели голосового тракта регулируемых размеров [13]. В этой модели воздушный путь имеет постоянную толщину 1.2 см. Общая длина зафиксирована на 17 см, а площадь поперечного сечения регулируется отдельно в двух 8.5 см секциях за счёт изменения расстояния до верхней и нижней стенки, как показано на Рис. 3. Толщина настолько мала, что модель можно считать 2-мерной по крайней мере до 8 кГц. Модель возбуждается разрядником, а излучаемый отклик был проанализирован спектрально. Рис. 4 показывает отклик для формы, грубо соответствующей гласной [e]. Были использованы два различных набора размеров сечения с соотношением размеров 3:1. Видно, что когда максимальный размер сечения был всего лишь 2 см, в отклике до 8 кГц было в общей сложности восемь формант, как и предсказывалось теорией акустической трубы с плоской волной. Для максимального размера 6 см три нижние форманты были очень похожи на те же для случая 2 см, но выше 3 кГц наблюдались несколько дополнительных резонансных режима, а также редкие глубокие провалы, вызванные антирезонансами.


Рис. 3. Простая акустическая модель, представляющая простейший голосовой тракт.

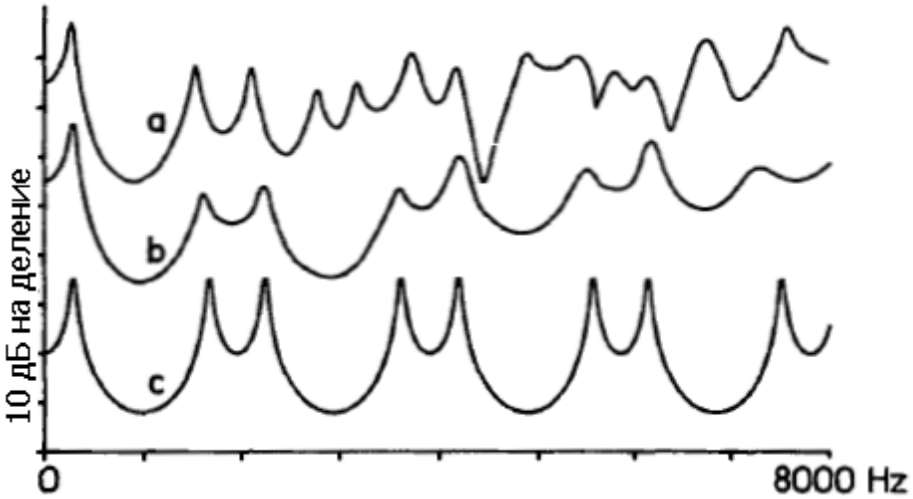



**Рис. 4. (a) Выходной спектр модели, показанной на Рис. 3, измеренный с помощью полосы анализа 50 Гц; максимальное сечение модели было 6 см. (b) Спектр, измеренный в (a), когда максимальное сечение было 2 см. (c) Расчётный отклик для модели идеальной акустической трубы с той же эффективной длиной и формы функции сечения как на Рис. 3, и всеми потерями со стороны рта.**

Из вышеизложенного видно, что для тех звуков, которые производятся каскадной моделью неразветвлённого голосового тракта, вероятно, будет возможно получить хорошее приближение к отклику голосового тракта до примерно 3 кГц, но что величина моделируемого отклика может существенно отличаться на более высоких частотах. Для слухового восприятия наиболее важные особенности в области 3 - 5 кГц - это уровни в областях спектра, примерно такие же широкие, как критические полосы [14]. Подходящие уровни могут быть достигнуты манипулированием позиций полюсов и большие изменения уровней, получаемых таким образом, хорошо иллюстрируются двумя наборами частот высших формант, использованных на Рис. 2. Раздел II.C Клатта [3] сообщает об использовании этого метода создания спектральных уровней выше 3 кГц, но нет очевидных причин, почему частоты полюсов должны быть такими же, как резонансные режимы реального речевого тракта. Тот факт, что предположение о плоской волне не является действительным выше 3 кГц для реалистичных размеров голосового тракта означает, что отклик в этой области будет в общем иметь больше резонансов, чем всеполюсная модель, а также антирезонансы, которые будут влиять на интенсивность сигналов на резонансных частотах. Поскольку этими резонансами двигают артикуляторы, следует ожидать движение сложным путём и антирезонансов, как иллюстрируется спектрограммой типичного естесвенного произношения, показанного на Рис. 5.

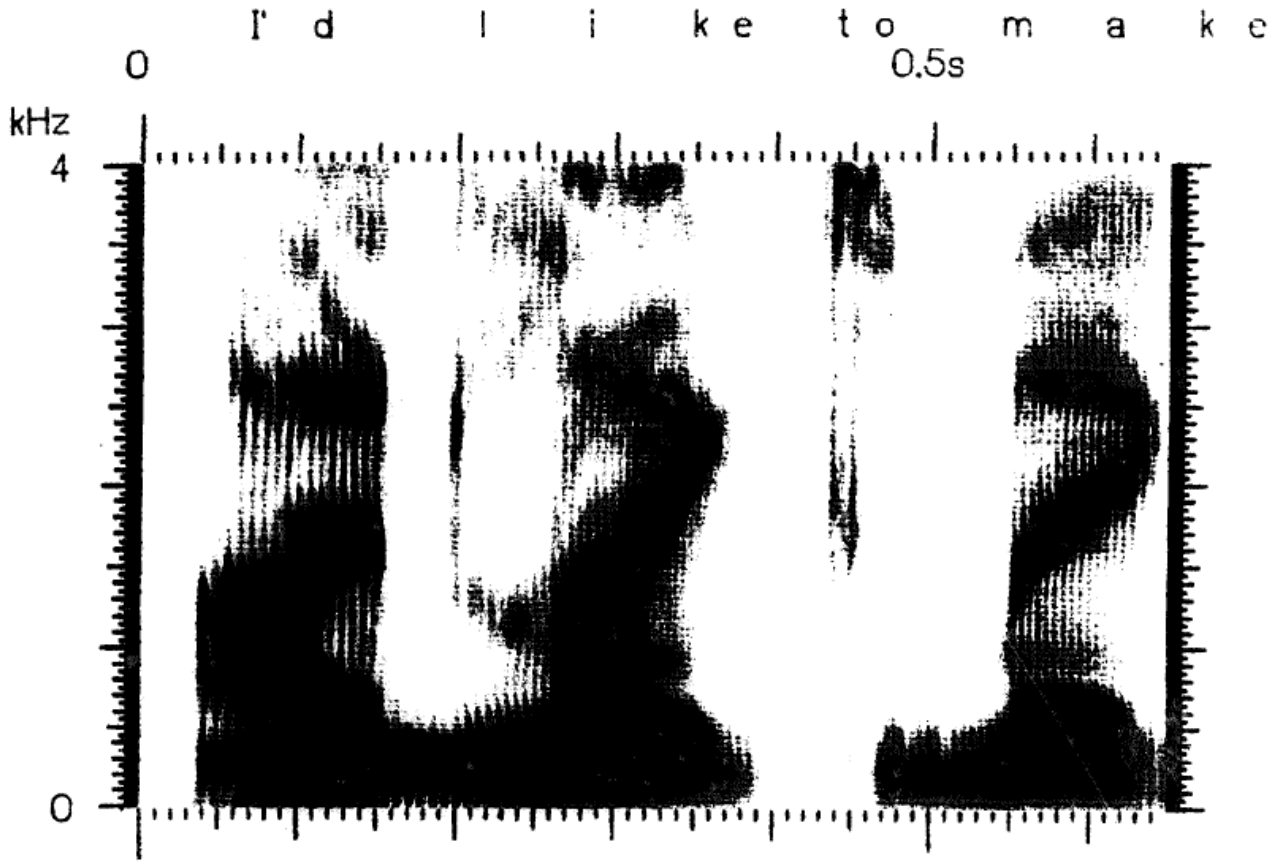



Рис. 5. Спектрограмма натуральной речи, иллюстрирующая сложную резонансную структуру между 3 кГц и 4 кГц.

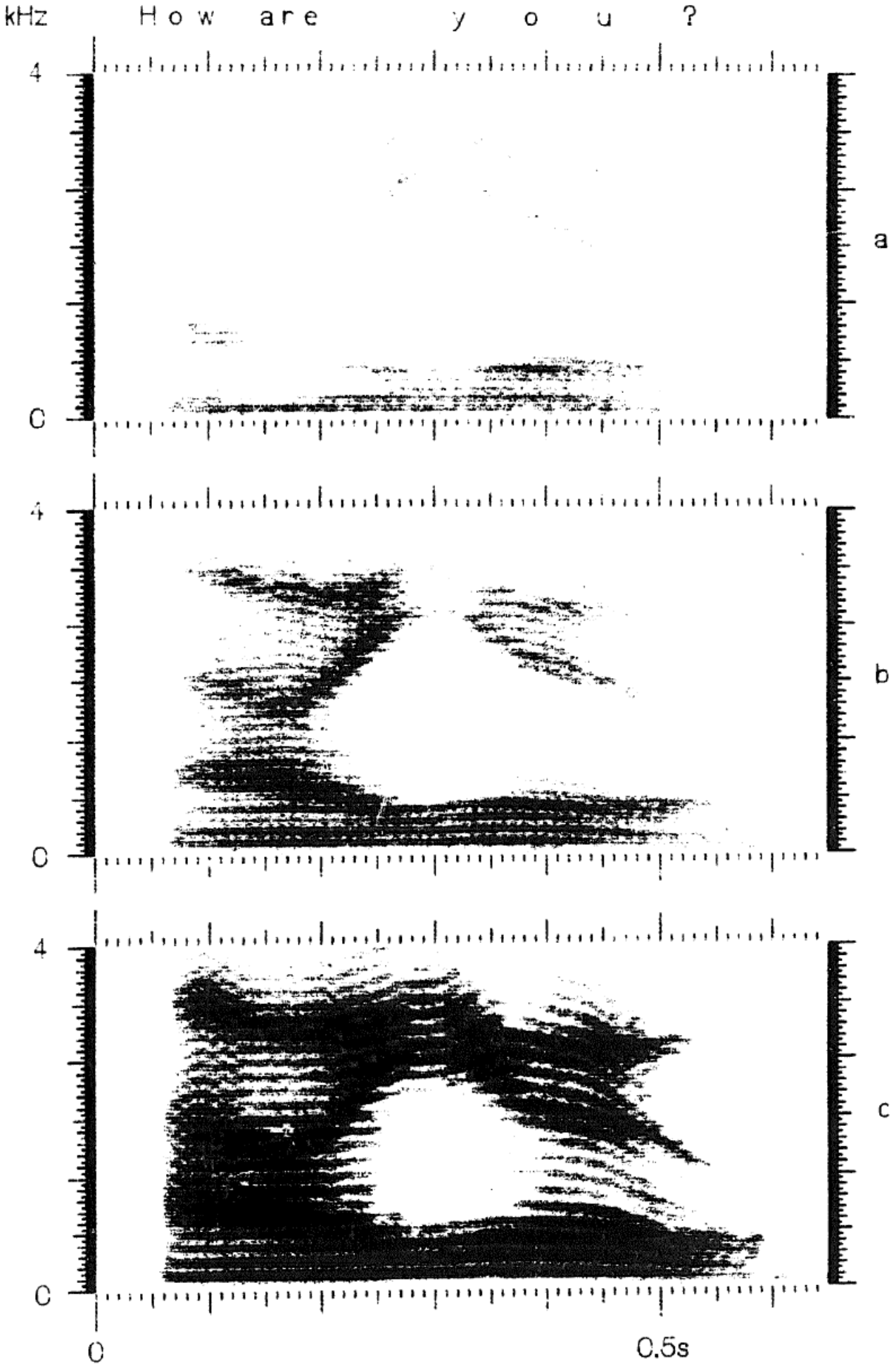
Критические зоны в этой части спектра шириной порядка 500 Гц, и, таким образом, мелкие детали сигнала в частотной области сами по себе, вероятно, не будут важны для восприятия. Однако, биения между компонентами спектральной картины в критической полосе приведут к сложной огибающей во времени в ответ на каждый импульс голосовой щели. Эта сложность временной структуры может быть субъективно обнаруживаема в некоторых обстоятельствах, но не может быть смоделирована с использованием расстояния между полюсами простого каскадного синтезатора.

