Моделирование сигнала

Методы моделирования сигнала в распознавании речи

В настоящее время существует три важных направления разработок в распознавании речи. Первое, наборы разнородных параметров, которые объединяют абсолютную спектральную информацию с динамической или спектральную информацию, зависящую от времени, становятся общими. Второе, сходные методы преобразования, часто используемые при нормализации и декорреляции параметров для уменьшения вычислительных затрат, становятся популярными. Третье, задача оценки параметров сигнала объединяется с процессом распознавания речи, т.о. более "умные" статистические модели спектра сигнала могут быть оценены с помощью обратной связи. В этой работе мы рассмотрим компоненты этих алгоритмов обработки сигнала. Эти алгоритмы приводятся как часть общего обзора задачи параметризации сигнала, которая делится на три направления: измерение, преобразование и статистическое моделирование. В соответствии с этой целью, в работу включено полное математическое описание каждого алгоритма.

Кемерово 2000

Содержание

I. Введение

Параметризация аналогового сигнала речи является первым шагом в процессе распознавания речи. Различные популярные методы анализа сигнала возникли как де-факто стандарт в литературе. Эти алгоритмы предназначены для выполнения параметрического представления речевого сигнала: параметры описывающие поведение человеческой слуховой системой. Естественно, эти алгоритмы специально разработаны для увеличения производительности системы распознавания речи.

Корни многих этих методов можно проследить в ранних разработках систем распознавания с настройкой на диктора. В настоящее время значительные части разработок сосредоточены на задачах распознавания независимо от диктора, но многие из этих параметризаций продолжают быть полезными. В распознавании речи независимо от диктора, все внимание сосредоточено на разработке описаний, которые не зависят от смены диктора. Предпочтительны параметры, которые представляют собой пики спектральных энергий звука, а не детали голоса определенного диктора.

В этой работе мы предполагаем, что синтаксический метод распознавания образов в распознавании речи состоит из двух фундаментальных операций: моделирование сигнала и сетевой поиск. *Моделирование сигнала* представляет собой процесс преобразования речевых отсчетов в вектора наблюдений, которые представляют события в вероятностном пространстве. *Сетевой поиск* является задачей поиска наиболее вероятной последовательности этих событий при некотором синтаксическом ограничении. Ниже представлены обзоры популярных подходов к моделированию сигналов в распознавании речи.

1.1 Парадигма модели сигнала

Моделирование сигнала может быть разделено на четыре основные операции: спектральная форма, спектральный анализ, параметрическое преобразование и статистическое моделирование. Полная последовательность шагов приведена на рис. 1. Первые три операции являются главными задачами в цифровой обработке сигналов. Последняя операция, тем не менее, часто подразделяется на систему моделирования сигнала и систему распознавания речи.

В разрабатываемых системах моделирования сигнала существуют три главных движущих силы. Первая, поиск параметров представляющих выдающиеся аспекты речевого сигнала, предпочтительно тех, которые аналогичны тем, что используются человеческой слуховой системой. Такие параметры часто называют ощутимозначимыми (perceptually-meaningful). Вторая, желательна параметризация, которая устойчива к изменениям в канале, диктора и преобразователя. Назовем ее задачей устойчивости или неизменности. Наконец, совсем недавно стали применять параметры описывающие спектральную динамику (изменение спектра во времени). Назовем ее задачей временной корреляции. С введением метода Марковского моделирования, который способен статистически моделировать развитие сигнала во времени, параметры, объединяющие абсолютные и дифференциальные измерения спектра сигнала, становятся все более общими.

Моделирование сигнала в настоящее время занимает менее 10% процессорного времени затрачиваемого в типичном приложении распознавания речи с большим словарем. Разница в затратах процессорного времени между различными подхода-

ми в моделировании сигнала измеряется в единицах процента от общего процессорного времени. Все усилия сейчас переместились на увеличение производительности и уменьшения количества степеней свободы. Параметризация, кратко описывающая сигнал, может быть легко просчитана на оборудовании с фиксированной точкой и может быть сжата простым методом квантования, который часто используют вместо других экзотических подходов. Требования к памяти часто перевешивают любые небольшие увеличения производительности в распознавании речи предоставляемые новыми методами моделирования сигнала.

Исторически, устойчивость к окружающему шумовому фону была главной движущей силой в разработке моделей сигнала. В действительности, множество моделей сигнала в настоящее время ведут свои корни от разработок приложений для шумной окружающей среды: голосовое управление военным оборудованием (распознавание речи в рубке) [1-2] и голосовое управление телефоном [3-6]. По мере того, как технологии распознавания речи становились все более "умными", сами системы распознавания обращают больше внимания на задачу помехоустойчивости, чем на моделирование сигнала. Следовательно, часто бывает трудно ограничить расширения алгоритма моделирования сигнала.

К тому же, те модели сигнала, которые успешно используются в одном случае могут быть просто не применимы в другом. Например, в системе распознавания речи независимо от диктора, созданной для применения в конкретной среде (например, распознавание непрерывно диктуемых чисел в приложениях компьютерной телефонии), определенные типы статистических вариаций канала и дикторов могут быть успешно предсказаны и приняты во внимание. В приложениях с настройкой на конкретного диктора или приложениях идентификации диктора, важно изучение уникальных характеристик пользователя и акустической среды пользователя. Хотя кажется, что эта разница может потребовать различных подходов к моделированию сигнала, методы описанные здесь хороши для обоих типов приложений.

1.2 Терминология

Во всем документе[1] мы будем избегать всеохватывающего термина "выделение характеристик" по двум причинам.

Во-первых, очень часто этот термин носит некое сопутствующее значение, которое уменьшает его информационное наполнение. Основные характеристики речевого сигнала очень сильно зависят от контекста. Ни один алгоритм выделения характеристик не может магически нормализовать все вариации наблюдаемого сигнала без предварительной информации о контексте звука. Мы часто выбираем алгоритмы, которые сохраняет спектральную вариацию данных, вместо тех, которые пытаются удалять данные о спектре в начальной стадии обработки. Нам надо, чтобы система распознавания речи работала со статистической вариацией данных.

Во-вторых, термин "выделение характеристик" подразумевает знание того, что мы ищем в сигнале. На ранних стадиях разработки систем распознавания речи не было ничего абсолютного. Качество модели сигнала должно было определяться в контексте задачи распознавания. В действительности, наилучший набор характеристик может быть функцией алгоритма распознавания и задачи. Конечная цель — сохранить эти изменения в данных, показывающих измерения в которых основные звуковые единицы могут выражаться.

[1] Так как ни одна область деятельности человека не обходится без терминологии, в этом разделе будут даны краткие описания часто используемых терминов.

Какой же термин мы в таком случае должны использовать? Мы предпочитаем использовать в этом качестве просто *модель сигнала*. Модель сигнала будет содержать три компонента: измерения — основные спектральные и временные измерения; параметры — параметрически отсортированные и сглаженные версии этих измерений; наблюдения — выход в определенном формате статистической модели параметров. Наблюдения модели сигнала естественно связаны с технологией распознавания речи. Эти внутренние компоненты показаны на рис. 1.

Теперь опишем все эти шаги более подробно. Отметим, к сожалению, что очень мало систем распознавания речи имеют возможность полностью сравнить все вариации моделей сигнала контролируемым способом. Следовательно, мотивация выбора определенного подхода не всегда научна. Значит предположим, что эта работа охватывает наиболее используемые в настоящее время модели сигнала и дадим несколько комментариев на соответствующие авторские претензии о достоинстве их подходов. Отличные работы и другая соответствующая информация на эту тему может быть найдена в [7-26][2].

[2] Литературные ссылки в этой работе отобраны в основном по их ценности для вводного курса по этой области, а не по их достоверности в качестве оригинальной ссылки на предмет. У нас не было намерений дискредитировать определенные исследования в этой области (хотя это, наверное. неизбежно). Отличные исчерпывающие описания приведенные в работе могут быть найдены в [27-29].

II. Спектральная форма

Спектральная форма включает в себя две основные операции: *аналого-цифровое преобразование* — преобразование сигнала из волны звукового давления в цифровой сигнал; *цифровая фильтрация* — выделение главной частоты сигнала. Процесс преобразования показан на рис. 2. Полное описание основных принципов оцифровки и АЦ-преобразования может быть найдено в [30]. Мы не останавливаемся на выборе частоты оцифровки сигнала, хотя этот выбор играет важную роль в задаче моделирования сигнала.

Микрофон, используемый в процессе АЦ-преобразования, обычно вносит нежелательные примеси в сигнал, например сетевой шум (звук с частотой 50 Гц от электрической проводки), теряет часть низких и высоких частот и нелинейные искажения. АЦ-преобразователь также вносит свои собственные искажения из-за нелинейной функции передачи и колебания постоянного смещения. Пример такого сигнала показан на рис. 3. Ослабление низких и высоких частот часто вызывает проблемы с алгоритмами последовательного параметрического спектрального анализа.

Из-за ограничения частотного диапазона аналоговых телекоммуникационных каналов и широкого использования 8 кГц оцифровки речи в цифровой телефонии, наиболее популярной частотой оцифровки речевого сигнала в телекоммуникациях является та же частота. Последующее возникновение широкополосных цифровых сетей вызовет появление новых телекоммуникационных приложений, которые будут использовать повышенную частоту оцифровки речевого сигнала. В других приложениях, в которых подсистема распознавания речи имеет доступ к высококачественному представлению речевого сигнала, используются следующие частоты 10 кГц, 12 кГц и 16 кГц. Такие частоты дают лучшее временное и частотное разрешение [31].

Главная цель процесса оцифровки заключается в получении данных речевого сигнала с высоким отношением *Сигнал/Шум*. В настоящее время телекоммуникационные системы обеспечивают значение этого коэффициента порядка 30 dB для приложений распознавания речи, что более чем достаточно для получения высокой производительности таких приложений. Изменения в преобразователях, каналах и фоновом шуме, тем не менее, вызывают определенные проблемы.

После преобразования сигнала, заключительным шагом цифровой постфильтрации является применение FIR фильтра (Finite Impulse Responce):

$$H_{pre}(z) = \sum_{k=0}^{N_{pre}} a_{pre}(k) z^{-k} \qquad (1)$$

Обычно используется одно-коэффициентный цифровой фильтр, известный как *взвешивающий* фильтр:

$$H_{pre}(z) = a_{pre} z^{-1} \qquad (2)$$

Стандартным диапазоном значений для a_{pre} — [-1.0, -0.4]. Значения, близкие к -1.0, которые могут быть эффективно обсчитаны с помощью аппаратуры с фиксированной точкой, наиболее часто используются в распознавании речи. Диапазон частот для взвешивающего фильтра (2) показан на рис. 4. Взвешивающий фильтр повышает спектр сигнала приблизительно на 20 dB (приращение величины по частоте).

Существует два общих объяснения преимуществ использования данного фильтра. Первое, секция речевого сигнала содержащая саму речь в действительности имеет негативный спектральный уклон (ослабление) приблизительно на 20 dB

из-за физиологических характеристик речевого тракта [28, 31]. Взвешивающий фильтр выполняет смещение этого уклона перед анализом спектра, повышая, таким образом, эффективность этого анализа [31, 32].

Альтернативное объяснение построено на том, что слух более чувствителен в регионе спектра выше 1 кГц. Взвешивающий фильтр усиливает эту область спектра, помогая алгоритму спектрального анализа в моделировании наиболее важных аспектов речевого спектра [31] (для подробностей обратитесь к разделу 3.3.4).

Отметим также, что такой взвешивающий фильтр поднимает частоты выше 5 кГц, области к которой слуховая система менее чувствительна. Частоты выше 5 кГц действительно заглушаются голосовым трактом и обычно имеют малый вес в системах распознавания речи.

Но был предложен более изощренный взвешивающий алгоритм. Один такой примечательный подход — адаптивное взвешивание, в котором спектральный уклон автоматически выравнивается [31] перед спектральным анализом. Другие алгоритмы используют формовые фильтры, которые заглушают области спектра известные своим шумом [33, 34]. Теперь алгоритмы классификации речь/шум базируются на адаптивной фильтрации [35]. Тем не менее, ни один из этих методов не имеет широкого применения в приложениях распознавания речи. В действительности, многие системы распознавания речи не используют стадию взвешивания и компенсируют спектральный уклон его учетом в статистической модели распознавания речи (см. раздел VI).

III. Спектральный анализ

В педагогических целях классифицируем типы спектральных измерений используемых в системах распознавания речи: *энергия* — измеряет рост спектральной (или временной) энергии сигнала; *спектральная амплитуда* — измеряет энергию на определенном интервале частот спектра. Типичный набор параметров в распознавании речи будет включать каждое из этих измерений. С недавних пор появилось возрождение интереса[3] к основной частоте, как к prosodic характеристике [36], для ее использования в распознавании речи для тональных языков (например, для китайского) или для многотональных языков (например, для японского), и для идентификации пользователя [37]. Кратко опишем поиск основной частоты и энергии и сфокусируемся на оценке спектральной амплитуды.

3.1 Основная частота

Основная частота[4] определена как частота с которой вибрируют голосовые связки во время речевого процесса [38, 39]. Основная частота (f_0) довольно долго была сложным параметром для ее определения из речевого сигнала. Раньше ею пренебрегали в системах распознавания речи по ряду причин, включая высокую вычислительную нагрузку необходимую для точной оценки; беспокойство, что ненадежная оценка может послужить барьером для достижения высокой производительности; сложность в комплексном характеристическом взаимодействии между основной частотой и suprasegmental phenomena.