## 3. Источник спектральных проблем в каскадных синтезаторах

В звуках голоса основным источником акустического сигнала возбуждения голосового тракта является квази-периодическая объёмная скорость воздуха, проходящего между голосовыми связками. Даже на пике воздушного потока в каждом цикле, когда складки находятся максимально далеко друг от друга, площадь голосовой щели значительно меньше, чем площадь поперечного сечения глотки, так что импеданс голосовой щели значительно выше, чем входной импеданс голосового тракта, за исключением, возможно, областей вблизи формантных частот. Таким образом, в первом приближении возбуждение вокального тракта можно рассматривать как практически не зависящее от конфигурации вокального тракта и определяемое только давлением под голосовой щелью и изменяющимися во времени свойствами голосовой щели. Предположение, что импеданс открытой голосовой щели достаточно высок, чтобы можно было

им пренебречь, однако, верно только для первого приближения, и существуют наблюдаемые эффекты, вызванные непосредственно импедансом голосовой щели и сопряжением подсвязочной системы [15 , 16], особенно в нижней части частотного диапазона речи. Конечно, эффективная объёмная скорость чуть выше гортани представляет собой истинное возбуждение голосового тракта, но конечный импеданс щели заставляет форму волны этой объёмной скорости быть изменённой свойствами акустической системы выше и ниже гортани [16,17] . Холмсом [18] было указано, что значительное возбуждение голосового тракта также возможно воздухом, перемещаемым в результате движения поверхности голосовых связок, также как и воздухом действительно протекающим через голосовую щель. Конечным результатом всех вышеописанных эффектов является то, что детальный спектр голосового возбуждения может значительно отличаться от стилизованного импульса потока голосовых связок, такого, как описанные Розенбергом [19], Ротенбергом и др. [20] и Титце [21], даже при том, что последний может адекватно представлять распределение общей спектральной энергии некоторых реальных сигналов возбуждения. Есть случаи, когда эти небольшие изменения вызывают заметные изменения в интенсивности формант в результате спадов спектра голосовой щели и совпадения формантных частот.

Даже изменение спектра возбуждения голосовой щели не является постоянным во времени. Скорость и резкость закрытия складок в значительной степени зависят от вокальных усилий. Мощность возбуждения на высших звуковых частотах в основном происходит из очень коротких временных отрезков около моментов, когда голосовые связки, наконец, вступают в контакт в заключительной фазе каждого цикла вибрации, и это искривляет форму волны объёмной скорости в этих точках, которые управляют силой такого возбуждения. Эта кривизна пропорциональна как объёмной скорости как раз перед закрытием, так и скорости движения голосовых связок при закрывании. Повышение вокального усилия заставляет обе эти переменные расти и поэтому вызывает заметное изменение мощности на высоких частотах. С другой стороны, сокращение импульсов головой щели при повышенном вокальном усилии [22] приводит в результате лишь к незначительному изменению общего объёма воздуха, несмотря на изменение пиковой скорости потока. Мощность на основной частоте, таким образом, гораздо более постоянна при изменении вокальных усилий. Спектрограммы на Рис. 6 иллюстрируют этот эффект.


kHz
How are you ?
4
0
a
4
0
b
4
0
c
0
0.5s

Для многоцелевых синтезаторов речи не требуется получения эффектов широкого спектра вокальных усилий, но в человеческой речи, как правило, есть некоторое изменение усилия даже в пределах каждой группы дыхания, также как между говорящими. Использование постоянной формы импульса голосовой щели в каскадном синтезаторе не позволяет воспроизвести изменения вокальных усилий и разных сложных эффектов импульса голосовой щели, поскольку относительные интенсивности формант автоматически устанавливаются формантными частотами; альтернатива для того, чтобы сделать спектр импульса возбуждения изменяющимся реалистичным образом, требует дополнительного управления и заметную дополнительную сложность.

# 4. Простые подходы к параллельному формантному синтезу

Как видно из приведённых выше рассуждений, в каскадном синтезаторе нельзя в общем случае полагать, что он даёт правильные относительные амплитуды формант, наблюдаемые в человеческой речи, даже для тех звуков, для которых он является наиболее подходящим теоретически. Из-за этих проблем стоит задуматься, в какой степени параллельные синтезаторы смогут их избежать. Учитывая возможные преимущества каскадного соединения для не-носовых сонорных, первый вопрос для рассмотрения, как много из недостатков параллельного соединения накладывается на эти звуки. Хорошо известно, что все-полюсную функцию передачи резонансной системы, такой, как каскадный синтезатор, можно разложить на простейшие дроби, в которых каждый член представляет отклик одного резонатора. В общем, числители членов являются линейными функциями комплексной переменной частоты, ***s***, но в частном случае, когда действительные части координат полюса являются равными, эти линейные члены сводятся к константам. Таким образом, можно видеть, что каскадное соединение резонаторов, моделирующих несколько формант, эквивалентно полосам частот, которые могут быть точно представлены параллельным соединением, при условии, что отдельные уровни формантных схем устанавливаются равными коэффициентам простейшей дроби. Значения коэффициентов усиления легко вычисляются из знания формантных частот и будут иметь противоположный знак для последовательных формант, тем самым подразумевая, что выходы параллельных формантных генераторов должны быть смешаны с использованием разных полярностей. Если формантные полосы не равны, но изменяются в пределах, нормальных для речевых сигналов, отклик параллельного варианта по-прежнему почти идентичен отклику каскадного, как показано на Рис. 7. Так что параллельное соединение имеет практический недостаток, заключающийся в том, что сигналы управления усилением должны быть предусмотрены для каждой форманты, но если эта же формантная система может быть использована для согласных звуков, общая сложность синтезатора по-прежнему может быть меньше, чем будет найдена в каскадном синтезаторе с отдельными схемами для согласных.

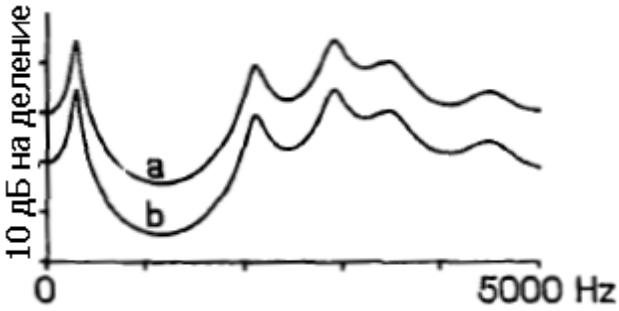



**Рис. 7. Иллюстрация, что неравные формантные полосы не мешают параллельному синтезатору очень близко приближаться к отклику каскадного соединения;**
**(a) каскадные форманты;**
**(b) параллельные форманты.**

На первый взгляд может показаться, что оптимальным способом использования параллельного формантного синтезатора является использование его с коэффициентами усиления, соответствующими моделированию каскадного вида, когда требуется каскадный отклик, и варьировать уровни по мере необходимости, как это требуется для амплитуд других формант. Такой выбор, конечно, подразумевал бы, что коррекция высшего полюса явно обеспечена в случае аналоговой реализации или реализации с высокой частотой сэмплирования. Несмотря на свою теоретическую привлекательность, такой подход не реален для аналоговых синтезаторов. Точный параллельный эквивалент пяти-формантного каскадного синтезатора имеет асимптотический наклон его амплитудно-частотной характеристики -60 дБ на октаву. Поскольку смешивание отдельных формантных сигналов происходит с разной полярностью, этот наклон достижим взаимокомпенсацией отдельных наклонов -12 дБ на октаву у верхних хвостов отдельных формант. В полном синтезаторе коррекция высшего полюса служит тому, чтобы привести обратно уровень высоких частот к тому же порядку, что и для низких частот. Однако, когда изменяется уровень одной из низкочастотных формант, взаимокомпенсация верхних хвостов уже не происходит и асимптотический наклон имеет только -12 дБ на октаву; усиление общей цепи выше пятой форманты вследствие этого значительно возрастает и могут быть существенные изменения формы спектра между формантами. Рис. 8 иллюстрирует этот эффект для трёх типичных гласных при увеличении на 6 дБ амплитуды F2.

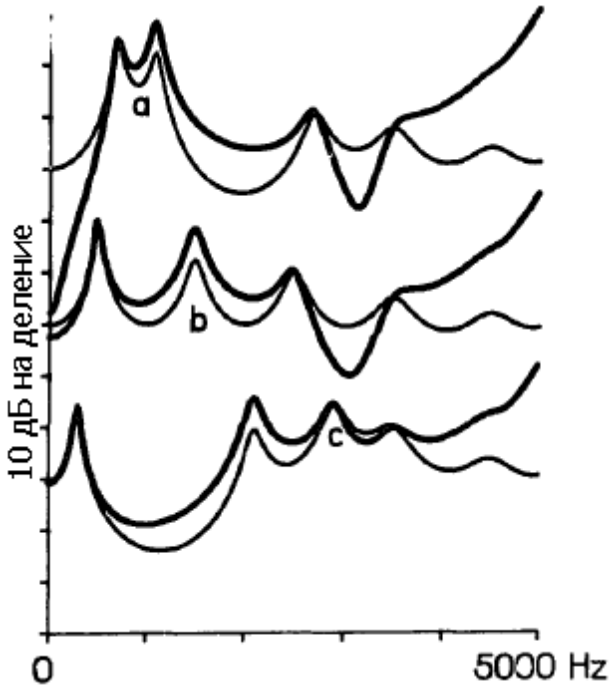



**Рис. 8. Отклик аналогового параллельного синтезатора с коррецией высшего полюса, когда амплитуда одной форманты изменяется от её теоретического значения для каскадного соединения, проиллюстрированный на трёх типичных гласных. В каждом случае тонкая линия является откликом каскадного соединения и жирная линия показывает отклик параллельного, когда уровень F2 увеличивается на 6 дБ.**
**(a) Низкий гласный заднего ряда;**
**(b) Идеальный нейтральный гласный;**
**(c) Высокий переднеязычный гласный.**

Из вышеизложенного видно, что этот метод реализации параллельного синтезатора позволяет получить точный аналог каскадного соединения, но требование точности управления амплитудой формант нереально высокое. Совсем небольшие ошибки введут нули в функцию передачи, которые могут заметно нарушить амплитуды спектральных компонентов между формантами, а большие изменения для достижения запланированных изменений амплитуд формант могут иметь катастрофические последствия для общей формы спектра.

Очевидный способ избежать этой чувствительности к амплитудным ошибкам заключается в удалении коррекции высшего полюса и перенастройке управления амплитудой формант для достижения того же спектрального отклика на формантных пиках. Отклик параллельного подключения теоретически отличается от него же в каскадной форме, но когда формантные амплитуды установлены должным образом, ошибки не очень большие. Результаты тех же вариаций амплитуд и гласных, как на Рис. 8, показаны на Рис. 9. Нежелательные эффекты на высокой частоте при изменении амплитуды гораздо меньше, чем это показано на Рис. 8, но нарушение формы спектра на других частотах очень заметно.