В настоящее время существует четыре основных класса подобных алгоритмов. Одним из первых появившихся алгоритмов является алгоритм, который использует множество периодических измерений сигнала для определения наличия голоса в сегменте и основной частоты. В специальной литературе этот алгоритм известен как алгоритм Голда-Рабинера (Gold-Rabiner) [40, 41]. Этот алгоритм все еще популярен в основном из-за его простоты и удобства надежной реализации. К сожалению, он не универсален.

Второй принадлежит Национальному Агентству Безопасности США, он является частью программы разработки защищенных цифровых телефонов работающих на голосовом кодировании с малым потоком данных и очень устойчив [42, 43]. Этот алгоритм базируется на функции разности средних значений [32] и на дискриминантном анализе многочисленных измерений сигнала. Является государственным стандартом США.

Третий класс алгоритмов похож по природе на предыдущий и базируется на концепциях динамического программирования [44]. Эти алгоритмы показывают высокую производительность в широком диапазоне окружающих сред, включая шумные телекоммуникационные каналы. Этот класс алгоритмов использует изощренную процедуру оптимизации, которая оценивает различные измерения корреляции и изменения спектра сигнала и одновременно вычисляет оптимальную основную частоту и определяет наличие голоса в сегменте.

---

[3] Естественно, основная частота довольно редко используется в практических системах распознавания речи.

[4] Этот раздел является обзором существующих работ по этому направлению. Детали алгоритмов не вошли в данную работу. Для этого советуем обратиться к [38, 39].

Наконец, мы подошли к алгоритму, который довольно редко используется в речевых системах реального времени, но часто применяется в исследованиях. Это алгоритм оперирующий кэпстром речевого сигнала [45]. Он популярен в настоящее время, как точный метод оценки основной частоты в тихих лабораторных условиях.

Для представления основной частоты часто используют логарифмическую шкалу вместо линейной. Для удобства определим измерение основной частоты как:

$$F(n) = \log_{10}(f_0(n)) \tag{3}$$

где n — дискретное время.

Обычно для голосового сегмента $50\text{Hz} \le f_0 \le 500\text{Hz}$. Для сегмента без голоса основная частота не определена и для удобства $F \equiv 0$. Часто основная частота нормализуется средним значением диктора f_o. или каким-нибудь физиологически мотивированным преобразованием номинального значения соответствующего сегмента с голосом.

3.2 Энергия

Использование какого-нибудь вида измерения энергии в распознавании речи в настоящее время является обычным явлением. *Энергия* довольно просто подсчитывается по формуле:

$$P(n) = \frac{1}{N_S}\sum_{m=0}^{N_S-1}\left(w(m)\cdot s\left(n - \frac{N_S}{2} + m\right)\right)^2 \tag{4}$$

где N_S — количество отсчетов сигнала использованных для подсчета энергии, s(n) — массив значений сигнала, w(m) — взвешивающая функция, n — индекс отсчета (дискретное время) центра окна.

Вместо того, чтобы использовать подсчитанную энергию напрямую, многие системы распознавания речи используют логарифм энергии умноженный на 10, заданный как энергия в децибелах, для эмуляции логарифмической реакции человеческой слуховой системы [46].

Взвешивающая функция в уравнении (4) работает как оконная функция. Теория окна была раньше очень активным направлением исследований в области цифровой обработки сигнала [31, 32]. Существует очень много типов окон, включая прямоугольное, Хамминга (Hamming), Ханнинга (Hanning), Блэкмана (Blackman), Бартлетта (Bartlett) и Кайзера (Kaiser). В настоящее время в распознавании речи используется исключительно окно Хамминга, причем окно Хамминга — это особый случай окна Ханнинга. В общем случае, окно Ханнинга задается так:

$$w(n) = \frac{\alpha_w - (1-\alpha_w)\cdot\cos\left(\frac{2\pi n}{N_S - 1}\right)}{\beta_w} \tag{5}$$

для $0 \le n \le N_S$ и $w(n) \equiv 0$ во всех других случаях; α_w — константа окна в диапазоне [0,1]; N_S — длительность окна в отсчетах. На практике $\alpha_w = 0.54$.

β_w — константа нормализации, заданная так, что значение квадратного корня окна равно единице.

$$\beta_w = \sqrt{\frac{1}{N_S}\sum_{n=0}^{N_S-1} w^2(n)} \qquad (6)$$

На рис. 5. показаны различные реализации окна Ханнинга.

На практике желательно нормализовать окно таким образом, чтобы энергия сигнала до наложения окна была бы приблизительно равна энергии сигнала после наложения окна. Уравнение (6) описывает такую нормализационную постоянную. Такой тип нормализации особенно удобен для применения при аппаратной обработке с фиксированной точкой. Обратите внимание, что вычислительные затраты на обработку окна относительно малы, так как оконные коэффициенты рассчитываются при инициализации.

Окно нужно для взвешивания отсчетов по отношению к центру окна. Эта характеристика совместно с перекрывающим анализом описанным далее выполняет важную функцию для получения сглаженных изменяющихся параметрических оценок. Очень важно чтобы ширина основной доли частоты окна была минимальна или чтобы оконный процесс мог иметь убывающий эффект при сегментном спектральном анализе. Хорошее описание по данному разделу можно найти в [31, 32, 47].

Энергия, как и большинство параметров системы распознавания речи, включая основную частоту, подсчитывается кадр за кадром. *Размер кадра* (T_f) задан как время (в секундах) в течении которого набор параметров верен. *Период кадра* определяет время между успешными подсчетами параметров. *Частота кадров*, еще один общий термин, это число кадров обработанных за секунду (в Герцах).

В уравнении (4), n — это длительность кадра в отсчетах. В практических системах длительность кадров выбирается в диапазоне от 10 мс до 20 мс. Значения из этого диапазона представляют выбор оптимального решения между частотой изменения спектра и сложностью системы. Соответствующая длительность кадра полностью зависит от скорости произношения в системах распознавания речи (скорости изменения вокального тракта). До тех пор пока речь (например, стоповые согласные или дифтонги) имеет острые спектральные переходы, которые могут заканчиваться спектральными пиками со сдвигами на 80 Гц/мс [31], кадры длительностью менее 8 мс не будут использоваться.

Тем не менее, важен интервал на котором вычисляется энергия. Количество отсчетов (N_S) использованных для подсчета известно как размер окна (в отсчетах). Размер окна (T_w) обычно измеряется в единицах времени (секундах).

Размер окна управляет средней, или сглаженной, суммой используемой в подсчете энергии. Длительность кадра и размер окна вместе управляют частотой с которой значения энергии показывают динамику сигнала [32]. Длительность кадра и размер окна обычно объединены в пары: размер окна в 30 мс используется вместе с длительностью кадра 20 мс, пока размер окна в 20 мс используется при длительности кадра 10 мс. В общем говоря, так как малая длительность кадра используется для захвата быстрой динамики спектра, то размер окна должен быть также малым, чтобы детали спектра не были чрезмерно сглажены.

Обработка кадров базируется на анализе показано на рис. 6. На этот тип анализа часто ссылаются как на *перекрывающий анализ*, потому что для каждого нового кадра изменяется только доля сигнала. Сумма перекрытия в известной мере управляет скоростью изменения параметров от кадра к кадру. Процент перекрытия можно вычислить по формуле:

$$\%\text{Overlap} = \frac{T_w - T_f}{T_w} \times 100\% \qquad (7)$$

где T_w — размер окна в секундах, T_f — длительность кадра. Если $T_w<T_f$, то процент перекрытия равен нулю.

Комбинация кадра длительностью 20 мс и окна размеров 30 мс соответствует 33% перекрытия. Некоторые системы используют 66% перекрытие. Одна причина такой большой суммы перекрытия в том, чтобы уменьшить сумму шума вносимого в измерения такими артефактами как размещение окна и нестационарный шум канала [32]. С другой стороны, чрезмерно сглаженные оценки могут скрыть истинные вариации сигнала.

Покадровый подсчет энергии может быть сделан рекурсивно [32]. Этот метод рассматривается как операция фильтрации квадратной амплитуды сигнала:

$$P(n) = -\sum_{i=1}^{N_a} a_{pw}(i) \cdot P(n-i) + \sum_{j=0}^{N_b} b_{pw}(j) \cdot s^2(n-j) \qquad (8)$$

где $\{a_{pw}\}$ и $\{b_{pw}\}$ — коэффициенты цифрового понижения уровня сигнала при проходе через фильтр. Чаще всего используются для первого фильтра (leaky integrator) $N_a=1$, $N_b=0$ или для второго (биквадратная секция) — $N_a=2$, $N_b=2$. Создание системы в соответствии с уравнением (8), выдающей сглаженные оценки энергии — классическая задача. Хорошее обсуждение этой задачи можно найти в [48].

На рис. 7 показано использование этих параметров. Внизу рисунка показан речевой сигнал. Первые четыре графика показывают контуры энергии для:

a) T_f = 5 мс, T_w = 10 мс;

b) T_f = 10 мс, T_w = 20 мс;

c) T_f = 20 мс, T_w = 30 мс;

d) T_f = 20 мс, T_w = 30 мс, использовалось окно Хамминга;

e) T_f = 20 мс, T_w = 60 мс;

f) Во второй секции был применен рекурсивный фильтр отсекающий 50 Гц шум.

Использование окна Хамминга, как показано на рис. 7, помогает получить сглаженные оценки энергии в регионах, где она быстро изменялась (обратите внимание на точку на рисунке указанную стрелкой). Обратите внимание, что при t=0.3 секунды есть небольшое повышение уровня энергии. Это повышение энергии видно только в случаях (a) и (e), в двух анализах с наибольшей возможностью обнаружения быстрых изменений в энергии сигнала.

Рекурсивный метод, показанный на рис. 7, выдает колебательный контур энергии. Второй фильтр не способен достаточно ослаблять высокие частоты в амплитуде сигнала / мгновенном контуре энергии. Значит, выходные данные имеют тенденцию быть слишком чувствительными к моментальному уровню энергии сигнала. По этой причине, такие фильтры часто используют после покадрового анализа, так как этом случае контур энергии слишком сглажен, чтобы с ним работать. Для аккуратной разработки этих схем требуется, чтобы скорость адаптации соответствовала приложению.

Рекурсивная формулировка используется при применении этих алгоритмов для адаптивного управления уровнем, оценки пиков энергии сигнала. и определении конечной точки сигнала. Уравнение (8) может быть применено после уравнения (4) для получения дополнительного сглаживания оценок энергии.

3.3 Спектральный анализ

Существует шесть основных классов алгоритмов спектрального анализа, использующихся в настоящее время в системах распознавания речи, все эти классы показаны на рис. 8. Методы набора фильтров (применяемые в аналоговых схемах) — были исторически первыми методами. Методы линейного предсказания были введены в 1970-х и доминировали до начала 1980-х. В настоящее время широко распространены методы Фурье-преобразования и линейного предсказания. В этом разделе мы обсудим все эти методы, начиная с набора цифровых фильтров.

3.1.1 Цифровой банк фильтров

Цифровой банк фильтров — одно из основных понятий в обработке речи. Банк фильтров можно рассматривать как упрощенную модель начальных этапов человеческой слуховой системы. Существует две мотивации для использования банка фильтров.

Первая, положение максимального смещения вдоль базилярной мембраны для таких стимулов как чистые тоны пропорционально логарифму частоты тона. Эта гипотеза является частью теории слуха называемой "теорией местоположения" [49].

Вторая, эксперименты с человеческим восприятием показали, что частоты сложных звуков в пределах определенной ширины некоторой номинальной частоты могут быть индивидуально выделены. Когда один из компонентов этого звука выпадает из полосы частот, его можно выделить. Мы рассматриваем такую полосу частот, как критическую полосу частот [50], обычно на 10..20% от центральной частоты звука.

Зададим распределение акустической частоты (f) по шкале "ощущаемых" частот следующим образом [51]:

$$\mathrm{Bark} = 13 \cdot \mathrm{a\,tan}\left(\frac{0.76 \cdot \mathrm{f}}{1000}\right) + 3.5 \cdot \mathrm{a\,tan}\left(\frac{\mathrm{f}^2}{(7500)^2}\right) \qquad (9)$$

Единицы измерения такой шкалы "ощущаемых" частот называются критическими частотами полосы или Bark-шкалой. Такая шкала показана на рис. 9a.

Более популярное приближение к этому типу распределения в распознавании речи известно как Mel-шкала [51]:

$$\mathrm{mel\_frequency} = 2595 \cdot \log_{10}\left(1 + \frac{\mathrm{f}}{700.0}\right) \qquad (10)$$

Mel-шкала пытается отобразить воспринятую частоту тона или высоту на линейную шкалу. Такая шкала показана на рис. 9b. На диапазоне частот 0..1000 Гц шкала линейна, после 1000 Гц — логарифмическая.

Выражение для критической полосы выглядит так:

$$\mathrm{BW}_{\mathrm{critical}} = 25 + 75 \cdot \left(1 + 1.4 \cdot \left(\frac{\mathrm{f}}{1000}\right)^2\right)^{0.69} \qquad (11)$$

Это преобразование может быть использовано для вычисления полосы на шкале воспринимаемых частот для фильтров на данной частоте по Bark- или Mel-шкалам. Функция критической полосы показана на рис. 9c.

Bark-шкалу и Mel-шкалу можно рассматривать как преобразование шкалы частот в линейную воспринимаемую шкалу. Объединение этих двух теорий известно как банк фильтров критической полосы. *Банк фильтров критической полосы* — это набор линейных фазовых FIR фильтров линейно размещенных вдоль Bark- или Mel-шкалы. Ширина полосы частот фильтров выбирается равной критической ширине полосы частот для соответствующего центральной частоты.

Один такой банк [52], ставший стандартом, показан в таб. 1.

Центральные частоты и ширина полос для таких фильтров расположены во втором и третьем столбцах таб. 1. Центральные частоты соответствуют тем частотам для которых уравнение (9) выдает целочисленный индекс. Ширина полосы частот подсчитывается с помощью уравнения (11).

Во многих приложениях распознавания речи первый фильтр не используют, так как его диапазон лежит вне возможностей АЦ-преобразователя. Часто оцифрованные данные содержат в соответствующем диапазоне частот слишком много шума. Ухудшенная телефоном речь обычно обрабатывается банком из 16-ти фильтров (индексы — 2..17).

Другой важный банк фильтров упоминаемый в литературе по распознаванию речи — банк фильтров базирующийся на Mel-шкале. Отношение частоты к диапазону частот для такого банка фильтров [26] дано в четвертом и пятом столбцах таб. 1. В этом случае, десять фильтров линейно распределены по диапазону 100..1000 Гц. Выше 1000 Гц расположено пять фильтров на каждом удвоении частоты (октавы). Эти фильтры размещены логарифмически (одинаковое расстояние по логарифмической шкале). Диапазон частот взят таким, что точка 3 dB является половиной между текущим значением и следующим или предыдущем. Затененные ячейки таблицы приведены только для сравнения. Обычно используется только первые 20 из банка фильтров.

Каждый фильтр в банке работает как линейный фазовый фильтр, таким образом групповая задержка всех фильтров равна нулю и выходные сигналы с фильтров будут синхронизированы во времени. Уравнения для случая линейного фазового фильтра могут быть объединены так:

$$S_i(n) = \sum_{j=-\frac{N_{FB_i}-1}{2}}^{\frac{N_{FB_i}-1}{2}} a_{FB_i}(j) \cdot s(n+j) \qquad (12)$$

где $a_{FB_i}(j)$ — j-ый коэффициент i-го фильтра критического диапазона. Для линейного фазового фильтра порядок фильтра обычно четный.

Пример обработки речи двумя фильтрами из такого банка показан на рис. 10. Показан речевой сигнал и выходные сигналы двух фильтров на 250 Гц и 2500 Гц соответственно. Обратите внимание, что энергия выходного сигнала варьируется в зависимости от типа произносимой речи. Это и есть основное качество банка фильтров: определенный фильтр может быть согласован с определенным классом произносимой речи.

Выходной сигнал с фильтра обычно обрабатывается любым из методов оценки энергии, которые были обсуждены выше. Цифровой банк фильтров наиболее часто используется в системах, которые пытаются эмулировать работу человеческой слуховой системы [53, 54]. Последовательные вычисления особенно удобны для последующей обработки в таких приложениях.

Выходным значением этого вида анализа будет вектор значений энергии (или пар энергия/частота) для каждого кадра данных. Эти данные обычно объединяются с другими параметрами, такими как общая энергия, для формирования вектора измерений сигнала. Банк фильтров разлагает сигнал на дискретный набор спектральных образов, которые содержат информацию аналогичную той, которая присутствует на верхних уровнях обработки в человеческой слуховой системе. Из-за того, что анализ базируется на линейной обработке (нелинейная техника будет обсуждена в разделе 3.3.4), он довольно устойчив к воздействию фонового шума.

Завершим этот раздел экскурсом в историю. За долго до того, как компьютеры стали способны выполнять сложные математические операции в реальном времени, аналоговые банки фильтров, похожие на вышеописанный цифровой банк фильтров, использовались в системах распознавания речи. Банки фильтров создавались из дискретных компонентов и нуждались в тщательной настройке путем подбора резисторов и конденсаторов. В те времена разработчики мечтали о днях, когда параметры системы распознавания речи можно будет настраивать программно. Аналоговый банк фильтров — один из самых старых методов используемых в распознавании речи, также самый дешевый метод до настоящего времени.

3.3.2 Банк Фурье-фильтров

Выше обсуждались преимущества использования разнообразно размещенных образцов частот. Один из легких и наиболее эффективных путей вычисления модели сигнала банка разнообразно размещенных фильтров — просто выполнить Фурье-преобразование сигнала и получить данные о желаемых частотах. *Дискретное Фурье-преобразование* (Discrete Fourier Transform) сигнала определяется так:

$$S(f) = \sum_{n=0}^{H_s - 1} s(n) \cdot e^{-j\left(\frac{2\pi f}{f_s}\right)n} \qquad (13)$$

где f — частота в Гц, f_s — желаемая частота, N_s — длительность окна в отсчетах.

Банк фильтров, использующий уравнение (13), может быть применен к спектру на частотах, которые указаны в таб. 1. Тем не менее, довольно часто полученный спектр имеет лучшее разрешение, чем указано в таб. 1, и каждый выходной сигнал банка фильтров (энергетическая спектральная величина) вычисляется как взвешенная сумма смежных величин:

$$S_{avg}(f) = \frac{1}{N_{os}} \sum_{n=0}^{N_{os}} w_{FB}(n) \cdot S(f + \delta f(f, n)) \qquad (14)$$

где N_{os} — количество отсчетов, использованных для подсчета среднего значения; $w_{FB}(n)$ — взвешивающая функция; $\delta f(f,n)$ — некоторая функция описывающая соседние с f частоты, которые были использованы при подсчете. Обратите внимание, что уравнение (14) — это всего лишь один из методов применения спектральной сглаживающей функции.

Усреднение часто применяется в области Mel-шкалы частот, если используется дискретное Фурье-преобразование, так как прирост вычислительных затрат при этом довольно мал. Усреднение также применяется в области логарифмических значений энергии, а не спектральных амплитуд. Преимущества использования усредненных значений в спектральном анализе показаны на рис. 11.

Быстрое Фурье-преобразование (Fast Fourier Transform) [55] также можно применить как альтернативный метод вычисления спектра сигнала. Быстрое Фурье-преобразование — вычислительно более эффективное применение дискретного

Фурье-преобразования, если принять во внимание, что спектр должен оцениваться на дискретном наборе частот, которые умножаются на f_s / N. Такие частоты называются ортогональными. Принципиальное преимущество быстрого Фурье-преобразования в том, что это преобразование очень быстрое: необходимо выполнить приблизительно $N \cdot \log(N)$ сложений и $\frac{N \cdot \log(N)}{2}$ умножений. К примеру, дискретное Фурье-преобразование требует примерно N^2 операций. Недостаток быстрого преобразования в том, что нелинейное распределение частот, см. таб. 1, должно учитывать ограничения налагаемые методом.

Часто применяют дополнительную обработку сигнала. Отчасти базируясь на нашем схематичном знании человеческого восприятия, мы предположим, что область больших амплитуд спектра взвешивается более интенсивно в человеческой слуховой системе, чем область малых амплитуд. В шумной обстановке, шум случайным образом влияет на наши оценки в области малых амплитуд спектра. Говоря по другому, мы более уверены в правильности наших оценок именно в области высоких амплитуд спектра.

По этим причинам мы всегда ограничиваем динамический диапазон спектра, как показано на рис. 12. Такое ограничение называется *порогом динамического диапазона*. Вместо того, чтобы использовать шумные оценки диапазона малых амплитуд спектра, мы просто не учитываем оценки ниже определенного порога. В методах, использующих Фурье-преобразование, это напрямую используется в пороговой функции спектральной величины, измеряемой в dB.

Важно, что спектральная оболочка имеет относительно плоскую форму перед применением таких пороговых алгоритмов. В противном случае, полезные порции спектра с низкой энергией могут быть ошибочно устранены. Вспомните, что, так как спектр речевого сигнала проседает на 20 dB, порог для низкочастотных энергий, где разница пиков спектральных амплитуд достаточно велика, может легко отсечь полезную энергию сигнала на высоких частотах. Позже мы обсудим боле изощренные методы с ограничением спектра, базирующиеся на методах параметрических моделей. Читатель может представить себе утилиту, позволяющую варьировать порог динамического диапазона как функцию уровня фонового шума в спектре или как функцию локальных спектральных пиков для локализации уровня фонового шума.

3.3.3 Кэпстральные коэффициенты

С момента их введения в начале 1970-х, на методы гомоморфической обработки сигнала [27] обратили повышенное внимание. Гомоморфические системы — это класс нелинейных систем, которые подчиняются общим принципам суперпозиции. Линейные системы, описанные выше, — специальный случай гомоморфических систем. Мотивация для использования гомоморфической обработки в анализе речи показана на рис. 13.

В обработке речи, гомоморфическая система, которую мы ищем, должна иметь следующее свойство:

$$D\left[\{x_1(n)\}^{\alpha} \bullet \{x_2(n)\}^{\beta}\right] = \alpha D[x_2(n)] + \beta D[x_2(n)]$$

Это операция суперпозиционного типа относящаяся к умножению, суммированию и экспоненте. Логарифмическая функция, конечно, подчиняется общему свойству суперпозиции.

Гомоморфические системы полезны при обработке звука [27], так как они предлагают методологию по отделению возбуждающего сигнала от голосового тракта. Текущие подходы в распознавании речи в основном занимаются моделированием характеристик вокального тракта. В линейной акустической модели речевого тракта, композитный спектр сигнала, измеренный с помощью Фурье-преобразования, содержит возбуждающий сигнал отфильтрованный линейным фильтром изменяющимся во времени, представляющем собой голосовой тракт.

Процесс разделения двух компонентов сигнала называют разверткой и описывают так:

$$s(n) = g(n) \otimes v(n)$$

где g(n) — возбуждающий сигнал; v(n) — импульсный ответ вокального тракта; ⊗ — развертка. Область частот этого процесса:

$$S(f) = G(f) \cdot V(f)$$

Если мы возьмем логарифм обоих частей, то получим:

$$\log(S(f)) = \log(G(f) \cdot V(f)) = \log(G(f)) + \log(V(f))$$

Следовательно, в области логарифмов, возбуждение и голосовой тракт наложены друг на друга и могут быть разделены, с помощью обработки сигнала (как минимум, в теории).

Для подсчета кэпстра, мы сначала подсчитываем логарифмы спектральных величин (усредненных, если необходимо) по уравнению (14). Потом подсчитываем инверсное Фурье-преобразование логарифмического спектра:

$$c(n) = \frac{1}{N_s} \sum_{k=0}^{N_s - 1} \log_{10} \left| S_{avg}(k) \right| \cdot e^{j \frac{2\pi}{N_s} kn}, \quad 0 \le \pi \le N_s - 1 \qquad (15)$$

где c(n) — это *кэпстр*. Мы ссылаемся на кэпстральные коэффициенты, которые подсчитываются с помощью Фурье-преобразования (или аналогового банка фильтров) как на *кэпстральные коэффициенты Фурье-преобразования*.

Обратите внимание на то, что c(0) в уравнении (15) представляет среднее значение спектра или RMS (root mean square) значения сигнала. Изначально, этот термин был важной частью вектора кэпстральных параметров. Позднее было замечено, что измерения абсолютной энергии сигнала были чем-то ненадежным[5] и использование c(0) не надо выделять. С того времени как параметрический вектор различных альтернативных измерений стал использоваться на других стадиях обработки сигнала, использование c(0) стало не нужным. Начиная с этого места, мы исключаем этот термин из нашего обсуждения последовательности кэпстральных коэффициентов.

Уравнение (15) также рассматривается как инверсное дискретное Фурье-преобразование логарифмического спектра. Вспомним, что логарифмическая величина спектра — реальная симметрическая функция. Следовательно, уравнение (15) можно упростить так:

$$c(n) = \frac{2}{N_s} \sum_{k=1}^{N_s} S_{avg}(I(k)) \cos\left(\frac{2\pi}{N_s} kn \right) \qquad (16)$$

[5] Некоторые системы [14] все еще используют некоторую форму абсолютной энергии с различными измерениями нормализованной энергии.

где c(n) в этом уравнении обычно исключается на порядок гораздо меньший чем N_s.; I(k) — функция распределения, которая сопоставляет целое k соответствующему образцу S_{avg}. Для удобства S_{avg} может быть также подсчитан, используя перегруженный алгоритм быстрого Фурье-преобразования, вместо необычно размещенного дискретного Фурье-преобразования.

Обратим внимание, что наше понятие кэпстра, используемое в обработке речи, немного отличается от классического определения комплексного кэпстра, упоминаемого в литературе [27, 28]. Тем не менее, наше определение несет всю значимую информацию для распознавания речи. Кэпстр, определенный в уравнении (16), может быть легко преобразован в кэпстр Mel-шкалы, применяя Фурье-преобразование для соответственно размещенных частот.

Кэпстры двух разных речевых сигналов показан на рис. 14. На рис. 14(a) и рис. 14(c) показаны фрагменты звука без речи и с ней соответственно. На рис. 14(b) и рис. 14(d) показаны соответствующие кэпстры. Малые отклонения кэпстра соответствует краткосрочному соотношению в речевом сигнале (сглаживание спектральной формы). Локальный максимум на больших отклонениях на рис. 14(d) показывает на периодичность (возбуждающая информация). Кэпстр на рис. 14(b) неголосового сегмента не показывает никакой периодичности. В спектральном анализе приложений распознавания речи обычно используют малый порядок (n<20).

Так как кэпстр подсчитывается с помощью нелинейного оператора (логарифмической функции), обычно подразумевают чувствительность к определенным типам шума и искажениям сигнала. Пока есть сложные шумовые процессы, фоновой акустический шум может вызывать проблемы. Кэпстральные параметры, получаемые из спектральных оценочных функций высокого разрешения или параметров спектра, часто используются приложениями для шумных помещений.