Сильная чувствительность на высокой частоте к изменению амплитуды, показанная на Рис. 8, возникает только в аналоговых или работающих с высокой частотой сэмплирования синтезаторах. Методу генерации формант на минимальной частоте сэмплирования присуща коррекция высшего полюса, что фактически означает, что для этого не требуется высокая

степень взаимокомпенсации откликов на краях в параллельном варианте. Отклики параллельного дискретного синтезатора с теми же изменениями амплитуд формант, как на Рис. 8, показаны на Рис. 10, и видно, что они немного лучше на высоких частотах, чем у показанных на Рис. 9, хотя эффекты на других частотах дают тот же тип нарушения спектральной формы.

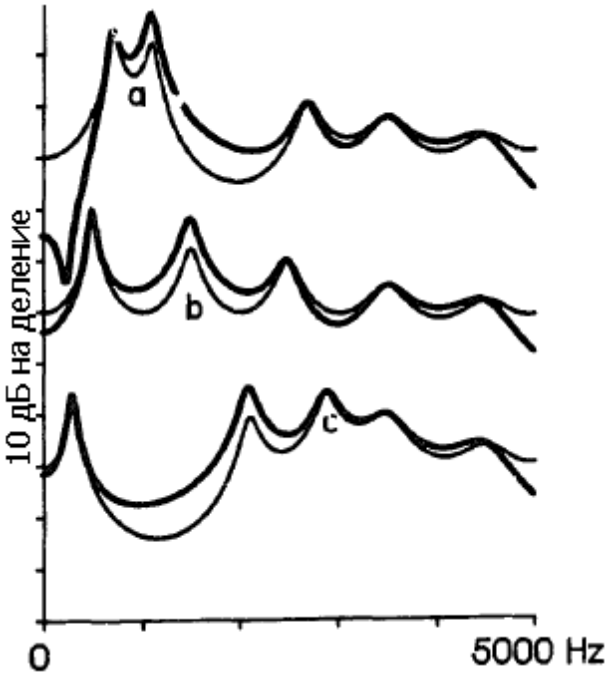



**Рис. 9. Отклики, эквивалентные тем же на Рис. 8, если не используется коррекция высшего полюса.**

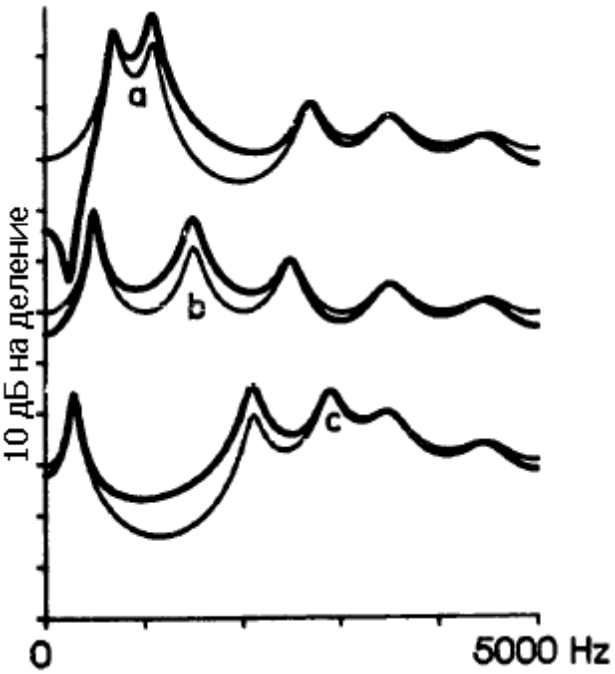



**Рис. 10. Отклики, эквивалентные тем же на Рис. 8, но дискретный синтезатор использует частоту дискретизации 10 кГц.**

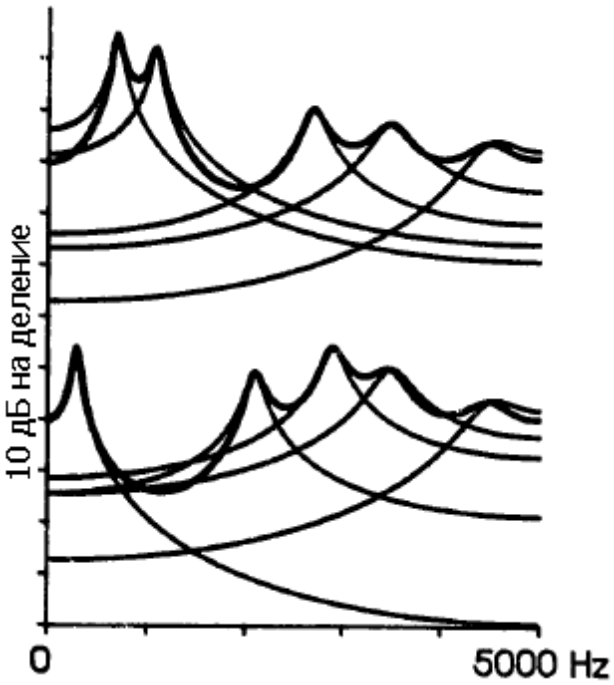



**Рис. 11. Отклики отдельных формантных генераторов параллельного дискретного синтезатора, используемого для моделирования отклика каскадного для двух гласных. Толстая линия представляет собой совокупный ответ в каждом случае.**

Рис. 11 показывает вклад каждого формантного генератора в параллельный эквивалент каскадного соединения для двух из гласных, использованных на Рис. 10. Видно, что спектральный уровень между формантными пиками в общем является результатом комбинации значительных вкладов от нескольких формант. Поскольку выходы генераторов формант должны быть подключены с разными знаками, чтобы сделать их сочетание корректным между формантными пиками, могут быть случаи, когда составной ответ на определённых частотах на самом деле меньше, чем индивидуальный вклад каждой из нескольких формант. Такая частичная взаимокомпенсация выходных сигналов от отдельных формантных генераторов является причиной сильно меньшего низкочастотного отклика, показанного на Рис. 9а и 10а, и очевидно, что большое изменение амплитуды любой форманты, как правило, будет иметь заметные последствия во многих областях спектра в дополнение к своей основному эффекту управления интенсивностью одной форманты. В области ниже частоты F1 спектральный уровень голосовой речи обычно высок, поэтому изменение низкочастотного уровня в результате изменения амплитуд высших формант, вероятно, даст субъективно нежелательные изменения мощности на низких частотах. Спектральные изменения между другими формантами, вероятно, будут менее важными, даже там, где произойдёт изменение уровня на несколько децибел, потому что такие ошибки будут скрыты для восприятия благодаря восходящему распространению маскировки [23] от интенсивных областей низкочастотных формант.

Конечно, главная причина для использования параллельного синтеза заключается в предоставлении возможности для гораздо больших изменений амплитуды формант, чем показано на Рис. 8 - 10, таких, как возникающие в результате изменения спектра источника и в ещё большей степени для различных согласных звуков. Например, для носовых согласных уровни мощности верхних формант на много децибел ниже их уровня для гласных, а для глухих

фрикативных очень слабы нижние форманты. Для успешного моделирования спектра этих звуков важно, что большое изменение амплитуды формант не должны производить неречеподобных неуместных эффектов удалённых по частоте от тех формант, управление которыми производилось. Изменение амплитуды формант - не единственные спектральные особенности, наблюдаемые во время произнесения человеком этих согласных звуков; также могут возникнуть глубокие провалы спектра как результат нулей передаточной функции, но провалы ***сами по себе*** имеют небольшое значение для восприятия. Наиболее важный эффект нулей функции передачи голосового тракта состоит в изменении величины функции передачи на частотах полюсов. Для параллельного синтезатора с явным контролем интенсивности каждой форманты главными требованиями являются:

(i) управление каждой амплитудой должно иметь свой основной эффект в области частот, близких к его собственной форманте;
(ii) должна быть разумная речеподобная интерполяция спектральной кривой между формантными пиками, кроме случаев, когда уровень является достаточно низким, и ошибки субъективно замаскированы мощностью соседних формант.

Первое из этих требований подразумевает, что отклик от какого-либо одного формантного генератора на частотах между парами удалённых формант должен быть низким по сравнению с желаемым общим откликом на этих частотах. Видно, что это условие не выполняется системами, использованными при создании Рис. 8 - 10, и поэтому необходимо отказаться от этих простых параллельных схем и пересмотреть основные требования к параллельной формантной системе.

# 5. Разработка универсальной параллельной формантной системы

В этом Разделе будет показано, что действительно хороший параллельный формантный синтезатор значительно отличается от простых систем, описанных в Разделе 4 11.

В модели речеобразования источник-фильтр с параллельными резонаторами удобно сделать сигналы управления амплитудами только теми параметрами, которые определяют мощность в спектре вокруг формантных пиков, потому что эти элементы управления могут быть связаны с непосредственно измеряемыми свойствами естественной речи. Такой выбор означает, что спектральная интенсивность сигнала возбуждения на формантных частотах должна быть независимой от вокальных усилий и от громкости речи. Эффекты изменения интенсивности и изменение спектра реального сигнала возбуждения, точка приложения возбуждения и структура голосового тракта все могут быть представлены в модели системы формантных фильтров.

Эти предположения всё ещё оставляют многие аспекты системы не определёнными. При системе синтеза, изображенной символически на Рис. 12, источник возбуждения представлен в виде генератора плоского спектра с фильтром, формирующим огибающую возбуждения; результирующий сигнал подаётся в параллельную формантную систему и в конце на фильтр. Предполагая, что процессы фильтрации можно рассматривать как стационарные и линейные (что является разумным приближением за исключением случаев, когда спектр меняется быстро), есть широкий выбор, как разделить комбинированное определение спектра между тремя составными частями. Выбор будет зависеть главным образом от простоты разработки системы формантных фильтров, а также будет находиться под влиянием таких факторов, как шум аналоговой цепи или цифровые шумы квантования, удобства описания и избегания неблагоприятных переходных эффектов, которые могут возникнуть, когда потребуются

быстрые изменения спектра. В системе формантных фильтров легче удовлетворить требованиям отклика отдельных формантных генераторов на краях, если максимальный отклик, необходимый для каждой из формант, примерно одинаков. Величина отклика от голосовой щели до губ неразветвлённого голосового тракта имеет примерно одинаковое максимальное значение для каждой из формант и поэтому представляется целесообразным, как и в системах, описанных в Разделе 4, указать, что для гласных система формантных фильтров должна быть способна вплотную приблизиться к отклику неразветвлённой акустической трубы (или её каскадного формантного эквивалента). Однако, в отличие от систем Раздела 4, он также должен быть пригодным для получения соответствующих спектральных форм для других звуков речи с сильно различающимися амплитудами формант.

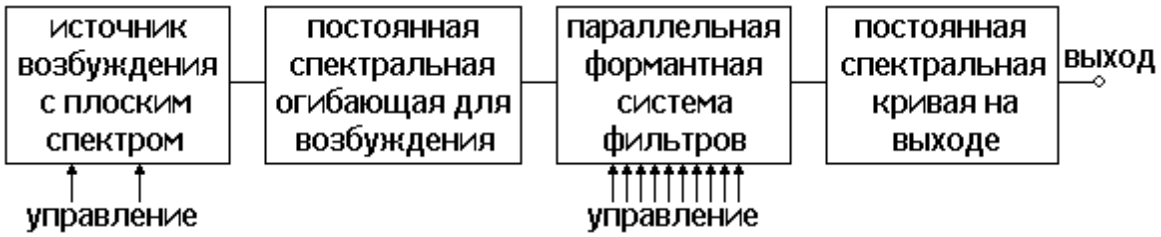



**Рис. 12. Блок-схема, иллюстрирующая получение огибающей спектра возбуждения, параллельную формантную систему и формирование выходного спектра.**

Причина, почему схемы, используемые для Рис. 8, 9 и 10, не были приемлемыми в том, что общий отклик был во многих местах зависящим от сочетаний значительных краевых откликов от нескольких формант. Можно преодолеть эти краевые эффекты, связывая каждый генератор основных формант с дополнительной цепью формирования спектра перед смешиванием сигналов отдельных формант. В идеале требования к таким фильтрам формирования спектра таковы:

(i) они должны иметь сильный по амплитуде отклик в диапазоне частот, разрешённом для соответствующей форманты (однако, не обязательно, что отклик должен быть постоянным в этом частотном диапазоне, потому что при изменении формантной частоты можно компенсировать колебания изменением сигналов управления амплитудами);

(ii) они не должны вызывать фазовых искажений, с тем, чтобы компоненты значительной амплитудой от соседних формантных генераторов объединялись в правильном фазовом соотношении.