3.3.4 Коэффициенты линейного предсказания

Теперь вернемся от методов Фурье-преобразования базирующихся на линейном спектральном анализе к классу методов параметрического моделирования, которые пытаются оптимально моделировать спектр как авторегрессионый процесс. Очень трудно преувеличить влияние параметрической модели введенной в начале 1970-х [56, 57]. Позже, практически каждая система обработки речи использовала подобный алгоритм, включая приложения для распознавания речи, ее компрессии и идентификации пользователя. Хотя в настоящее время параметрические модели менее популярны, они все еще широко распространены в системах компрессии. Параметрические модели послужили импульсом для перехода к значительно более мощным методам статистического моделирования. в распознавании речи. В этом разделе мы обсудим вычисление параметрической модели основанной на теории наименьшей квадратичной ошибки. Этот метод известен как линейное предсказание (Linear Prediction).

Корни линейного предсказания, как и алгоритма наименьшей квадратичной ошибки, идут из различных областей: задачи системной идентификации в современных управляющих системах, временной анализ для экономических приложений, методы максимальной энтропии, квантовая физика, геофизика, адаптивная фильтрация и спектральная оценка в обработке сигнала. Теория линейного предсказания хорошо освещена в литературе [31, 32, 57]. Ниже мы кратко опишем механизм модели линейного предсказания и ее применение в распознавании речи.

Дан сигнал s(n), мы пытаемся смоделировать сигнал как линейную комбинацию предыдущих отсчетов сигнала. Зададим собственную модель сигнала:

$$s(n) = -\sum_{i=1}^{N_{LP}} a_{LP}(i) \cdot s(n-i) + e(n) \tag{17}$$

где N_{LP} — количество коэффициентов модели (порядок предсказания); a_{LP} — коэффициенты линейного предсказания; e(n) — функция ошибки модели (различие между предсказанным значением и реально измеренным значением).

Одно очевидное достоинство этой модели — она точна, так как мы должны иметь возможность предсказать будущие значения сигнала, основываясь на текущем наборе измерений[6]. Ошибка говорит нам о качестве модели (если ошибка мала, то модель точная). Также можно показать, что модель линейного предсказания эффективно моделирует спектр сигнала как сглаженный спектр [31].

Уравнение (17) может быть переписано в z-представлении и показано как операция линейной фильтрации:

$$E(z) = H_{LP}(z) \cdot S(z)$$

где E(z) и S(z) — z-представление ошибочного сигнала и речевого сигнала соответственно, а

$$H_{LP}(z) = 1 + \sum_{i=1}^{N_{LP}} a_{LP}(i) \cdot z^{-i}$$

или

$$H_{LP}(z) = \sum_{i=1}^{N_{LP}} a_{LP}(i) \cdot z^{-i} \tag{18}$$

где $a_{LP}(0) \equiv 1$ — инверсный фильтр линейного предсказания.

Учитывая то, что квадратичная ошибка должна быть минимальна (желателен поиск решения, которое дает нам минимальную ошибку энергии), коэффициенты уравнения (18), кроме $a_{LP}(0)$, могут быть получены из следующей матрицы:

$$\bar{a}_{LP} = \underline{\Phi}^{-1} \bar{\phi} \tag{19}$$

где

$$\bar{a}_{LP} = [a_{LP}(1), \quad \ldots, \quad a_{LP}(N_{LP})] \tag{20}$$

$$\underline{\Phi} = \begin{bmatrix} \phi_n(1,1) & \phi_n(1,2) & \ldots & \phi_n(1,N_{LP}) \\ \phi_n(2,1) & \phi_n(2,2) & \ldots & \phi_n(2,N_{LP}) \\ \ldots & \ldots & \ldots & \ldots \\ \phi_n(N_{LP},1) & \phi_n(N_{LP},2) & \ldots & \phi_n(N_{LP},N_{LP}) \end{bmatrix}, \tag{21}$$

$$\bar{\phi} = [\phi_n(1,0), \quad \phi_n(2,0), \quad \ldots, \quad \phi_n(N_{LP},0)], \tag{22}$$

и

$$\phi_n(j,k) = \frac{1}{N_s} \sum_{m=0}^{N_s-1} s(n+m-j) \cdot s(n+m-k) \tag{23}$$

[6] Наивный читатель может предположить использование таких алгоритмах в экономике. И не ошибется, такие алгоритмы изначально использовались в экономических приложениях.

Эти уравнения называются ковариантным методом, где $\underline{\Phi}$ — ковариантная матрица, $\phi_n(j,k)$ — ковариантная функция для s(n).

Существует три основных пути подсчета коэффициентов предсказания: ковариантные методы базируются на ковариантной матрице (также известной как чистый метод наименьших квадратов) и периодические (или гармонические) методы. Более подробно различия в этих подходах можно посмотреть в [31, 57]. Автокорреляционный метод в распознавании речи используется почти исключительно из-за его вычислительной эффективности и присущей ему стабильности. Автокорреляционный метод всегда производит фильтр предсказания, чьи нули лежат внутри круга в z-плоскости.

Модифицируем уравнение (23) для автокорреляционного метода:

$$\phi_n(j,k) = \phi_n(0,|j-k|) \tag{24}$$

или

$$R_n(k) = \frac{1}{N_s}\sum_{m=0}^{N_s-1-k} s(n+m)\cdot s(n+m-k) \tag{25}$$

где $R_n(k)$ — автокорреляционная функция. Это упрощение происходит из-за наложения ограничений оценочного интервала в диапазоне [0, N-1] и предположения, что значения, лежащие вне диапазона, равны нулю.

Из-за того, что длина должна быть конечной, наложение окна (см. уравнение (5)) на сигнал имеет важное значение. Обычно для этого используют окно Хамминга. Приложения, использующие такое окно, высвечивают проблемы вызываемые быстрыми изменениями в сигнале на краях окна. При перекрывающем анализе гарантируется гладкий переход от кадра к кадру оцененных параметров.

Это упрощение позволяет предсказать коэффициенты для подсчета, эффективно используя рекурсию Левинсона-Дарбина [37]:

Инициализация:

$$E_{LP}^{(0)} = R_n(0) \tag{26}$$

for (i=1; i<=N_{LP}; i++)

{

$$k_{LP}(i-1) = \frac{R_n(i) + \sum_{j=1}^{i-1} a_{LP}^{(i-1)}(j)\cdot R_n(i-j)}{E_{LP}^{(i-1)}} \tag{27}$$

$$a_{LP}^{(i)}(i) = k_{LP}(i-1) \tag{28}$$

for (j=1; j<i-1; j++)

{

$$a_{LP}^{(i)}(j) = a_{LP}^{(i-1)}(j) + k_{LP}(i-1)\cdot a_{LP}^{(i-1)}(i-j) \tag{29}$$

}

$$E_{LP}^{(i)} = \left(1 - k_{LP}(i-1)^2\right)\cdot E_{LP}^{(i-1)} \tag{30}$$

}

Эти уравнения подсчитывают коэффициенты предсказания пропорционально N^2_{LP} и позволяет все вычисления линейного предсказания выполнять примерно так $N_s N_{LP} + 3N_s + N_{LP}^2$.

Модель сигнала в действительности — это инверсия $H_{LP}(z)$:

$$S_{LP}(z) = \frac{G_{LP}}{H_{LP}(z)} \tag{31}$$

где G_{LP} — модель прироста:

$$G_{LP} = \sqrt{E_{LP}^{(N_{LP})}} \tag{32}$$

Обратите внимание на то, что условие прироста можно выразить следующей формулой:

$$G_{LP} = \sqrt{R_n(0)\prod_{i=1}^{N_{LP}}\left(1 - k_{LP}(i-1)^2\right)} \tag{33}$$

Условие прироста позволяет сопоставлять спектр модели линейного предсказания со спектром оригинального речевого сигнала. Модель линейного предсказания, описываемая уравнением (19), является нормализованной, так как значения предсказанных коэффициентов не зависят от энергии сигнала.

Есть три важных наблюдения по линейному предсказанию. Первое, промежуточные переменные $\{k_{LP}\}$, используемые при вычислении, называются *коэффициентами отражения*. Их связь:

$$0 \le \left|k_{LP}(i-1)\right| \le 1, \quad \forall 1 \le i \le N_{LP} \tag{34}$$

Это очень полезный результат для приложений хранения и компрессии, использующих модель линейного предсказания. Например, коэффициенты линейного предсказания для приложений работающих с речью обеспечивают сжатие с качеством 30бит/модель без значимой деградации [58]. Коэффициенты линейного предсказания могут нормально храниться в таком виде и мы можем достичь значительного порядка компрессии оригинальных данных. Это немаловажно для приложений распознавания речи, которые хранят большое количество моделей.

Второе, итеративный подход обсчитывает все модели порядка $1 \le i \le N_{LP}$. Это очень удобно использовать в приложениях обработки сигнала, которые оценивают порядок модели как часть задачи. Обычно, в приложениях распознавания речи, порядок модели — это фиксированный системный параметр.

Третье, с повышением порядка точность модели улучшается. Уравнение (30) представляет собой энергию ошибки. Из этого уравнения видно, что ошибка монотонно уменьшается при росте порядка модели. Модель пытается совпасть с общим спектром насколько это позволяет порядок.

Данный факт показан на рис. 15, где показан речевой спектр и два спектра соответствующие моделям линейного предсказания. Обратите внимание на то, что при увеличении порядка, модель все лучше совпадает с оригинальным спектром. При малом порядке, видно только большие пики, а при большом — видно более мелкие детали спектра.

Как было рассмотрено выше, спектральная модель не точна в областях сигнала с низкими энергиями. Мы желаем установить порог динамического диапазона как было сделано для рис. 12. Есть несколько методов осуществить подобное в модели линейного предсказания:

1. Стабилизированный ковариантный метод [59], который уменьшает динамический диапазон спектра;

2. Взвешивающий метод [60], который немного расширяет границы диапазона модели линейного предсказания;

3. Стабилизированный автокорреляционный метод [44], в котором в автокорреляционную функцию добавляется немного шума.

Последний из этих методов прост и эффективен. Автокорреляционная функция, представленная уравнением (25), модифицируется до применения линейного предсказания следующим образом:

$$\begin{aligned} R_{nw}(0) &= (1+\gamma_{nw})R_n(0) \\ R_{nw}(i) &= R_n(i) \quad i>0 \end{aligned} \tag{35}$$

Порог динамического диапазона обычно указывается в dB:

$$\gamma_{nw_{dB}} = 10\log_{10}\gamma_{nw} \tag{36}$$

Типичное значение порога динамического диапазона — 10 dB.

Процесс стабилизации эквивалентен добавлению некоррелированного белого шума в речевой сигнал перед линейным предсказанием. Он нужен для того, чтобы избежать моделирования точных нулей в спектре в модели линейного предсказания. Это показано на рис. 16. Обратите внимание, что некоторое искажение в форме спектрального сглаживания вводится в модель в области высоких энергий спектра (расширение границ диапазонов резонансов модели линейного предсказания иногда вызывает соседний спектральный резонанс и сворачивается в один широкий диапазон).

Позвольте нам сделать одно важное замечание по модели линейного предсказания. Добавляя информацию об основной частоте и энергии к коэффициентам линейного предсказания, можно полностью реконструировать аудио версию речевого сигнала [56]. Изучение параметрического описания, особенно моделей распознавания речи, очень полезно для задачи диагнозирования [61]. Некоторые параметрические преобразования, такие как кэпстральные коэффициенты, не совпадают с оригинальными данными линейного предсказания. Следовательно, довольно трудно оценить достоверность набора параметров.

Исторически, системы распознавания речи сначала использовали параметры линейного предсказания напрямую. Теперь же разработаны более продвинутые преобразования этих параметров. Тем не менее, важно помнить, что создание точной модели линейного предсказания — самый важный из первых шагов в спектральном анализе. Так как метод линейного предсказания является нелинейной операцией, производительность системы в шумной обстановке иногда падает. По этим причинам, некоторые системы все еще используют банк Фурье-фильтров.

На рис. 17 показан процесс моделирования линейного предсказания, разбитый на различные условия [62]. Обратите внимание, что при уменьшении размера кадра (и соответствующего уменьшения размера окна) увеличивается временное разрешение спектрограмм. Кадр размеров в 20 мс используется практически в каждой системе распознавания речи. Позднее, когда внимание исследователей сфокусировалось на фонетическом распознавании, стал использоваться кадр размером порядка 10 мс. Продолжается улучшение временного разрешения по мере развития технологии фонетического распознавания.

3.3.5 Амплитуды банка фильтров на основе линейного предсказания

Можно объединить понятие банка фильтров, описанного в разделе 3.3.1, с моделью линейного распознавания. Амплитуды банка фильтров — это амплитуды банка фильтров, проистекающие от дискретной формы спектральной модели линейного предсказания при определении частот банка фильтров. Проницательный читатель может спросить: "В чем же преимущество данного метода?"

Было оспорено, что модель линейного предсказания, или модель высокого разрешения, как ее часто называют, дает более устойчивые спектральные оценки [57]. Спектральное сглаживание, присущее в модели линейного предсказания, дает более стабильные параметры для последующих этапов обработки. Тем не менее, распознавание речи и технологии цифровой обработки сигналов совершенствуются и различия в этих подходах не такие большие как может показаться. Как же мы сможем получить образец спектра модели линейного предсказания? Техника просчета амплитуд банка фильтров по модели линейного предсказания включает в себя прямую оценку этой модели:

$$S_{LP}(f) = \frac{G_{LP}}{\sum_{i=0}^{N_{LP}} a_{LP}(i) \cdot e^{-j2\pi \cdot (f/f_s)}} \tag{37}$$

где f_s — частота оцифровки. Этот метод требует порядка 4p+3 операций умножения/присваивания на одно значение частоты. Как описано в разделе 3.3.1, спектр обычно перегружен и средние оценки делаются для актуальных амплитуд банка фильтров.

Другой популярный подход — подсчитать спектр энергии из автокорреляции импульсного отклика $H_{LP}(z)$. Импульсный отклик $H_{LP}(z)$ можно напрямую подсчитать из коэффициентов линейного предсказания [32]:

$$R_{LP}(k) = \begin{cases} \sum_{m=0}^{N_{LP}-|k|} a_{LP}(m) \cdot a_{LP}(m+|k|), & |k| \le N_{LP} \\ 0, & |k| \ge N_{LP} \end{cases} \tag{38}$$

Плотность энергетического спектра может быть успешно подсчитана через автокорреляционную функцию при учете того, что автокорреляционная функция является четной реальной функцией. Следовательно, ее Фурье-преобразование тоже реальное:

$$S_{LP}(f) = R_{LP}(0) + 2\sum_{k=1}^{N_{LP}} R_{LP}(k) \cdot \cos\left(2\pi k \frac{f}{f_s}\right) \tag{39}$$

Уравнение (38) требует всего приблизительно $N_{LP}^2 - \frac{3}{2} \cdot N_{LP}$ операций умножения/присвоения, в то время как уравнение (39) — всего N_{LP}.

Похожим образом можно обработать нелинейные спектры с помощью соответствующего выбора набора частот банка фильтров. Хотя модель линейного предсказания работает со сглаженными спектрами, ее также выгодно использовать при экстремальных спектрах, такими что острые пики в частотном отклике будут точно характеризованы банком фильтров, который имеет тенденцию грубо квантовать спектр.

3.3.6 Линейно предсказанные кэпстральные коэффициенты.

Наконец, мы обсуждаем нашу последнюю технику измерения сигналов. Вспомним, что в последней главе мы применили модель линейного предсказания для просчета амплитуд банка фильтров. Другим логическим шагом в этом направлении будет использование модели линейного предсказания для подсчета кэпстральных коэффициентов. Снова, проницательный читатель может спросить: "Можно ли подсчитать кэпстральные коэффициенты напрямую из модели линейного предсказания?"

Если линейный фильтр предсказания стабилен и гарантируется его стабильность в автокорреляционном анализе, то логарифм инверсного фильтра может быть выражен как энергетические серии в z^{-1} [63]:

$$C_{LP}(z) = \sum_{i=0}^{N_c} c_{LP}(i) \cdot z^{-i} = \log H(z) = \log \frac{G_{LP}}{\sum_{j=0}^{N_{LP}} a_{LP}(j) \cdot z^{-j}} \qquad (40)$$

Мы можем найти коэффициенты, дифференцируя обе стороны выражения и назначая коэффициентам результаты полиномов. Это делается с помощью следующей рекурсии [32,56,64]:

Инициализация:

$$c_{LP}(1) = -a_{LP}(1) \qquad (41)$$

for (i=2; i<=N_c; i++)

{

$$c_{LP}(i) = -a_{LP}(i) - \sum_{j=1}^{i-1} \left(1 - \frac{j}{i}\right) a_{LP}(j) \cdot c_{LP}(i-j) \qquad (42)$$

}

Коэффициенты $\{c_{LP}\}$ — линейно предсказанные кэпстральные коэффициенты.

Исторически, коэффициент $c_{LP}(0)$ был определен как логарифм энергии ошибки линейного предсказания [31]. Теперь же, отметим, что так как энергия рассматривается как отдельный параметр, то нет нужды включать его в вышеприведенное уравнение. Можно считать кэпстральную модель нормализованной, похожей на модель линейного предсказания, в которой $c_{LP}(0) = \log 1 = 0$. Мы обсудим этот момент позже (раздел 5.2), когда будем сравнивать модели.

Есть только одна маленькая сложность в рекурсии кэпстральных коэффициентов. Мы не указываем число кэпстральных коэффициентов (N_c) для подсчета. Так как в действительности они являются результатом инверсного Фурье-преобразования импульсного отклика модели линейного предсказания, а модель линейного предсказания сигнала является фильтром бесконечного импульсного отклика, мы можем, теоретически, подсчитать бесконечное число кэпстральных коэффициентов. Тем не менее, число подсчитанных кэпстральных коэффициентов обычно сравнимо с числом коэффициентов линейного предсказания: $0{,}75p \le N_c \le 1{,}25p$.

Подсчитанные с помощью вышеописанной рекурсии кэпстральные коэффициенты отражают шкалу линейных частот. Единственный недостаток этого метода в том, что мы должны работать немножко больше для введения понятия нелинейной шкалы частот. Предпочтительный подход базируется на методе, используемом для сдвига частот при проектировании цифрового фильтра. В этом методе используется

очень важное преобразование в цифровой обработке сигналов: билинейное преобразование [27]:

$$f_{new} = 2\pi\frac{f}{f_s} + 2\tan^{-1}\left(\frac{\alpha_{bt}\sin\left(2\pi\frac{f}{f_s}\right)}{1-\alpha_{bt}\cos\left(2\pi\frac{f}{f_s}\right)}\right) \qquad (43)$$

где α_{bt} — параметр сдвига частоты. Когда $0{,}4 \le \alpha_{bt} \le 0{,}8$, частотный сдвиг билинейного преобразования подобен Mel-шкале, это показано на рис. 18. Обычно, $\alpha_{bt} = 0{,}6$.

Мы можем использовать это преобразование в одном из нескольких мест подсчета линейно предсказанных кэпстральных коэффициентов: в автокорреляционной функции, в параметрах предсказания или в линейно предсказанных кэпстральных параметрах. В [20] показано, что последующая обработка кэпстральных коэффициентов — это наиболее эффективный и самый простой метод.

Уравнение (43) — это процедура области частот. Если обеспечено использование подхода z-преобразования, то вычисление потребует инверсное Фурье-преобразование кэпстральных коэффициентов. Вместо этого мы предпочитаем использовать прямое рекурсивное вычисление, используя кэпстральные коэффициенты. Такая рекурсия успешно существует [65].

Эта рекурсия может быть рассмотрена как последовательность каскадных линейных операций применения фильтра безразличных к сдвигам:

for (n=0; n<=N_c; n++)
{

$$c_{bt}^{(n)}(0) = \alpha_{bt}\left[c_{bt}^{(n-1)}(0) - 0\right] + c_{LP}(N_c - n) \qquad (44)$$

$$c_{bt}^{(n)}(1) = \alpha_{bt}\left[c_{bt}^{(n-1)}(0) - 0\right] + \left(1-a^2\right)\cdot c_{bt}^{(n-1)}(0) \qquad (45)$$

for (k=2; k<=N_{bt}; k++)
{

$$c_{bt}^{(n)}(k) = \alpha_{bt}\left[c_{bt}^{(n-1)}(k) - c_{bt}^{(n)}(k-1)\right] + c_{bt}^{(n-1)}(k-1) \qquad (46)$$

}
}

где все начальные условия равны нулю. Так как $c_{LP}(0) = 0$, обработка может начинаться с $c_{LP}(1)$ при n=0.

В этой рекурсии сначала мы делаем итерацию по всем $c_{bt}(k)$ и обновляем эти значения для каждого n (вторая итерация). Результатом N_c итераций будут коэффициенты преобразования. Эта рекурсия требует порядка $N_c \times N_{bt}$ операций. Это похоже по сложности на линейное предсказание.

Снова нам необходимо принимать во внимание усечение. Кэпстральная билинейно преобразованная последовательность, которая является результатом усеченного кэпстра обработанного нелинейным частотным преобразованием, практически бесконечна по длительности. Тем не менее, на практике, если кэпстральная последовательность имеет конечную длину, результирующая преобразованная последовательность будет асимптотической экспоненциальной задержкой. Следовательно, есть возможность ограничить преобразованную последовательность маленьким искажением. Обычно, $N_{bt} \le 1{,}25N_c$.

На рис. 19 мы демонстрируем объединение эффектов кэпстрального анализа и билинейно преобразованных коэффициентов. Показан речевой спектр для сигнала, оцифрованного с частотой 8 кГц, вместе с его моделью линейного предсказания 10-го порядка. Логарифм величины спектра кэпстральных коэффициентов также показан для кэпстрального анализа 12-го порядка. Также показан логарифм величины спектра для билинейно преобразованного кэпстра. На этом примере 12 кэпстральных коэффициентов были преобразованы в 16 билинейно преобразованных коэффициентов.

Собственная сумма компрессии для билинейного преобразования является до некоторой степени функцией частоты оцифровки. В некоторых примерах работающих с частотой оцифровки 16 кГц [20, 66] используется фактор компрессии равный 0,6. Эффективный диапазон речевого сигнала при 16 кГц оцифровке мал (максимум энергии приходится на нижнюю четверть шкалы частот для сонорантны).

Мы видим, что по ряду причин билинейное преобразование не так полезна при частоте оцифровки в 8 кГц. Во-первых, речевой сигнал теперь занимает почти весь доступный диапазон. Есть минимум "свободного пространства" (на диапазоне 3..4 кГц), чтобы использовать растяжение спектра. Можно выделить полезную информацию, если присутствуют экстремальные суммы масштабирования.

Во-вторых, фактор компрессии должен быть минимален, чтобы поддержать приближение Mel-шкалы. Также, при уменьшении фактора уменьшается сумма сжатия. Отметим, что это преобразование не будет лучшим способом приближения Mel-шкалы.

Мы обсудили все основные методы измерения сигнала, которые используются в настоящее время в системах распознавания речи. Далее, мы обсудим как эти параметры будут сглаживаться и конкатенировать для формирования параметров сигнала.

IV. Преобразования параметров

В предыдущем разделе, мы обсудили несколько методов обсчета абсолютных измерений. В этом разделе мы обсудим следующее звено цепочки операций показанных на рис. 1 — преобразование параметров. Параметры сигнала генерируются из измерений сигнала с помощью двух основных операций — дифференцирования и конкатенирования. Выходной результат обработки является вектором параметров, который содержит ряд оценок сигнала. Обзор этих операций, которые составляют вектор параметров показан на рис. 20.

4.1 Дифференцирование

С увеличением вычислительной мощности компьютеров в начале 1980-х, использование вспомогательных измерений речевого спектра в системах динамического искажения времени стало осуществимо. Как часть продолжающегося движения к лучшему выделению временных изменений в сигнале, в модель сигнала была добавлена производная высокого порядка измерений сигнала [14, 18, 21]. Прежде обсужденные абсолютные измерения можно представить как производные нулевого порядка. Теперь, мы исследуем добавление производных этих измерений для нашей модели сигнала.

В цифровой обработке сигнала существует несколько способов в котором производная первого порядка может быть аппроксимирована. Три популярных приближения показаны ниже [27, 67, 68]:

$$s(n) \equiv \frac{d}{dt} s(n) \approx s(n) - s(n-1) \tag{47}$$

$$s(n) \equiv \frac{d}{dt} s(n) \approx s(n+1) - s(n) \tag{48}$$

$$s(n) \equiv \frac{d}{dt} s(n) \approx \sum_{m=-N_d}^{N_d} m \cdot s(n+m) \tag{49}$$

Обратите внимание, что мы отбросили излишний фактор нормализации в этих уравнениях. Первые два уравнения известны как обратное и прямое дифференцирование, а уравнение (49) представляет приближение линейного фильтра фазы как идеального дифференциатора. Этот подход часто называют регрессионным анализом.

Выходной сигнал из этого процесса дифференцирования известен как *дельта-параметр*. Производные второго порядка могут быть также аппроксимированы, применяя уравнение (49) к выходному сигналу дифференциатора первого порядка, как это показано на рис. 20. Такой выходной сигнал называют *дельта-дельта параметром*. Ясно, что этот процесс можно продолжать и дальше.

Уравнение (47) похоже на уравнение (2). Вспоминая, что основной целью взвешивающего фильтра является усиление высокочастотных порций спектра, мы должны знать, что дифференцирование является "шумным" процессом. Дифференцирующие фильтры имеют тенденцию усиливать шум в измерениях сигнала. Часто бывает нежелательно подсчитывать производные сглаженных параметров, а не чистые измерения, таким образом, уровень шума в измерениях на выходе уменьшается.

Существует несколько популярных способов достичь такого результата. Регрессионный анализ, уравнение (49), сплайн интерполяция и дифференцирование на

ограниченном диапазоне — одни из таких способов. Так как уравнение (49) подсчитывает разницу симметрично расположенных вокруг отсчета во время n, используется набор из N_d предыдущих отсчетов для подсчета текущего значения. Следовательно, некоторое измерение сглаживания присуще таким подсчетам.

Есть два способа использования уравнения (49). Много систем в настоящее время [2, 66, 113] использует простую разницу первого порядка: N_d=1. Эти системы обычно работают с кадрами в диапазоне 10..20 мс. Диапазон времени на котором берется производная довольно мал: $\Delta T_f = 40$ мс. Вторая группа систем [11, 67] используют $5 \le N_d \le 7$, а таких системах кадр равен 8..10 мс. Диапазон на котором подсчитывается производная немного больше: $\Delta T_d = 56..75$ мс.