(iii) они должны обеспечить существенное ослабление любых компонентов, которые обладают значительной амплитудой вне диапазона частот формантного генератора, который производит их;

(iv) уровни усиления/частотные характеристики должны меняться постепенно на частотах, для которых выходы имеют значительную амплитуду, чтобы избежать нарушений в объединённом отклике полного синтезатора.

Использование такого набора фильтров, хотя он предотвращает неприятные отдалённые краевые эффекты, также приносит свои собственные проблемы. В естественной речи, когда два форманты близки друг к другу, их амплитуды значительно возрастают; скорость падения амплитудного отклика на краях такой пары формант является такой же, как у двухрезонаторного полосового фильтра, и гораздо больше, чем у одного резонатора. В параллельной реализации этот эффект может быть достигнут только путём взаимокомпенсации формантных откликов и поэтому отклики любых дополнительных фильтров, включённых в формантные генераторы, должны быть достаточно похожи в тех местах, где возможно смешивание формант, чтобы эта взаимокомпенсация всё ещё происходила. Поэтому необходимо пятое условие:

(v) фильтры для соседних формант должны иметь очень похожие отклики в частотных

областях около возможного смешивания формант.

Это дополнительное условие не вызывает каких-либо проблем при смешивании высших формант, но в случае смешивания F1 - F2 встречается с фундаментальной трудностью. В каскадном синтезаторе на низких частотах передаточная функция имеет единичную величину и в её точном параллельном эквиваленте единичная величина определяется комбинацией нижних хвостов отликов всех отдельных резонаторов, как показано на Рис. 11. Отклики F2 и F4 имеют противоположную полярность к таковым на F1, F3 и F5. В случае звука подобного [a], где F1 и F2 очень интенсивны и близки по частоте, правильная величина на низкой частоте может быть получена только при наличии высокой степени взаимокомпенсации между нижними хвостами F1 и F2. Вышеприведённое условие (iii) требует, что для глухих фрикативных не должно быть значительного ослабления нижнего хвоста F2, и таким образом взаимокомпенсация откликов на частотах, ниже F1 - F2, требуемая в некоторых гласных, происходить не будет.

Возможным решением этой проблемы является предоставление специальной схемы для поддержания необходимого уровня низкочастотного сигнала, зависящего от амплитуд и частот F1 - F5, и это решение объясняется в Разделе 6 20. Предполагая, что таким образом можно избежать трудностей с низкочастотным откликом, требованиям к фильтрам спектральной огибающей в дискретных синтезаторах не очень трудно удовлетворить КИХ фильтрами довольно небольшого порядка. Однако, для практических синтезаторов речи требование (ii) может быть несколько ослаблено. Так как человеческая слуховая система не чувствительна к фазе отклика, если искажение групповой задержки мало, достаточно определить (ii), как:

(ii) фазы откликов должны быть такими, чтобы компоненты значительной амплитудой от различных формантных генераторов объединялись в правильном соотношении фаз.

Этот ослабленный набор условий может, фактически, быть выполнен достаточно хорошо набором очень простых фильтров. Для всех формант, кроме F1, основной эффект требования (iii) состоит в том, что величина отклика должна быть очень небольшой ниже частоты F1, такой, чтобы мог быть адекватно синтезирован спектр глухого согласного. Это условие может быть выполнено с помощью фильтра с одним нулём в своей функции передачи в начале ***s***-плоскости, то есть с помощью простой дифференцирующей цепочки. Такой фильтр также достаточно хорошо удовлетворяет всем другим условиям для этих формант, если может быть выбрана подходящая характеристика фильтра для F1, чтобы заставить его выход надлежащим образом объединиться с сигналом F2. Когда формантный генератор используется с дифференцирующей цепочкой, очень легко, как для дискретной, так и аналоговой конструкции, добиться того, чтобы максимум амплитуды огибающей импульсной характеристики по большей части не зависел от частоты форманты и ширины диапазона (см. Рис. 13). Для многих приложений эта величина является наиболее удобным определением амплитуды формант для F2 и выше, так как она легко оценивается из анализа естественной речи и может быть непосредственно использована в синтезаторе.

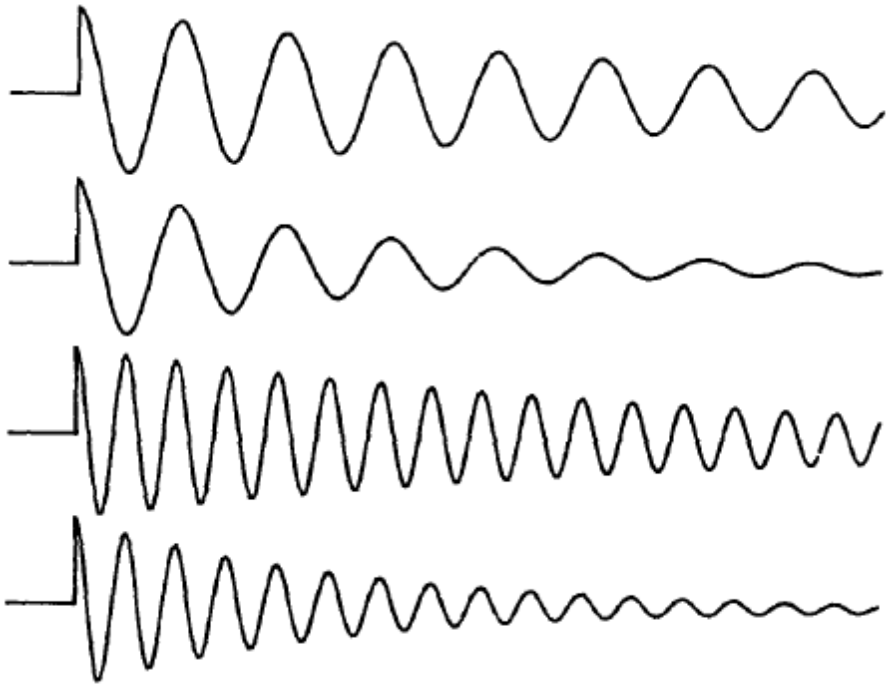


**Рис. 13. Продифференцированные импульсные характеристики формантных генераторов для разных формантных частот и полос, все определены как имеющие одинаковую амплитуду форманты.**

Поскольку коррекции высшего полюса нет, отклик только что описанного параллельного соединения будет быстро падать выше самой верхней явной форманты, чем должен в идеальной резонансной модели, если не используется дискретная форма с минимальной частотой дискретизации. Этот эффект может быть преодолён путём добавления одного или нескольких высокочастотных резонансов в цепь формирования спектра самых высокочастотных параллельных формант и эта договоренность была использована в откликах, показанных на Рис. 16 и 17. Однако, как описано в Разделе 2[3], поскольку предположение о плоской волны выше 3 кГц является недопустимым, в человеческом голосовом тракте существуют дополнительные резонансы и антирезонансы. Поэтому, вероятно, более реалистично просто использовать один или два достаточно широкополосных фильтра, содержащий несколько резонаторов для представления формы спектра выше 3 кГц, и отказаться от любых попыток моделировать каскадный отклик в этой области частот [13].

Для фильтра, формирующего спектральную кривую для F1, величина низкочастотного отклика должна быть приблизительно не зависящей от частоты, так чтобы генератор F1 мог сделать правильный тип вклада в общий отклик на низкой частоте. Выше частоты F1 необходимо, чтобы фазовый сдвиг фильтра приближался к 90° для достижения правильного фазового соотношения с выходом дифференцирующей цепочки, используемой с F2. Единственный вещественный нуль, на позиции около -600 или -700 Гц, даёт разумное приближение к этим условиям, хотя сдвига фаз между частотой F1 и низкими частотами F2 не совсем достаточно для достижения хорошего сложения с нижним хвостом F2 в этом регионе. Подъём 6 дБ на октаву, который этот нуль вызывает на высших звуковых частотах, не нарушает условия (iii), так как резонатор F1 сам по себе даёт гораздо большее затухание верхнего хвоста, чем необходимо, чтобы избежать проблем при синтезе высокочастотных формант. Начало подъёма 6 дБ на октаву в области ниже F2 на самом деле помогает, так как последующий дополнительный выход от генератора F1 частично компенсирует ослабление нижнего хвоста F2, вызванное нулём фильтра F2. В этом случае физический смысл сигнала,

управляющего амплитудой F1, не такой же, как для высших формант. Если то же самое определение амплитуды форманты необходимо для F1, коррекция управления амплитудой должна быть рассчитана исходя из знания формантной частоты. Однако, значение амплитуды, проиллюстрированное Рис. 13, не очень полезно для низких формантных частот из-за трудностей оценки огибающей волны естественной форманты, когда её амплитуда заметно затухает в течение одной половины цикла резонансной частоты, и в таком синтезаторе удобнее задать амплитуду сигнала F1 указав усиление генератора F1 на нулевой частоте.

Существует небольшое преимущество, если к фильтру F1 добавлен дополнительный всепропускающий фазовращатель для получения небольшого дополнительного сдвига фазы в области между F1 и F2. Выбор -640 Гц для реального нуля F1, и пары реального нуля и полюса на +/- 270 Гц даёт фазовую характеристику, показанную на Рис. 14, и лишь незначительно отличается от 90° в важном диапазоне частот.

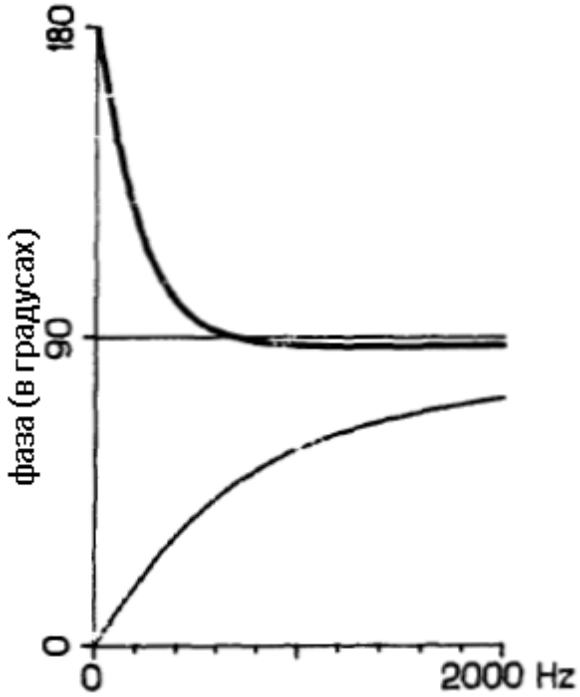



**Рис. 14. Фазовый сдвиг предложенного фильтра спектральной огибающей F1. Тонкая линия является откликом только с одним вещественным нулём, а жирная линия показывает эффект от добавления дополнительной коррекции фазы.**

## 6. Низкочастотный отклик параллельных синтезаторов

В голосовом тракте человека, если в нём не полностью или почти полностью не препятствуют для взрывных или фрикативных согласных, объём воздуха, входящего через голосовую щель, равен совокупному объёму воздуха, выходящего через рот и нос. Отсюда следует, что ниже частоты самого низкого резонанса передаточная функция голосового тракта для не шумящих звуков близка к единичной величине, независимо от фактически произносимого звука. Импеданс голосовой щели во время звучания высок и за исключением шумящих поток воздуха в любой момент незначительно изменяется фактической конфигурацией голосового тракта. Амплитуды самых низкочастотных составляющих речевого сигнала (основной компонент, а для говорящих взрослых мужчин также и вторая гармоника)

имеют поэтому почти постоянную амплитуду во время разговора, или изменяются только очень медленно, с изменениями в голосовом усилии или тоне. Эти компоненты обычно содержат наибольшую мощность в речевом сигнале, а поскольку они так мало зависят от голосового тракта, они почти не способствуют разборчивости речи и, фактически, отфильтровываются в телефонных сетях общего пользования. Однако, важно, чтобы они присутствовали в высококачественной речи, и мой опыт работы с синтезом речи показал, что если самые низкочастотные компоненты присутствуют, люди-слушатели более чувствительны к неестественным изменениям их уровня, чем к аналогичного размера (в дБ) уровням ошибок на формантных частотах.