Частотные отклики некоторых реализаций дифференциального фильтра показаны на рис. 21. Низкочастотная часть каждого фильтра создана для аппроксимации линейной функции (пилообразная функция), которая несет полезную высокочастотную информацию (характеристика временных изменений). Отметим, что когда порядок дифференциатора увеличивается, фильтр начинает взвешивать высокие частоты и привносит большие пульсации в спектр. Сами затухающие высокие частоты рассматриваются как форма уменьшения шума — после определенной точки высокочастотная информация считается ненадежной и не используется.

4.2 Конкатенация

В большинстве методов измерения наиболее общим способом, который мы обсудили, является использование линейной фильтрации для достижения сглаживания параметров. Большинство систем обрабатывают данные таким образом, что операции можно легко объяснить в терминах теории линейной фильтрации. В этом разделе мы опишем это понятие в форме матричного оператора.

Зададим матрицу измерений сигнала следующим образом:

$$\underline{X} = \begin{bmatrix} x(0,0) & x(0,1) & \dots & x(0, N_x - 1) \\ x(1,0) & x(1,1) & \dots & x(1, N_x - 1) \\ \dots & \dots & \dots & \dots \\ x(N_f - 1,0) & x(N_f - 1,1) & \dots & x(N_f - 1, N_x - 1) \end{bmatrix} \qquad (50)$$

где x(n,m) — m-ое измерение сигнала в кадре n (или время $\left(n + \frac{1}{2}\right)T_f$); N_f — общее число кадров в сигнале; N_x — общее число измерений сигнала в кадре.

Матрица измерений сигнала ($\underline{X}$) содержит все измерения сигнала за все время. Во многих практических системах сигнал обрабатывается кадр за кадром в реальном времени. Накопление сигнала в одной большой матрице добавляет задержку в систему. Тем не менее, в исследовательских целях удобно рассматривать модель сигнала как матрицу измерений.

Обратите внимание, что матрица измерений сигнала обычно содержит смесь измерений: энергию и набор кэпстральных коэффициентов. N_x — это размерность вектора, который объединяет эти измерения. С этой точки зрения мы рассматриваем эти измерение как группу, а не индивидуально, и не указываем особые виды измерений. Тем не менее, в некоторых видах анализа полезно объединять измерения вместе в соседние столбцы матрицы $\underline{X}$ (для облегчения субматричных операций).

Зададим две вспомогательные матрицы для параметризации процесса сглаживания. Во-первых, зададим матрицу запаздывания или задержки (τ), которая будет представлять временную задержку:

$$\underline{\tau} = \begin{bmatrix} \tau(0,0) & \tau(0,1) & \dots & \tau(0,\tau_0-1) \\ \tau(1,0) & \tau(1,1) & \dots & \tau(1,\tau_1-1) \\ \dots & \dots & \dots & \dots \\ \tau(N_p-1,0) & \tau(N_p-1,1) & \dots & \tau(N_p-1,\tau_p-1) \end{bmatrix} = \begin{bmatrix} \tau_0 & \tau_1 & \dots & \tau_{N_p-1} \end{bmatrix} \quad (51)$$

где τ_i соответствует i-му вектору запаздывания, N_p — общее число параметров сигнала. Вектор задержек может быть разной размерности, в зависимости от того сколько измерений было использовано при вычислении каждого определенного параметра сигнала, матрица задержки в действительности является вектором векторов. Следовательно, мы задаем измерение каждого ряда в терминах N_{τ_i}.

Далее, зададим взвешивающую матрицу ($\underline{W}$), которая содержит веса фильтров применяемых к измерениям. Эти веса напрямую соответствуют значениям в матрице задержки. Взвешивающая матрица задается так:

$$\underline{W} = \begin{bmatrix} w(0,0) & \dots & w(0,N_{\tau_0}-1) \\ \dots & \dots & \dots \\ w(N_p-1,0) & \dots & w(N_p-1,N_{\tau_p}-1) \end{bmatrix} = \begin{bmatrix} \overline{w}_0 & \overline{w}_1 & \dots & \overline{w}_{N_p-1} \end{bmatrix} \quad (52)$$

где $\overline{w}_i$ — вектор i-го коэффициента, чья размерность равна соответствующему вектору в $\underline{\tau}$.

Также определим индексирующий вектор ($\bar{I}$), который для каждой строки в $\underline{W}$ задает соответствующий столбец в $\underline{X}$:

$$\bar{I} = \begin{bmatrix} I_0 & I_1 & \dots & I_{N_s-1} \end{bmatrix} \quad (53)$$

Определим процесс фильтрации вектора измерений как оператор псевдосвертки:

$$\underline{V} = \underline{X} \otimes \underline{W},$$

где оператор $\otimes$ определен так:

for (n=0; n<=N_f-1; n++)
{
  for (i=0; i<=N_x-1l i++)
  {

$$V[n,i] = \sum_{j=0}^{N_{\tau_j}-1} W(i,j) \cdot X[n+\tau(i,j), I_i] \quad (54)$$

  }
}

Уравнение (54) представляет последовательность операций линейной фильтрации для каждого элемента вектора параметров сигнала для каждого кадра сигнала. Это выражено с помощью гибкой схемы индексирования, которая принимает во внимание, что различные типы характеристик требуют применения различных фильтров

Обратите внимание, что с использованием массива индексов ($\bar{I}$) можно получить многочисленные параметры из одного и того же измерения, т.е. среднюю энергию и дельта-энергию из одного измерения. Также матрица коэффициентов $\underline{W}$ может быть использована для реализации всех операций фильтрации, которые мы обсудили прежде, включая дифференцирование, усреднение и взвешивание. Уравнение (54) будем называть *конкатенацией* — создание единственного на кадр вектора параметров, который будет содержать все желаемые параметры сигнала.

Некоторые параметры, такие как энергия, часто нормализуются перед вычислением уравнения (54). Проще разделить энергию на максимум наблюдаемого значения или вычесть логарифм энергии. Этот подход имеет недостаток — необходимо ждать пока вся последовательность будет идентифицирована или обработана. Такая задержка часто неприемлема для систем реального времени.

Максимальные величины энергии можно также подсчитать с помощью рекурсивного метода фильтрации, который описан в разделе 3.2. В этом случае, покадровые оценки энергии часто фильтруются схемой адаптивного управления уровнем. Данная схема также привносит некоторую задержку, так как алгоритму требуется время для реакции на изменения сигнала. Время адаптации для таких оценочных функций порядка 0,25 с [66].

В то время, когда системы распознавания речи были очень просты, модели сигнала часто состояли из сильно сглаженных параметров. "Шумовые параметры" — это параметры усиливают динамику в спектре, считается, что они ненадежны. С возникновением Марковской модели, которая привнесла математический базис для характеризации последовательных (или временных) аспектов сигнала, уверенность в использовании динамических характеристик выросла. В настоящее время, динамические характеристики считаются существенными при разработке хорошего фонетического распознавания [69].

Дифференциальные параметры стали популярнее, когда ученые нашли модель сигнала инвариантную к резким изменениям в поведении говорящего (иногда эти изменения индуцируются приложением) [2, 70]. Например, в приложениях навигации звуковой сигнал может значительно изменяться если говорящий испытывает физическое или умственное напряжение. Эти изменения часто обнаруживаются в абсолютных спектрах, а не в дифференциальных. Считается, что параметры, полученные из дифференциальной спектральной информации, делают систему более независимой от роста изменений сигнале.

Прежде чем завершить этот раздел, давайте обсудим одну определенную форму взвешивания параметров работающую с кэпстральными коэффициентами. Ранние исследования методов кэпстральной обработки предлагали средства для выполнения операций линейной фильтрации непосредственно с кэпстральными коэффициентами, чтобы расширить эти части кэпстра, которые представляют информацию о голосовом тракте [71]. Этот метод известен как *лифтеринг*, термин образовался из-за сходства с теорией линейной фильтрации.

Процесс лифтеринга прост и выглядит так:

$$c_{\text{Lift}}(m) = c(m) \cdot w_{\text{Lift}}(m), \tag{55}$$

где

$$w_{\text{Lift}}(m) = 1 + \frac{N_c}{2} \sin\frac{\pi m}{N_c} \tag{56}$$

Уравнение (55) описывает оконную операцию, временная шкала кэпстра называется Que-частотой. Уравнение (56) описывает взвешивающую (или оконную) функцию. Здесь, мы просто отметим, что она является статической взвешивающей функцией, которую можно напрямую применить к любому набору кэпстральных коэффициентов. В следующем разделе, мы обсудим метод подсчета таких взвешиваний в статистической манере.

V. Статистическое моделирование

В последнем разделе мы обратили внимание на задачу статистического моделирования параметров сигнала. В этом разделе мы считаем, что параметры сигнала были генерированы некоторым многомерным случайным процессом. Постараемся изучить природу такого процесса. Создадим модель для данным, оптимизируем (или обучим) модель, а затем измерим качество приближения. Мы будем располагать только той информацией, которая появится на выходе модели. По этой причине, выходной вектор параметров называется *наблюдением сигнала*. Набор таких векторов для всего сигнала называется матрицей наблюдений сигнала.

Этот последний шаг обработки является первым шагом статистического моделирования в распознавании речи. Методы, описанные ниже, являются лишь основными подходами. Системы распознавания речи используют чрезвычайно интеллектуальные статистические модели — это одна из основных функций системы. Тем не менее, методы, представленные ниже, будут полезны для применения в широком спектре приложений обработки сигнала и формируют основу для более продвинутых алгоритмов. Обзор различных типов преобразований, обсуждаемых ниже, показан на рис. 22.

5.1 Многомерные статистические модели

Как мы упомянули выше, в типичном наборе разнородных параметров сигнала мы объединяем величины, такие как энергия и кэпстральные коэффициенты, которые имеют разные размерности: диапазон и вариация энергии гораздо больше чем диапазон и вариация кэпстральных коэффициентов. Изменения производных времени энергии гораздо больше чем изменения производных для кэпстральных коэффициентов. Если мы сравним два вектора параметров, используя такой простой оператор как Евклидово расстояние, то в результате будут доминировать члены с большими амплитудами и вариациями, хотя верная информация может лежать в параметрах с малыми амплитудами.

Подобно этому, если мы рассмотрим измерение с помощью амплитуд банка фильтров, то легко понять, что амплитуды соседних ячеек будут согласованы друг с другом. В действительности, банк фильтров специально разработан так, чтобы получать такой тип корреляции.

Мы можем проиллюстрировать эту задачу с выполнением прямых сравнений в пространстве векторов, в котором размерности имеют неравные вариации по сравнению с простым двухмерным примером показанном на рис. 23(a). В оригинальной системе координат расстояние между точками **a** и **b** равно расстоянию между точками **c** и **d** (размерности одинаковы). Также, мы будем считать, что первое расстояние будет больше второго, так как у него больше процент вариации параметра. Примечание — аргумент показанный на рис. 23 предполагает, что наблюдаемая вариация не привносит большого шума и что она, в действительности, представляет собой значимую вариацию параметра.

До тех пор пока решение вышеприведенной задачи вариации взвешивания прямое, устранение корреляции будет очень тонким процессом. Нам надо устранить корреляцию из наших измерений по двум причинам.

Во-первых, корреляция привносит избыточность. Действительное число "верных" параметров необходимых для описания информации может быть гораздо меньше числа измерений. Мы можем достичь некоторого уровня компрессии с по-

мощью выбора поднабора характеристик или линейной комбинации характеристик. Посторонние измерения в задаче обработки сигнала часто являются источником проблем.

Во-вторых, мы используем простой метод сравнения векторов. Присутствие коррелированных параметров делает разработку оптимального статистического показателя более сложной.

5.1.1 Предварительные преобразования

Существует прямой метод декорреляции параметров в статистически оптимальном виде для многомерной гауссовой обработки. Зададим многомерное гауссово распределение вероятности:

$$\iota(\bar{\nu}) = \aleph[\bar{\nu},\ \bar{u}_\nu,\ \underline{C}_\nu] = \frac{1}{\sqrt{(2\pi)^{N_\nu}|C_\nu|}} \cdot \exp\left[-\frac{1}{2}(\bar{x}-\bar{u}_\nu)\cdot\underline{C}_\nu^{-1}(\bar{x}-\bar{u}_\nu)^+\right] \quad (57)$$

Предположим, что наши параметры принадлежат этому типу статистической модели или, другими словами, наши параметры можно смоделировать с достаточной точностью таким процессом.

Мы можем просчитать линейное преобразование, которое будет одновременно нормализовывать и декоррелировать параметры. Зададим вектор преобразования:

$$\bar{y} = \underline{\Psi}(\bar{\nu} - \bar{\mu}_\nu) \quad (58)$$

где $\bar{\nu}$ — входной вектор параметров, $\bar{\mu}_\nu$ — средняя величина входного вектора параметров, $\underline{\Psi}$ — *предварительное преобразование* [18, 72], основанное на нашем желании того, чтобы выходной вектор был некоррелированным случайным гауссовым вектором. Для того, чтобы получить такой результат, надо выполнить:

$$\underline{\Psi} = \underline{\Lambda}^{-\frac{1}{2}}\underline{\Phi}^+ \quad (59)$$

где $\underline{\Lambda}$ — диагональная матрица собственных значений, $\underline{\Phi}$ — матрица собственных векторов ковариантной матрицы $\bar{\nu}$.

Полное рассмотрение уравнения (59), истинного значения собственных векторов и других тонкостей остались за границами нашего внимания. Тем не менее, мы не можем преуменьшить важность необходимости статистической нормализации параметров [73, 74]. Собственные значения и собственные векторы описанные выше являются ключом ко всему вычислению, так как они описывают линейное преобразование пространства входящих векторов в новое пространство с подсчитанными Евклидовыми дистанциями.

Собственные значения и собственные вектора соответствуют следующему выражению:

$$\underline{C}_\nu = \underline{\Phi}\underline{\Lambda}\underline{\Phi}^+ \quad (60)$$

где $\underline{C}_\nu$ — ковариантная матрица для $\underline{\nu}$, каждый элемент которой находится с помощью следующей формулы:

$$C_\nu(i,j) = \frac{1}{N_f}\sum_{m=0}^{N_f-1}(\nu_m(i)-\mu_\nu(i))\cdot(\nu_m(j)-\mu_\nu(j)) \quad (61)$$

Мы немного задержали описание вычисления собственных значений и собственных векторов, так как это довольно сложная алгоритмическая операция. До тех

пор, пока эта процедура будет иметь простую интерпретацию в линейной алгебре, программировать это не стоит. Обратите внимание, что ковариантная матрица является реальной, симметричной матрицей и что ковариантные матрицы для речевых параметров хорошо изучены и их решение давно известно.

Есть одно очень важное упрощение уравнения (60), которое надо обсудить. Если параметры некоррелированы, то ковариантная матрица в уравнении (60) сворачивается в диагональную матрицу:

$$\underline{\Psi} = \begin{bmatrix} \frac{1}{\sigma_{\nu(0)}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_{\nu(1)}} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \frac{1}{\sigma_{\nu(N_\nu-1)}} \end{bmatrix} \tag{62}$$

где $\sigma_{\nu(i)}$ — стандартное отклонение i-го компонента вектора параметров $\bar{\nu}$. Легко распознать нормализацию параметров по их стандартным отклонениям, этим самым получение каждого элемента параметра одинаково при вычислении (см. рис. 23).

Многократно наблюдалось, что определенный набор параметров, называемый кэпстральными коэффициентами, может быть рассмотрен (в приближении) как некоррелированный [66, 74]. Это удобно, так как значительно уменьшает количество параметров, которые необходимы для оценки системы. Так называемые "вариативно-взвешенные кэпстральные коэффициенты" очень популярны в настоящее время в системах распознавания речи. Другие параметры, такие как амплитуды банка фильтров, показывают очень сильную корреляцию первых недиагональных компонентов. Иногда это выгодно при желании уменьшить вычислительный шум в системе и уменьшить ее сложность, при аппроксимации ковариантной матрицы, например, ленточная матрица [18].

Техника уменьшения шума — это процедура в которой пренебрегают наименее значащими характеристиками. Если мы зададим N_ν собственных значений λ_0, λ_1, … $\lambda_{N_\nu-1}$, отсортированные по уменьшению порядка, получим важное отношение:

$$\sum_{i=0}^{N_\nu-1} \lambda_i = \text{trace} \quad C_\nu = \sum_{i=0}^{N_\nu-1} \sigma_i^2 \tag{63}$$

Из этого отношения видно, что собственные значения и вариации процесса относительны.

Можно задать сумму вариаций для каждой пары собственное значение – собственный вектор:

$$\zeta_i = \frac{\lambda_i}{\sum_{i=0}^{N_\nu-1} \lambda_i} \times 100\% \tag{64}$$

и общий процент вариации для первых N_y размерностей (одна размерность соответствует одно паре с.вектор – с.значение):

$$\zeta_{N_y} = \frac{\sum_{j=0}^{N_j-1} \lambda_j}{\sum_{i=0}^{N_v-1} \lambda_i} \times 100\% \qquad (65)$$

В этом случае, матрица преобразования $\underline{\Psi}$ является прямоугольной $N_y \times N_v$, но не квадратной, матрицей. Уравнения (64) и (65) часто используются для принятия решения о том, сколько же размерностей принимать во внимание.

Если одномерное гауссово распределение — недостаточная модель, то не забывайте, что мы можем моделировать данные как взвешенную сумму гауссова распределения [74]. Мы не обсуждаем этот аспект, но известно, что можно получить асимптотически хорошее совпадение с распределением параметров с такой линейной комбинацией функций гауссовой плотности.

Процесс предварительного преобразования помогает понять рис. 24, на котором показаны три собственных вектора преобразования для случая речи телефонного качества. Были подсчитаны амплитуды банка фильтров из банка фильтров Mel-шкалы. Обратите внимание, что каждый собственный вектор пытается смоделировать различные аспекты спектра речи. Несколько первых собственных векторов матрицы преобразования имеют тенденцию моделировать большие спектральные характеристики канала, которые постоянно изменяются в телекоммуникационных приложениях. Заметьте, что эта взвешивающая функция, при применении амплитуд банка фильтров, может быть рассмотрена как операция фильтрации во временной области (заключение области частот в окно эквивалентно фильтрации временной области).

До одностороннего введения энергии в многомерную гауссову модель, необходимо обучить матрицу преобразования. Обычно для этого собирают данные и ковариантную статистику из большого количества речевых данных. Это своего рода искусство, так как мы должны быть уверены, что наша статистика точно совпадает со статистикой реального процесса, которые мы моделируем. Недостаточный набор данных при обучении (слишком мало данных или условия записи, которые не совпадают с условиями эксплуатации) может привести к неточным преобразованиям.

5.1.2 Векторное квантование

Вспомним тот факт, что в гауссовом процессе во все моменты, кроме первого и второго моментов, значение и вариация равны нулю. Для речевых параметров, моменты высокого порядка часто являются значимыми[7], обеспечивая подтверждение того, что гауссова модель источника не всегда может быть достаточной. Мы можем выполнить непараметрическое описание данных просто предположив дискретное распределение вероятности произвольной формы и заставив систему запомнить эту форму распределения.

Такой тип модели дискретного распределения называется *векторным квантованием* [22]. Одним из самых убеждающих аргументов для использования этого ме-

[7] Моменты высокого порядка также имеют значительное влияние на речевую обработку, хотя ясно, что из этих вычислений получаются ненулевые моменты. Использование спектральных оценок высокого порядка и статистические оценки к распознаванию речи все еще является областью активных исследований. Мы предполагаем, что моменты высокого порядка могут содержать информацию о спектральном поведении сигнала. Эта информация может быть полезной для таких приложений как идентификация говорящего.

тода является то, что он базируется на модели речевоспроизведения [77], который форму речевого тракта ключом измерений. С этой точки зрения, существует малый набор физически возможных форм речевого тракта (или элементарных звуков) в языке. Следовательно, мы должны иметь возможность моделировать наш непрерывно изменяющийся вектор с конечным набором векторов, которые представляют эти уникальные формы речевого тракта. Если можно использовать измерения, которые нормализованы в соответствии с индивидуальной физиологией говорящего (длина речевого тракта, громкость и т.д.), то можно моделировать все формы вокального тракта для всех говорящих в пределах малого словаря.

Другой аргумент базируется на теории искажения частоты [22], т.е. мы должны иметь возможность смоделировать наш вектор параметров, используя конечное число дискретных значений с бесконечно малой ошибкой. Главный возникающи2й вопрос таков: "Сколько потребуется символов?" Как мы увидим, число векторов при использовании такой модели в распознавании речи обычно лежит в диапазоне 32..1024, в зависимости от поставленной задачи.

Зададим векторное квантование как смесь величин: матрицы $N_{vq} \times N_y$, *словаря векторного квантования* $\underline{Q}$ и дискретной символьной вероятности распределения $p(\bar{q}_i)$. Элементы последней $(0 \le i \le N_{vq} - 1)$ являются вероятностями наблюдения данного вектора параметров (или ряда) в $\underline{Q}$. Обозначим процесс векторного квантования входного вектора $\bar{y}$ как $Q[\bar{y}]$.

Также необходимо определить меру расстояния (или меру сходства): средство определения расстояния между двумя векторами. Более подробно опишем этот процесс в следующем разделе, а сейчас просто зададим меру сходства так:

$$D(\bar{y}_1, \bar{y}_2) = f(\bar{y}_1, \bar{y}_2) \tag{66}$$

Мы увидим, что евклидово расстояние — это одна из общих функций применяемых в уравнении (66).

Процесс векторного квантования содержит две главные проблемы. Первая, проблема обучения — как сделать оценку Q[ ] минимальной, такую как искажение из-за введения замены входного вектора на вектора из словаря? Вторая, существует проблема квантования — как оценить вероятность нахождения соответствия в словаре для $\bar{y}$? Последняя проблема является по существу задачей распознавания шаблонов, мы стремимся увеличить $P(\bar{y} \mid Q)$.

Последняя задача относительно проста: надо выбрать индекс i в соответствии с правилом ближайшего соседа:

$$i = \arg\min\left[D(\bar{y}, \bar{q}_j)\right], \quad 0 \le j \le N_{vq} \tag{67}$$

и

$$P(\bar{y} \mid Q) = p(\bar{q}_i) \tag{68}$$

$P(\bar{y} \mid Q)$ можно оценить с помощью подсчета вероятности выбора для каждого вектора в словаре (для этого надо использовать большую базу данных для обучения).

Первая проблема немного сложнее. Необходима обучающая последовательность — обычно используется одна и та же база речевых данных. Никакого близкого решения не существует для просчета оптимального набора векторов в словаре. К счастью, есть несколько итеративных методов для поиска в словаре.

Наиболее популярный из них — это K-MEANS алгоритм [76]. Его имя намекает на тот факт, что этот алгоритм старается собрать данные в K-группы и заменить данные группы значением группы. Этот процесс показан на рис. 25. Главными параметрами алгоритма является число векторов в словаре (N_{vq}) и некоторое завершающее условие. Мы используем ограничение на число итераций (M_{max}) и порог изменения для среднего искажения ($\Delta\varepsilon_{vq}$).

Алгоритм:

Инициализация

Определим набор из N_{vq} начальных векторов как $\bar{q}_i^{(0)}$

$\varepsilon^{(0)} = 1$

for (m=1; m<=M_{max}; m++)

{

$\eta(j) = 0, \quad 0 \le j \le N_{vq}$

for (n=0; n<N_f; n++)

{

$i = \arg\min\left[D\left(\bar{y}_n, \bar{q}_j^{(m-1)}\right)\right], \quad 0 \le j \le N_{vq}$

$\bar{q}^{(m)}[j] = \bar{q}^{(m)}[j] + \bar{y}_n[j], \quad 0 \le j \le N_y$

$\eta(i) = \eta(i) + 1$

}

$\bar{q}_i^{(m)} = \frac{\bar{q}_i^{(m)}}{\eta[i]}, \quad 0 \le i \le N_{vq}$

$\varepsilon^{(m)} = \varepsilon^{(m)} + Q[\bar{y}_n] \quad 0 \le n \le N_f$

if $\frac{\varepsilon^{(m)}}{\varepsilon^{(m-1)}} < \Delta\varepsilon_{vq}$ **break**

}

Завершение

$p(\bar{q}_i) = \frac{\eta[i]}{N_{db}}, \quad 0 \le i \le N_{vq}$

В итоге $Q^{(m)}$ будет содержать словарь.

Инициализация является важной частью этого процесса. Начальные условия для кластерных центров должны охватывать все пространство данных. Простая итеративная процедура [79] для выбора таких центров является поиском N_{vq} начальных векторов в данных, которые находятся друг от друга на расстоянии ε. Изначально, ε имеет некоторое большое значение, затем медленно уменьшается до тех пор, пока не будет найдено N_{vq} векторов на требуемом расстоянии.

Так как K-MEANS алгоритм является итеративным и так как мы должны установить условия для начального положения кластерных центров, этот алгоритм не гарантирует приближения к оптимальному решению. Итерационная процедура может быть перехвачена в локальном максимуме и выдать не совсем оптимальное решение. Тем не менее, на практике задача сходимости решается довольно быстро. Даже для больших словарей, сходимость часто достигалась через десять итераций.

Алгоритм, представленный здесь, подсчитывает центры новых кластеров как арифметическое среднее элементов в кластере. Он эффективен в смысле вычислительных затрат и использования памяти. Другие предложенные стратегии пересчета

[78] базируются на критерии минимума-максимума. Эти алгоритмы на практике обычно сравнимы по производительности с алгоритмом приведенном выше.

Качество словаря может быть подсчитано с помощью среднего искажения всей обучающей базы данных:

$$\varepsilon_{avg} = \frac{1}{N_f} \sum_{n=0}^{N_f-1} D(\bar{y}_n, Q[\bar{y}_n]) \qquad (69)$$

Это значение подсчитывается с каждым шагом в K-MEANS алгоритме. Среднее искажение обычно логарифмически уменьшается в зависимости от размера словаря (см. рис. 26).

После подсчета словаря процесс квантования становится задачей поиска. Правило ближайшего соседа является линейным поиском, требующим N_{vq} сравнений расстояний на входящий вектор. Есть возможность уменьшить время поиска с помощью создания структурированного словаря, хотя словарь в таком случае становится немного неоптимальным. Существует два популярных варианта K-MEANS алгоритма, которые создают структурированные словари. LBG алгоритм [77] создает словарь в виде дерева из N-уровней (бинарное дерево). Процедура описанная в [80] делает квантование структурной решетки.

Позднее возник новый класс алгоритмов базирующийся на технологиях нейронных сетей и называемый LVQ алгоритмы (Learning Vector Quantizers) [81]. Эти алгоритмы объединяют задачи обучения и распознавания в одну массивную параллельную вычислительную структуру базирующуюся на архитектуре нейронных сетей. Известно, что LVQ метод теоретически имеет возможность производить оптимальное квантование. Тем не менее, этот метод еще не показал значительную производительность при решении задач распознавания речи.