С помощью каскадного синтезатора единичная величина низкочастотной характеристики достигается автоматически и управление уровнем требуется только для представления низкочастотной интенсивности голосового источника. Если импульсы голосового источника имеют соответствующую форму и тракт не имеет носового ответвления, влияющего на амплитуды самых низких двух или трёх формант, то звук действительно "сразу выходят правильным" для гласных.

Как уже было описано в Разделе 5 16, в параллельной конфигурации необходимы дополнительные схемы формирования спектра с помощью формантных генераторов, если для некоторых согласных звуков в области F1 должен быть достигнут достаточно низкий уровень. Характеристики передачи высоких частот этих схем для F2 и более высоких формант делают усиление синтезатора на низкой частоте зависящим только от установленной амплитуды F1, и поэтому значительно различаются, так как управление амплитудой регулируется для получения разных интенсивностей F1, необходимых для различных гласных и звонких согласных. Холмс [17] описал параллельный синтезатор, в котором был предусмотрен дополнительный резонатор ниже частоты F1 для улучшения моделирования низкочастотных областей голосовой речи, и, в частности, для дополнительного спектрального пика, часто наблюдаемого в диапазоне 250-300 Гц для носовых гласных. Из-за этого последнего использования этот дополнительный резонанс для удобства называют носовой формантой, или FN. Если реальная координата ***s***-плоскости этого резонанса достаточна велика (скажем, около -150 Гц), его низкая резонансная частота приводит к характеристике усиления, которая имеет лишь небольшой резонансный пик, и является более похожей на характеристику фильтра низких частот. Этот выход добавляется параллельно с нижним хвостом отклика F1 и изменяя его интенсивность можно управлять амплитудами самых низких спектральных компонентов голосовой речи, независимо от настройки амплитуды F1. В синтезе, описанном Холмсом [17], для оптимизации синтезированного произношения были использованы все доступные сигналы управления и в последствии управление амплитудой FN было использовано для такой коррекции низких частот, а также для его основной цели - улучшения носовых гласных. Успех синтеза, о котором сообщалось в той работе, является доказательством того, что использование независимого контроля низкочастотным звуковым спектром может преодолеть проблемы поддержания правильного уровня в этом регионе.

Основной недостаток синтезатора Холмса 1973 года в том, что интенсивность низких частот, которая в человеческой речи почти полностью зависит от интенсивности голосового источника и меняется только очень незначительно с артикуляционными изменениями, была получена в результате комбинированных характеристик резонаторов F1 и FN, с помощью независимого управления их амплитудой. Очевидно, что использованию синтезатора способствовало бы, если бы для определения уровня низкой частоты была использована непосредственно одна из регулировок амплитудой, как это делает управление амплитудой голоса в каскадном синтезаторе.

Холмс [24] описал изменённую конфигурацию синтезатора, которая обеспечивает такую возможность. Новый сигнал управления амплитудой, известный как ALF, используется, как показано на Рис. 15, таким образом, что он контролирует сумму сигналов возбуждения F1 и FN. Если собственные низкочастотные уровни обоих резонаторов равны, их совокупный выход в этой области спектра будет определяться только ALF. Степень, в которой ALF правильно определяет комбинированные сигналы F1 и FN на высоких частотах в пределах полосы пропускания FN, зависит в основном от формы амплитудной характеристики FN и сходства фазовых характеристик FN и F1. Со специальной схемой формирования F1, описанной в Разделе 5 16, существует фазовый сдвиг сигнала F1 на 180° на нулевой частоте, что требует изменения полярности схемы FN, чтобы произвести соответствующее подключение сигнала. Если это изменение полярности обеспечено и координатой полюса FN является -150 +/- j200 Гц, разность фаз между F1 и FN достаточно мала для амплитуды общего сигнала, который будет почти полностью определяться ALF до частоты около 300 Гц, за исключением случаев, когда F1 имеет более низкую частоту.

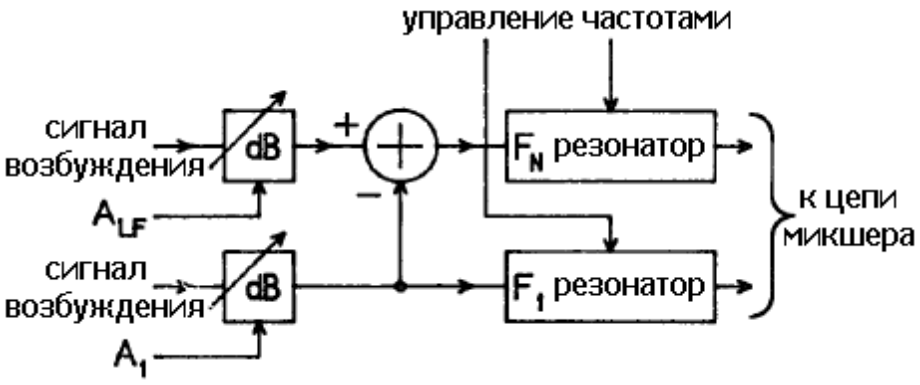



**Рис. 15. Блок-схема для иллюстрации использования сигнала управления ALF.**

Амплитудная характеристика полного параллельного синтезатора этого типа для различных гласных показана на Рис. 16 и 17. Этот синтезатор предназначен для покрытия диапазона частот до 4 кГц, что достаточно для большинства практических применений. Приближение к каскадному соединению для всех этих гласных при правильно установленной амплитуде управления так близко, что не может быть показано графически (ошибки заметно меньше, чем 1 дБ, на всех частотах до 4 кГц). Рис. 16 показывает эффект увеличения амплитуды F2 на 6 дБ, как на Рис. 8 - 10, и видно, что спектральные эффекты гораздо более приемлемы, особенно на низких частотах. Рис. 17 показывает, что вариации интенсивности формант, которые могут возникнуть в результате изменения вокального усилия, не вызывают резких изменений формы спектра между резонансными пиками или ниже частоты F1.

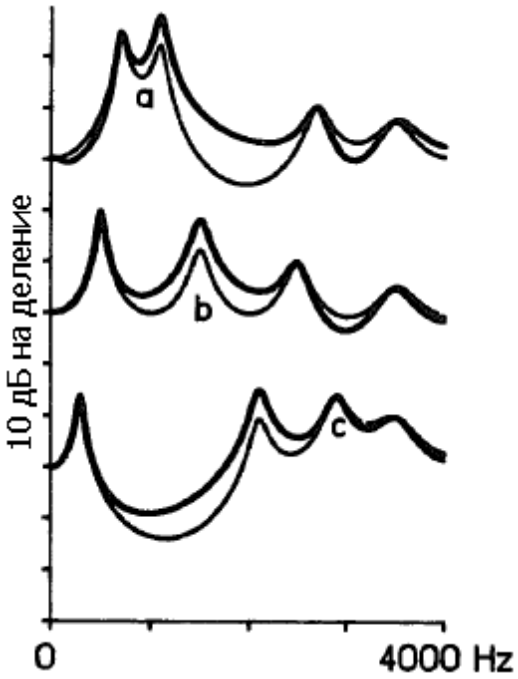



Рис. 16. Отклик предложенного параллельного синтезатора для тех же условиях, которые были использованы для Рис. 8-10.

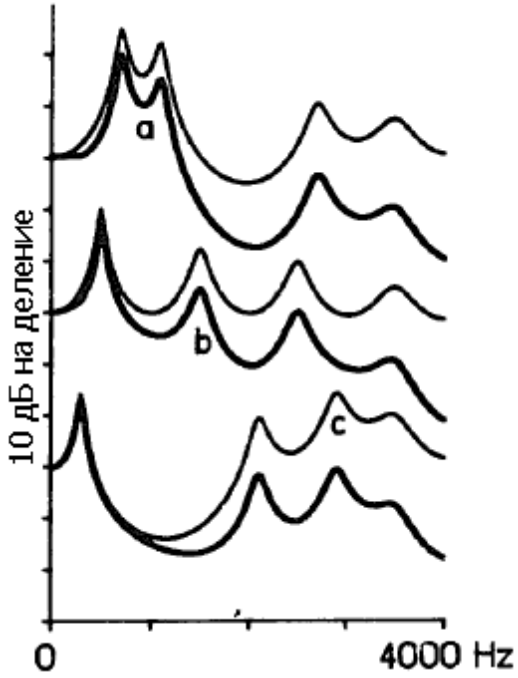



Рис. 17. Влияние на отклик синтезатора изменений амплитуд формант, чтобы смоделировать изменения вокальных усилий. Тонкие линии представляют нормальное вокальное усилие, а толстые линии показывают уменьшенное вокальное усилие.

Когда резонатор FN действительно необходим для его номинальной цели моделирования спектрального пика около 300 Гц в носовых согласных и носовых гласных, реальная координата -150 Гц слишком велика. Компромиссное значение, скажем, -90 Гц было найдено, чтобы быть всё ещё совместимым с довольно хорошим моделированием каскадного отклика без носовых, но могло бы также произвести разумные копию низкочастотного спектрального пика в носовых звуках. Для ещё лучшего моделирования низкочастотной области для полного спектра голосовых звуков может быть необходимо разработать какое-то более тщательно продуманное расположение фильтров, но конфигурация, описанная здесь, как показано, способна создавать речь, которая субъективно очень близка к естественной.

# 7. Сигналы возбуждения для параллельных синтезаторов

## 7.1. Голосовое возбуждение

Как уже было предложено в Разделе 5 16, желательно сделать, чтобы уровень возбуждения на частотах формант параллельных синтезаторов не зависел от вокальных усилий и уровня громкости, но не очевидно, как определение интенсивности голосового возбуждения должно быть связано с основной частотой. Мощность голосового возбуждения могла бы быть контролируемой, чтобы сохранить одинаковую среднюю плотность мощности спектра при изменении F0. Такой выбор подразумевал бы, что амплитуда импульса возбуждения должна быть обратно пропорциональна квадратному корню из F0. Возможные альтернативы в том, что интенсивности гармоник должны быть независимы от F0, или что мощность каждого импульса возбуждения должна быть независимой от F0. Последнее определение является наиболее удобным для вписывания в определение амплитуды высших формант, приведённое в Разделе 5 16. Такой выбор означает, что средняя мощность голосового возбуждения может быть равной глухому возбуждению только на одной частоте, которая может быть выбрана как некоторое среднее значение величины F0.

Соображения, представленные в Разделе 5 16, привели к решению, что для гласных формантная система должна быть способна предоставлять отклик каскадного синтезатора. В Разделах 5 16 и 6 20 было установлено, что с помощью универсальной параллельной системы можно вплотную приблизиться к каскадному отклику, при условии, что используется специальный резонатор низкой частоты, названный FN. В звуках голоса изменения для компенсации вариаций вокального усилия и любых других вариаций спектра возбуждения могут быть сделаны с помощью управления амплитудой формант, чтобы изменять относительные интенсивности всех формант по отношению к уровню низкочастотного сигнала. Такое соглашение, конечно, не обеспечивает контроль над всеми спектральными свойствами голосовой речи на основе от гармоники-к-гармонике, а также не определяет точную временную структуру краткосрочного спектра кроме той, что определяется периодичностью возбуждения. Поэтому важно, чтобы любые такие особенности, которые могут быть значимыми для восприятия, были бы предусмотрены некоторым другим способом.

Хотя они имеют очень небольшое значение для разборчивости речи, несколько самых нижних гармоник голосовой речи (то есть обычно ниже частоты F1) содержат большую часть мощности сигнала и имеют очень большое влияние на воспринимаемое качество. Приемлемости синтетической речи, несомненно, вредит, если эта область спектра большой мощности включена, но не хорошо моделируется (для комментариев по этому аспекту в

отношении LPC вокодеров смотрите, например, Ванг, Сяо и Маркел [25]). Стилизованное представление формы волны голосовых связок, предложенное Розенбергом [19], показано на Рис. 18. Показывая спектр дважды дифференцированной формы сигнала, Рис. 19 показывает, как спектр импульса такой формы отличается от наклона -12 дБ на октаву, что часто цитируют в качестве представляющего типичный импульс голосовых связок. Видно, что для данного сигнала вторая гармоника на 5.6 дБ сильнее, а четвёртая гармоника на 2.9 дБ сильнее, чем подходит для наклона -12 дБ на октаву. Детальные характеристики низких частот, так как они весьма значимы для восприятия, поэтому должны быть сохранены, если требуется достигнуть необходимого качества низких частот. Холмс [17] обсудил моделирование импульса голосовых связок по отношению к формантным синтезаторам и рекомендовал сигнал возбуждения на основе второй производной по времени типичного импульса объёмного потока в голосовой щели. Для импульсов голосовой щели с наклоном -12 дБ на октаву на высоких частотах дважды дифференцированный сигнал будет иметь примерно плоский спектр, но и для других импульсов была описана спектрально-сглаживающая процедура, которая примерно сохраняла относительные амплитуды близких гармоник в спектре и сохраняла фазовую структуру кратковременного спектра.

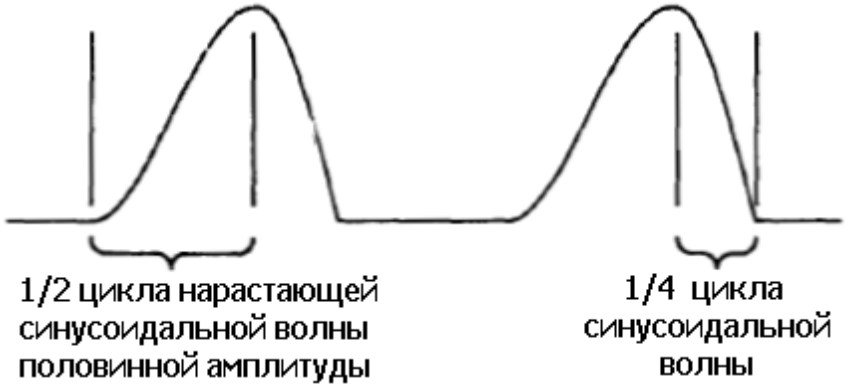



Рис. 18. Стилизованный импульс голосовой щели, составленный из синусоидальных сегментов.


Рис. 19. Уровни нескольких самых низких гармоник во второй производной сигнала по времени, показанного на Рис. 18 относительно среднего уровня высших гармоник.

Если только что описанный сигнала возбуждения используется в качестве источника на Рис.

12, выход должен быть способен хорошо моделировать звук речи, из которого была получена форма импульса возбуждения, при условии, что фильтры возбуждения и выходные фильтры выбраны подходящим образом. Совместное действие этих фильтров должно было бы компенсировать двойное дифференцирование возбуждения и смоделировать эффект излучения, и должно таким образом дать -6 дБ на октаву во всём диапазоне частот речи, вплоть до, скажем, 50 Гц.

В целом, практически невозможно обеспечить сигналы возбуждения, соответствующие широкому диапазону форм импульса голосовой щели, и степень, до которой такое моделирования является требуемым, зависит от приложения. Холмсом [17] были выдвинуты некоторые доказательства, чтобы предположить, что некоторые формы импульса голосовой щели могут быть типичными для определённых говорящих, но есть многочисленные доказательства из сигналов, получаемых обратной фильтрацией [22,26,27], что по крайней мере низкочастотные особенности формы импульса голосовой щели во многом аналогичны для большинства людей, и что кроме основной частоты наиболее важным параметром является длительность импульса (или соотношение открытая/закрытая). Таким образом, представляется вероятным, что импульс голосовой щели, получаемый Холмсом через процесс спектрального выравнивания из речи типичного говорящего, может оказаться подходящим для большинства целей. Если при соответствующей длительности такой импульс будет создавать подходящее соотношение между интенсивностями нескольких самых низких гармоник, но это также даст такой род тонких спектральных деталей и временной структуры, который необходим, чтобы вызвать часто наблюдаемое вторичное возбуждение формант иным, чем смыканием связок [18]. (фраза кажется незаконченной, но так в оригинале) Чтобы представить широкое изменение длительности импульса щели, что происходит в человеческой речи, особенно с помощью вокальных усилий, полезно иметь возможность изменять длительность импульсов возбуждения независимо от их частоты повторения. Однако, нет необходимости изменять общую форму спектра с помощью вокального усилия, так как основные эффекты такого изменения формы предоставляются управлением амплитудой формант.

Для приложений, не требующих большого естественного качества речи, можно сделать экономичные приближения к спектру импульса голосовой щели, в частности, на низких частотах с помощью отклика минимально-фазового фильтра второго порядка нижних частот. Импульсная характеристика такого фильтра с частотами полюсов на -100 + j100 Гц показана на Рис. 20 и не очень отличается от инвертированной по времени формы упрощённого человеческого импульса голосовой щели (то есть импульсы голосовой щели достаточно хорошо определяются максимально-фазовой функцией второго порядка). Искажение групповой задержки, вызванное инвертированием во времени, является незначительным для восприятия; поэтому ясно, что импульсная характеристика фильтра с низкой частотой сопряжения пары полюсов и двух нулей в начале ***s***-плоскости могла бы быть использована вместо дважды дифференцированного стилизованного импульса голосовой щели. Эффект изменения длительности импульса щели мог бы быть обеспечен перемещением позиций полюсов в фильтре.

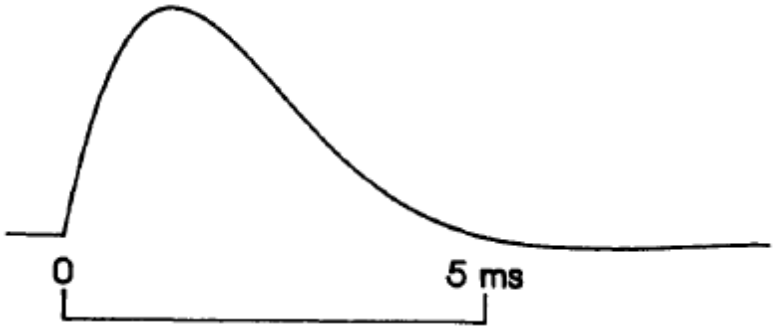



Рис. 20. Импульсная характеристика простого фильтра низких частот второго порядка, подходящего для моделирования спектра голосового источника.

## 7.2. Глухое возбуждение

В случае глухих звуков возбуждение не имеет преобладания мощности низких частот, что происходит во время звонких, и точка возбуждения обычно не в голосовой щели. В результате для таких звуков интенсивность низкочастотных формант, вообще говоря, гораздо меньше, чем интенсивность высокочастотных формант. Для многих фрикативных и взрывных близость сужения голосового тракта такова, что полость за сужением почти полностью акустически изолирована от передней части голосового тракта. F1 тогда связана с этой полостью, выступая в качестве резонатора Гельмгольца, но изоляция и тот факт, что точка возбуждения находится в передней части резонатора приводит к тому, что этот резонанс имеет очень незначительное влияние на общую передаточную функцию. Альтернативный способ рассмотрения того же эффекта состоит в том, чтобы рассматривать полюс F1 функции передачи голосового тракта как почти полностью отменённый нулём, поскольку возбуждение может рассматриваться как серии с задней полости, которая имеет очень высокое сопротивление в резонансе [15]. Таким образом, существуют противоположные требования к форме низкочастотного спектра для звонких и глухих звуков. В первом случае F1, как правило, интенсивна и форма спектра возбуждения, даже после дифференцирующего эффекта излучения, повышает интенсивность на нижнем хвосте F1. Для глухих звуков F1 очень слаба или даже не обнаруживается, и спектральная интенсивность ниже частоты F2 падает очень быстро. Если та же формантная система будет использоваться для обоих типов звуков и у амплитуд формант должен быть одинаковый физический смысл в обоих случаях возбуждения, спектральная огибающая должна быть одинаковой во всём диапазоне формантных частот. Однако, наклон -6 дБ на октаву, предлагаемый для комбинированного отклика фильтров формирования спектра ниже частоты F1, очевидно, не подходит для глухих звуков и поэтому желательно, чтобы для таких звуков на типичной частоте F1 низкочастотный подъём был бы уменьшен. Впоследствии будет небольшое изменение в физическом значении управляющего сигнала амплитудой F1 на низких частотах F1, но так как F1, как правило, крайне слаба в глухих звуках, это изменение не имеет практического значения.

## 7.3. Смешанное голосовое и глухое возбуждение

В отличие от систем с отдельными резонаторами для глухих звуков, в экономической полностью параллельной системе необходимо смешать оба звонких и глухих возбуждения в одних и тех же резонаторах для звонких фрикативных, аффрикатов и взрывных. Однако, когда требуется смешанное возбуждение, отношение мощности звонких и глухих зависит от частоты форманты. Для F3 и F4 характерно быть почти полностью глухими, когда F1 полностью

звонкая. Так как степень звонкости изменяется от полностью глухих к полностью звонким, необходимо заменять мощность глухих мощностью звонких постепенно, начиная с нижней форманты. Схема, которая обладает этими свойствами, была описана Холмсом [17] и работает следующим образом. Каждый формантный генератор имеет своё собственное управление смешиванием возбуждения, которое поддерживает постоянную мощность возбуждения, так как смесь звонких и глухих компонентов изменяется. Сигналы управления микшированием получаются от общего управления "степени звонкости", но каждый из них связан с индивидуально выбранным смещением, так что для заданной степени звонкости они могут предоставлять разные комбинации из различных формант. Полный диапазон управления отдельных смесителей (то есть от полностью глухих к полностью звонким) требует управления входным диапазоном равного только одной трети полного диапазона основного управления звонкостью и сигналы управления этих микшеров, выходящих за границы их ограниченного входного диапазона, заставляют микшер обеспечивать только звонкое или глухое возбуждение соответственно. Сделав диапазон смещений перекрывающим две трети диапазона управления звонкостью, получаем требуемую характеристику, как показано на Рис. 21.


**Рис. 21. Характеристики системы смешивания возбуждения в зависимости от изменения звонкости.**

Как было указано Рабинером [4] и Клаттом [3], при смешанном возбуждении в человеческой речи фрикативный шум имеет амплитудную модуляцию на частоте звучания, поскольку через сужение голосового тракта движется пульсирующий поток воздуха; у синтезаторов, которые они описали, было предусмотрено моделирование этого эффекта. Хотя такая модуляция, несомненно, часто наблюдается в речевых сигналах и широкополосных спектрограммах, ещё не является адекватно установленным, что это субъективно значимо. Примеры синтеза, описанные Холмсом [17], не использовали модуляцию щелью источника шума, но всё же приводили к результатам, которые в самых жёстких условиях прослушивания были почти неотличимы от естественных примеров речи, которые были смоделированы. Никакие судьи, которые были в состоянии обнаружить различия между естественным и синтетическим, не идентифицировали звуки со смешанным возбуждением как причину их суждений. Однако, если требуется, модуляция может быть представлена в простой форме, как описано Клаттом [3].

# 8. Практические соображения о динамическом диапазоне

Снова ссылаясь на Рис. 12, в настоящее время предполагается, что совокупный эффект возбуждения формирующего фильтра и выходного фильтра должен дать -6 дБ на октаву во всём диапазоне частот речи для звонких звуков, но что этот наклон должен применяться только выше частоты F1 для глухих звуков. Таким образом, необходимо, чтобы два источника возбуждения имели разные фильтры формирования возбуждения. Сигнал квантования или соображения о шуме в цепи позволяют предположить, что все формантные генераторы должны

иметь одинаковый диапазон уровней сигнала, и таким образом на выходном фильтре должен быть наклон -6 дБ на октаву. Однако, в генерации человеческой речи последняя операция, излучение через рот и нос, на самом деле вызывают наклон +6 дБ на октаву. Допустима перегруппировка порядка линейных процессов фильтрации, при условии, что спецификация фильтра не изменяется во времени, но это предположение, очевидно, не вполне оправдано для любой формантной системы фильтров. Эксперименты с системами этого типа показали, что применение -6 дБ на октаву к выходу изменяемой формантной системы может ошибочно подчеркнуть переходные процессы, усиливая низкую частоту в формантных фильтрах, производя таким образом случайные слышимые "удары" на выходе. Однако, выбором постоянной времени интегратора формирования спектра для получения точки разрыва в пределах 600 - 700 Гц и помещением оставшегося подъёма низких частот в фильтр формирования голосового возбуждения эти эффекты сделаны незаметными. Такие результаты выбора означают, что выходной фильтр обеспечивает всё необходимое для формирования спектра глухих звуков, так что фильтр глухого возбуждения может быть опущен. Этот выбор фильтрующих механизмов проиллюстрирован на Рис. 22.

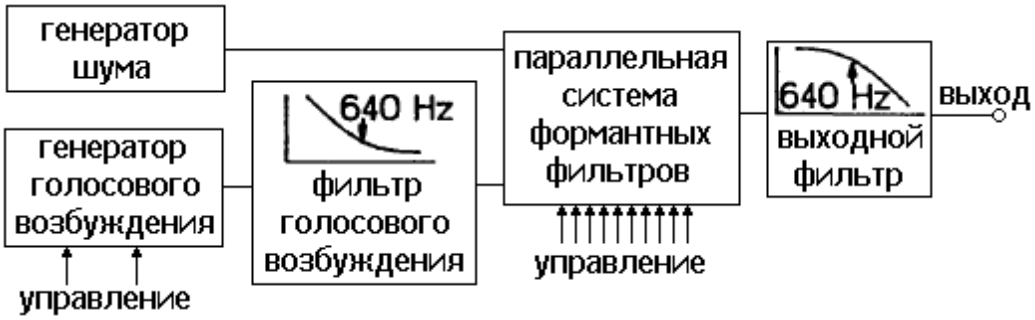



Рис. 22. Блок-схема отдельных механизмов фильтрации для голосового и глухого возбуждения.

## 9. Производительность и выводы

Формантный синтезатор, включающий особенности, объяснённые в этой работе, постепенно эволюционировал в Объединенном Исследовательское Подразделении Речи (Joint Speech Research Unit) в течение последних 20 лет. В своей последней форме он существует как программа на Фортране и оборудовании специального назначения реального времени. Он лишь немногим отличается от более ранней формы, описанной Холмсом [17], и упрощённая блок-схема параллельной системы формантных фильтров, которая включает в себя особенности, имеющие отношение к этой работе, показана на Рис. 23. У версии на Фортране есть условия для простого изменения многих аспектов проектирования и обе версии имеют много специальных особенностей, которые могли бы применяться одинаково хорошо и в каскадном синтезе и поэтому здесь не были упомянуты. Подробная информация о разработке не подходит для этой работы, но она была описана в отдельном докладе [28].

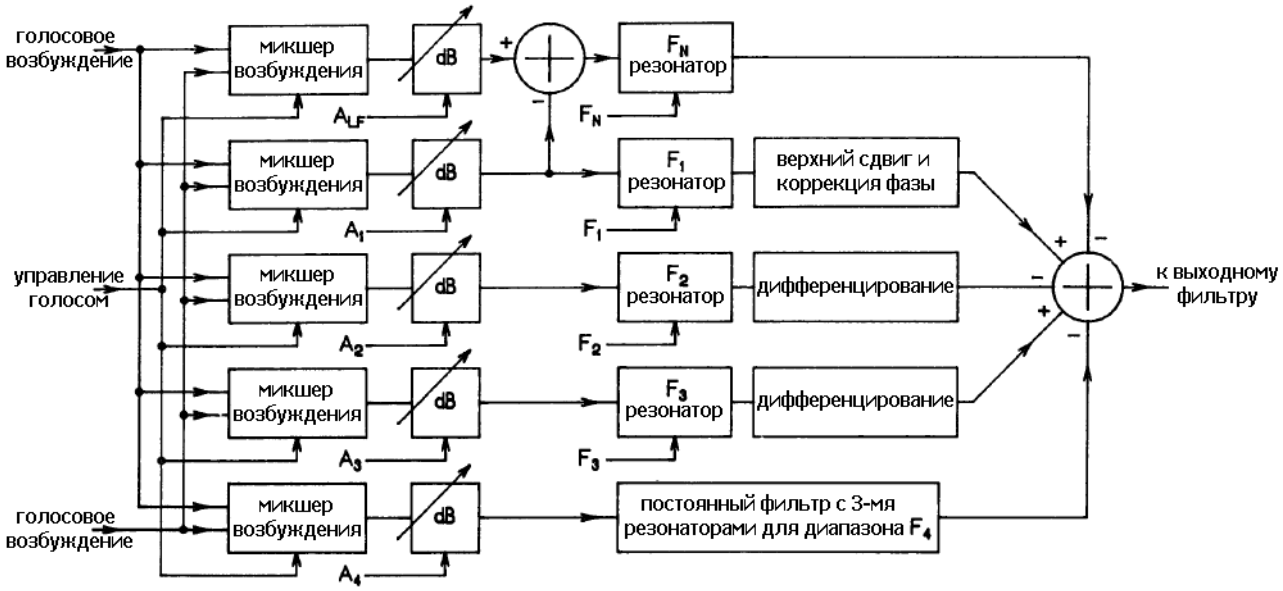



**Рис. 23. Упрощенная блок-схема параллельной формантной системы фильтров синтезатора JSRU.**

Синтезатор JSRU охватывает диапазон частот до 4 кГц и широко применяется для формантного вокодера [29], синтеза по правилу [30] и высококачественного копирования естественной речи с помощью оптимизированных вручную сигналов управления [31]. Хотя по синтезу женских и детских голосов было проделана некоторая работа, этот синтезатор почти всегда использовался для речи взрослых мужчин. Поэтому не представляется возможным сделать авторитетные заявления о то, что он подходит для других типов голоса, хотя аргументы в этой работе должны применяться одинаково хорошо для других случаев, если надлежащим образом изменить диапазоны частот.

Используя ручную оптимизацию, появилась возможность делать копии речи взрослого мужчины субъективно настолько близких к естественной в диапазоне 4 кГц, что большинству судей необходимо повторять прослушивание парных сравнений, чтобы решить, что есть что, и некоторые опытные судьи на самом деле делали неправильно выбор. Показанная на Рис. 24 пара спектрограмм иллюстрирует, насколько близко типичное синтетическое предложение может приближаться к естественной речи. На Рис. 25 приведены спектральные сечения для не-носовых гласных, плавного звука, носового согласного и период придыхания окончания глухого.

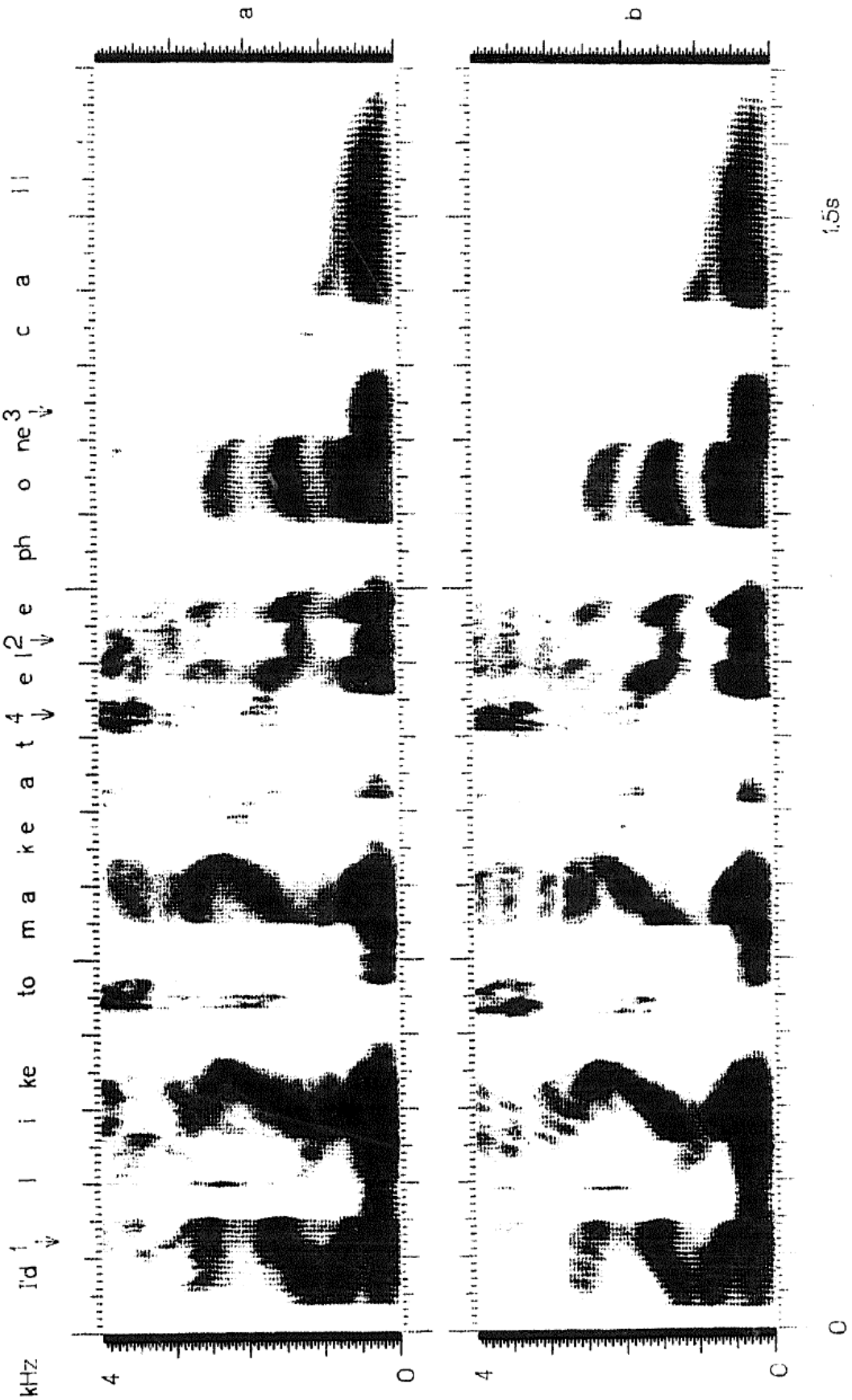



Fig. 24. (a) Спектрограмма естественной речи;
(b) Спектрограмма на выходе синтезатора JSRU.

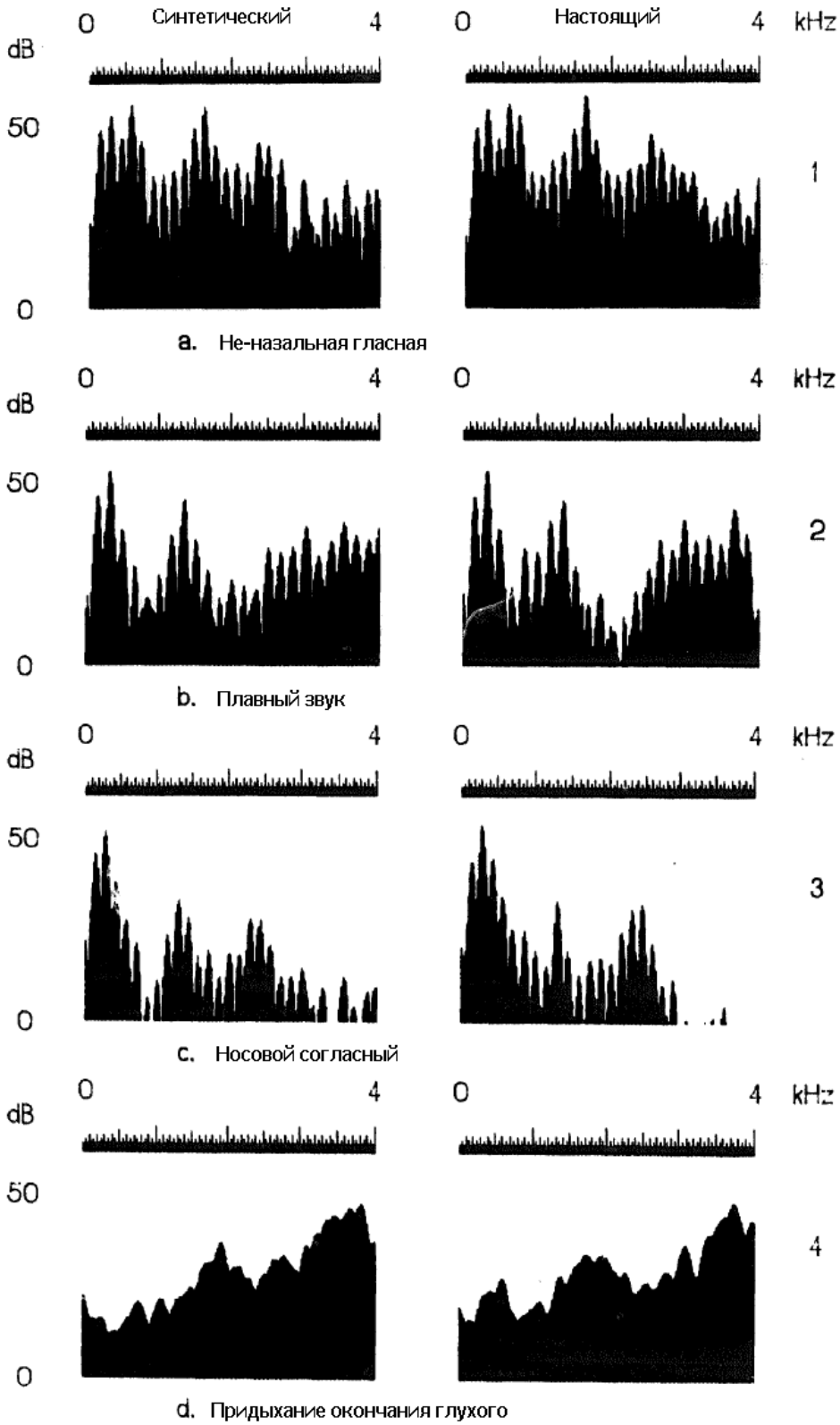



**Рис. 25. Спектральные сечения в местах, отмеченных на Рис. 24.**

Пока та же степень внимания к деталям управляющего сигнала не сделана для каскадного

синтезатора, невозможно утверждать, насколько близкая копия человеческой речи тем самым может быть достигнута. Однако, даже при создании гласных, тот факт, что каскадный синтезатор не может легко обеспечить изменения вокального усилия является, очевидно, серьёзным недостатком. Поскольку приближение плоской волны ломается на высоких частотах, никакая форма модели не представляет правильно сигнал выше 3 кГц, но параллельный синтезатор может представить его субъективно важные особенности способом более непосредственно связанным с требованиями слухового восприятия (например, путём управления спектральными уровнями в критических полосах). Для применений, требующих компоненты сигнала выше 4 кГц, могут быть легко предоставлены один или два дополнительных параллельных фиксированных частотных канала. Этот документ подтверждает, что параллельная форма намного больше подходит для согласных звуков, как и раньше объяснял Клатт [3].

Основные различия между параллельными и каскадными синтезаторами можно просуммировать следующим образом.

(i) параллельный синтезатор может использовать ту же систему источник/фильтр для моделирования всех видов звуков речи и поэтому может соотносить синтезированную речь с непосредственно измеряемыми свойствами человеческой речи, независимо от типа создаваемого звука. Эта функция, вероятно, самое сильное преимущество параллельного синтеза, в частности, для вокодерных приложений.
(ii) Для тех звуков, для которых каскадная модель хорошо подходит (например, гласные), параллельная форма нуждается в дополнительной информации управления. Однако, для синтеза по правилам расчёт необходимых амплитуд  формант тривиален и возможность варьировать их позволяет моделировать изменение вокального усилия. Последняя возможность не так удобно реализуема в каскадном синтезаторе.
(iii) Необходимо уделить пристальное внимание конструктивным особенностям параллельного синтеза для достижения хороших результатов, изложенных в этой статье. Хотя осуществление синтеза гласного более сложное, чем в каскадной форме, общая сложность заметно меньше, чем в полном каскадном синтезаторе с адекватными отдельными переделками для согласных звуков.
(iv) В параллельном синтезе нет необходимости ни в каких специальных мерах предосторожности для обеспечения примерно одинакового максимального уровня сигнала в каждом формантном генераторе. В результате легче добиться адекватной характеристики в условиях шума или погрешности квантования, чем в каскадном типе.

Моё мнение, что сочетание всех факторов, упомянутых в этой статье, даёт преимущества и по характеристике и по простоте реализации твёрдо в пользу параллельных формантных синтезаторов для исследования восприятия речи, формантных вокодеров и речевого машинного вывода.

# Литература

[1] W. Lawrence, "The synthesis of speech from signals which have a low information rate", in: W. Jackson, ed., Communication Theory, Butterworths, London, 1953.
[2] G. Fant, IVA (Royal Swedish Academy of Engineering Sciences, Stockholm), Vol. 24, 1953, pp. 331-337.
[3] D.H. Klatt, "Software for a cascade/parallel formant synthesizer" J. Acoust. Sot'. Am., Vol. 67, 1980, pp. 971-995.
[4] L.R. Rabiner, "'Digital-formant synthesizer for speech-synthesis studies", J. Acoust. Sot'. Am., Vol. 43, 1968, pp. 822-828.

[5] G. Fan( and J. Martony, "Instrumentation for parametric synthesis (OVE II)", STL-QPSR-2, Royal Institute of Technology, Stockholm, 1962, pp. 18-19.

[6] J. Anthony and W. Lawrence, "A resonance analogue speech synthesizer", Proc. 4 Int. Cong. Acoust. Copenhagen, 1962.

[7] J.D. Markel and A.H. Gray, Linear Predwtion of Speech, Springer, Berlin, 1976.

[8] J. Makhoul and L. Cosell, "LPCW: an LPC vocoder with linear predictive spectral warping", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Philadelphia, 1976, pp. 466-469.

[9] H.W. Strube, "Linear prediction on a warped frequency scale", J. Acoust. Sot'. Am., Vol. 68, 1980, pp. 1071-1076.

[10] G. Fant, The Acoustic Theory of Speech Production, Mouton, The Hague, 1960.

[11] G. Fant, "Acoustic analysis and synthesis of speech with applications to Swedish", Ericsson Techn., Vol. 1, 1959, pp. 1 - 108.

[12] H. Wakita, "Direct estimation of vocal tract shape by inverse filtering of acoustic speech waveforms", IEEE Trans. Audio and Electroacoust., Vol. AU 21, 1973, pp. 417-427.

[13] J.N. Holmes, "Requirements for speech synthesis in the frequency range 3-4 kHz", F.A.S.E. Symposium on Acoustics and Speech, Venice, Vol. I, 1981, pp. 169-172.

[14] B. Scharf, "Critical bands", in: J.V. Tobias, ed., Foundations of Modern Auditory Theory, Academic Press, New York 1970.

[15] J.L. Fianagan, Speech Analysis, Synthesis and Perception, Springer, Berlin, 1972.

[16] G. Fant, "'The source filter concept in voice production". F.A.S.E. Symposium on Acoustics and Speech, Venice, Vol. II, 1981, pp. 39-55.

[17] J.N. Holmes, "The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer", IEEE Trans. Audio and Eiectroacoust., Vol. AU 21, 1973, pp. 298-305.

[18] J.N. Holmes, "Formant excitation before and after glottal closure", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Philadelphia, 1976, pp. 39-42.

[19] A.E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels", J. Acoust. Soc. Am., Vol. 49, 1971, pp. 583-590.

[20] M. Rothenberg, R. Carlson, B. GranstrOm and J. Lindovist-Gauffin, "'A three-parameter voice source for speech synthesis", in: G. Fant, ed., Speech Communication, Almovist and Wiksell, Stockholm, Vol. 2, 1975, pp. 235-243.

[21] I.R. Titze, "Parameterization of the glottal source function and glottographic waveforms", J. Acoust. Soc. Am., Vol. 68, 1980, p. S71.

[22] M. Rothenberg, "A new inverse-filtering technique for deriving the glottal air flow waveform during voicing", J. Acoust. Soc. Am., Vol. 53, 1973, pp. 1632-1645.

[23] M.R. Schroeder, B.S. Atai and J.L. Hall, "Objective measure of certain speech signal degradations based on masking properties of human auditory perception", in: B. Lindblom and S. Ohman, eds., Frontiers of Speech Communication Research, Academic Press, London, 1979, pp. 217-229.

[24] J.N. Holmes, "Avoiding unwanted low-frequency level variations on the output of a parallel-formant synthesizer", J. Acoust. Soc. Am., Vol. 68, 1980, p. S18.

[25] D.Y. Wong, C.C. Hsiao and J.D. Markel, "Spectral mismatch due to preemphasis in LPC analysis/synthesis", IEEE Trans. Acoustics, Speech and Signal Processing, Vol. ASSP 28, 1980, pp. 263-264.

[26] J.N. Holmes, "An investigation of the volume velocity waveform at the larynx during speech by means of an inverse filter", Proc. 4 Int. Cong. Acoust., Copenhagen, 1962.

[27] M.J. Hunt, J.S. Bridle and J.N. Holmes, "Interactive digital inverse filtering and its relation to linear prediction methods", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Tulsa, 1978, pp. 15-18.

[28] J.M. Rye and J.N. Holmes, "'A versatile software parallel formant speech synthesizer", JSRU Research Report, No. 1016, 1982.

[29] J.N. Holmes, "Parallel formant vocoders", Proc. IEEE EASCON Conference, Washington,

1978, pp. 713-718.

[30] J.N. Holmes, R.D. Wright, J.W. Yates and M.W. Judd, "Extension of the JSRU synthesis-by-rule system", Proc. 9 Int. Cong. Acousr, Madrid 1977.

[31] J.N. Holmes, "Synthesis of natural-sounding speech using a formant synthesizer", in: B. Lindblom and S. Ohman, eds., Frontiers of Speech Communication Research, Academic Press, London, 1979, pp. 275-285.