5.2 Измерение расстояний

Сейчас мы обсудим задачу, которая является основой распознавания речи: измерение расстояний. Очень интересно рассмотреть эту проблему с исторической точки зрения. Во-первых, что такое измерение расстояний? Измерение расстояний должно иметь следующие свойства:

- неотрицательность:
  $D(\bar{x}_1, \bar{x}_2) > 0, \quad x_1 \neq x_2$
  $D(\bar{x}_1, \bar{x}_2) = 0, \quad x_1 = x_2$
- симметрия:
  $D(\bar{x}_1, \bar{x}_2) = D(\bar{x}_2, \bar{x}_1)$
- треугольное неравенство:
  $D(\bar{x}_1, \bar{x}_3) \leq D(\bar{x}_1, \bar{x}_2) + D(\bar{x}_2, \bar{x}_3)$

Измерение евклидового расстояния возможно наиболее известный метод, который удовлетворяет вышеприведенным условиям.

Один из первых введенных методов в распознавание речи базировался на наименьшей предсказанной ошибке и принципах сравнения спектра. Такой метод известен как измерение логарифма вероятности [83]. Этот метод подсчитывает энергию разности спектров двух наборов линейно предсказанных параметров. По

существу оценивается вероятность того, что тестовые данные сгенерированы статистической моделью базирующейся на указанном наборе линейно предсказанных параметров. Часто этот метод называют вероятностным измерением расстояния. Формула для измерения расстояния таким методом выглядит так:

$$D(\bar{y}_1, \bar{y}_2) = \frac{a_{LP_1^+} R_2 a_{LP_1}}{a_{LP_{21}^+} R_2 a_{LP_2}} \tag{70}$$

где R_2 — это автокорреляционная матрица использующаяся для генерации линейно предсказанных параметров для $\bar{x}_2$.

При выполнении задания, такого как векторное квантование (или распознавание речи) в котором циклически сравниваются тестовые данные, деноминатор в уравнении (70) можно исключить или подсчитать один раз. Нумератор может быть эффективно вычислен по уравнению [31]:

$$D(\bar{y}_1, \bar{y}_2) = \frac{\sum_{k=0}^{N_{LP}-1} R_1(k) R_2(k)}{a_{LP_{21}^+} R_2 a_{LP_2}} \tag{71}$$

где

$$R_a(k) = \sum_{i=0}^{N_{LP}-k} a_{LP}(i) \cdot a_{LP}(i+k) \tag{72}$$

где $R_a(k)$ — автокорреляция импульсного отклика инверсного фильтра линейного предсказания.

Видно, что измерение логарифма вероятности является асимметричным (нарушение нашего предыдущего определения измерение расстояния). Тем не менее, эта асимметрия не является значимой проблемой. Но этот метод сейчас используется редко. Тем не менее, он был важным шагом на пути к адаптации вероятностной основы измерения расстояния в распознавании речи.

Раз измерение логарифма вероятности хорошо подходит к линейно предсказанным коэффициентам, то какой вид измерения мы должны использовать для наших параметрических векторов? Одним из элегантных методов измерения в статистической обработке сигналов является вывод расстояния Махаланобиса (Mahalanobis) [72]. В этом выводе показано, что вероятность того, что вектор, принадлежащий многомерному гауссовому распределению, можно выразить как взвешенное евклидово расстояние:

$$D(\bar{y}, \bar{\mu}) = (\bar{y} - \bar{\mu}) \cdot \underline{C}^{-1} \cdot (\bar{y} - \bar{\mu})^+ \tag{73}$$

где μ и $\underline{C}$ — среднее значение и ковариация распределения. Очевидно, что если мы работаем с предварительно преобразованными векторами, то ковариантная матрица будет тождественной матрицей, а из уравнения (73) получаем евклидово расстояние в квадрате.

По этой причине в настоящее время наиболее часто применяется евклидово измерение расстояний, так как много используемых наборов параметров базируется на нечеткой декорреляции параметров, такие как кэпстральные коэффициенты. Также, в настоящее время системы распознавания речи уже развились для того, чтобы работать с этими видами преобразований [73, 74].

Во многих приложениях, таких как векторное квантование, есть возможность использовать факторную форму евклидового расстояния:

$$D(\bar{y}_1, \bar{y}_2) = \|\bar{y}_1 - \bar{y}_2\|^2 = (\bar{y}_1 - \bar{y}_2)(\bar{y}_1 - \bar{y}_2)^+ = \|\bar{y}_1\|^2 + \|\bar{y}_2\|^2 - 2(\bar{y}_1 \cdot \bar{y}_2) \quad (74)$$

Мы видим, что евклидово расстояние является суммой величин векторов без двойного произведения. Предположим, что мы желаем векторно проквантовать входной вектор $\bar{y}_1$. Пусть $\bar{y}_2$ — каждая запись в словаре с которой будет сравниваться входной вектор. Первое слагаемое будет постоянным по отношению к каждой записи в словаре, его можно не принимать во внимание. Второе слагаемое — это скаляр, который можно добавить в результат третьего слагаемого. Если каждая запись в словаре имеет одинаковую величину, то второе слагаемое можно тоже отбросить.

Третье слагаемое должно вычисляться для каждой записи. Этот тип вычисления можно использовать в большинстве приложений речевой обработки, которые используют векторное квантование и распознавание речи[8].

Мы завершили наше обсуждение методов моделирования сигналов в распознавании речи. Мы показали, что модель сигнала может быть собрана из последовательности трех операций: измерения сигнала, сглаживания параметров и статистического моделирования. В следующем разделе мы обсудим как эти концепции можно применить в распознавании речи и прокомментируем относительные достоинства этих методов.

[8] Однажды консультант одной знаменитой компании, занимающейся суперкомпьютерами, сказал мне, что "все в жизни можно привести к одной из двух операций: умножению или векторному умножению/сложению". В ходе своих интенсивных исследований по оптимизации кода для суперкомпьютеров он обнаружил, что 99% всего времени выполняются одна из этих операций.

VI. Практические примеры

В настоящее время существует большое количество комбинаций моделей, обсужденных нами выше. Позвольте начать этот раздел простым перечислением существующих систем распознавания речи. Полный обзор дан в таб. 2. Таблица разделена на четыре секции: авторство (для ссылок), измерение сигналов, параметры сигнала и статистические модели. Мы обсудили каждую из этих тем в предыдущих разделах этого документа. Обратите внимание на то, что список аббревиатур и их значение даны после таблицы. Так как таблица достаточно велика, то мы сначала обсудим ее содержимое.

6.1 Обзор таблицы

Авторская секция таблицы является простым алфавитным идентификатором. Ссылки ассоциированы с записями, которые содержать достаточно информации о модели сигнала используемой в самой последней опубликованной системе распознавания. Многие из систем упомянутых в таб. 2 развились со временем.

Второй столбик содержит краткое описание типа приложения. "Малый" обозначает задачу распознавания речи с использованием малого словаря, обычно менее 100 слов. Непрерывное распознавание цифр и распознавание букв принадлежит этой категории. "Средний" (порядка 1000 слов) и "Большой" (более 5000 слов) — соответствующие задачи распознавания.

Также мы квалифицируем приложения по акустическому фону. "Офис" обозначает систему использующую базу данных созданную либо в нормальной офисной обстановке (обычно около 70 dB SPL), либо в акустически изолированной комнате. "Тел" обозначает использование в телекоммуникационных приложениях, использующих стандартные аналоговые телефонные линии. "Воен" обозначает военные приложения, которые включают в себя широкий диапазон фонового шума и речевых стилей.

Секция измерения сигнала делится на пять частей. Первая содержит частоту оцифровки в системе. Хотя в настоящее время большинство систем может программно переключать частоту оцифровки, эта запись показывает частоту оцифровки экспериментальной базы данных. Это сделано для сравнения.

Следующие три столбца содержат информацию по спектральному анализу: a_{pre} — постоянная взвешивающего фильтра первого порядка; "Кадр" — длительность кадра при анализе; "Окно" — размер анализирующего окна. Мы специально не указываем точным тип окна наложенного на сигнал перед спектральным анализом, так как все системы представленные здесь используют некую форму обобщенного окна Ханнинга (большинство использует окно Хамминга, в то время как остальные пользуются окном Ханнинга).

Последний столбик в этой секции с названием "Спектральный анализ" относится к последовательности операций, которые включены в измерение. Например, под записями "AT&T [7]", LP(8) и CEP(12) понимается, что для получения кэпстральных измерений сигнала применили сначала анализ линейного предсказания восьмого порядка, затем кэпстральный анализ двенадцатого порядка. Большинство систем, которые используют линейно предсказанные кэпстральные коэффициенты, будут обладать несколькими записями в этом столбце, показывая, что были выполнены анализ линейного предсказания и кэпстральный анализ.

Следующая секция занимает восьмой столбик таблицы и содержит элементы вектора параметров сигнала. Обычно запись содержит набор абсолютных измерений, таких как "Mel-Cep", и производные времени, такие как "D-Cep". Обратитесь к списку аббревиатур после таблицы.

Последняя секция таблицы (столбцы 9 и 10) содержит описание используемой статистической модели. Тип метода распознавания речи использованного в модели сигнала показан в основном для указательной цели[9].

[9] Эти методы не были обсуждены в этом документе. существует два основных класса: Скрытые марковские модели (СММ) и нейросети (НС). См. [27].

VII. Заключение

Мы представили вашему вниманию несколько популярных методов анализа сигналов в применении к задаче распознавания речи, что выявило важность точного спектрального анализа и статистической нормализации. Разница между этими методами невелика, по сравнению со всеми аспектами задачи распознавания. Все методы работают с основными понятиями: изменяющиеся со временем данные, преобразования и нормализация параметров.

Обзор современных систем показывает, что кэпстральные коэффициенты полученные с помощью Фурье-преобразования доминируют в этой области. Линейное предсказание отошло на второй план. Вектор параметров сигнала содержащий кэпстральные коэффициенты, первая производная от кэпстральных коэффициентов, энергия и производная от энергии стали стандартом. Вариативное взвешивание этого вектора параметров является наиболее популярным методом нормализации.

Все еще остались несколько вопросов: устойчивость к шуму, невосприимчивость к частоте оцифровки, универсальность в распознавании. Интересно отметить, что несмотря на кажущиеся обширные алгоритмические различия этих методов, многие из них успешно используются. Довольно часто ощутимые различия систем распознавания лежат вне пределов моделей сигналов.

Нельзя сказать, что проблема распознавания речи полностью решена. Производительность систем распознавания речи все еще отстает от производительности человеческой слуховой системы. Например, в задаче распознавания цифр, когда словарь мал и все отдано акустическому моделированию, производительность системы как минимум на два порядка ниже производительности человека [73, 115]. В шумной обстановке, в телекоммуникационных системах, пространство между производительностью системы и человеком очень велика.

Главной движущей силой в настоящее время является минимизация числа степеней свободы системы. Так как современные системы распознавания речи имеют большое число свободных переменных (порядка 10000 переменных) недостаточный объем обучающих данных является настоящей проблемой. В течении всех лет изучения задачи распознавания речи мы узнали одну важную вещь: плохо обученные параметры часто становятся причиной значительного снижения производительности системы. Следовательно, методы уменьшающие количество параметров, подобно вариативному взвешиванию, более предпочтительны из-за их статистической оптимальности (предварительные преобразования), но им необходим гораздо больший объем обучающих данных.

VIII. Ссылки

1. D.S. Pallet, "Speech Results on Resource Management Task", in *Proceedings of the February 1989 DARPA Speech and Natural Language Workshop*, Morgan Kaufman Publishers, Inc., Philadelphia, PA, USA, pp.18-24, February 1989.
2. D. Paul, "The Lincoln Robust Continuous Speech Recognizer", in *Proceedings IEEE International Conference on Acoustic, Speech, and Signal Processing*, pp. 556-559, Glasgow, Scotland, May 1989.
3. J.G. Wilpon, R.P. Mikkilineni, D.B. Roe, and S. Gokcen, "Speech Recognition: From the Laboratory to the Real World", *AT&T Technical Journal,* vol. 69, no. 5, pp. 14-24, October 1990.
4. J.G. Wilpon, D.M. DeMarco, R.P. Mikkilineni, "Isolated Word Recognition Over the DDD Telephone Network – Results Of Two Extensive Field Trials", in *Proceedings IEEE International Conference on Acoustic, Speech, and Signal Processing,* pp. 55-57, New York, NY, USA April 1988.
5. B. Wheatly and J. Picone, "Voice Across America: Toward Robust Speaker Independent Speech recognition For Telecommunications Applications", *Digital Signal Processing: A Review Journal*, vol. 1, no. 2, pp. 45-64, April 1991.
6. J. Picone, "The Demographics of Speaker Independent Digit Recognition", in *Proceedings IEEE International Conference on Acoustic, Speech, and Signal Processing,* pp. 105-108, Albuquerque, New Mexico, USA, April 1990.
7. J.G. Wilpon, C.H. Lee, and L.R. Rabiner, "Improvements in Connected Digit recognition Using Higher Order Spectral and Energy Features", in *Proceedings IEEE International Conference on Acoustic, Speech, and Signal Processing,* pp. 349-352, Toronto, Ontario, Canada, May 1991.
8. T. Matsuoka and K. Shikano, "Robust HMM Phoneme Modeling For Different Speaking Styles", in *Proceedings IEEE International Conference on Acoustic, Speech, and Signal Processing,* pp. 265-268, Toronto, Ontario, Canada, May 1991.
9. Y.T. Lee, "Information-Theoretic Distortion Measures for Speech Recognition: Theoretical Considerations and Experimental Results", *IEEE Transactions on Signal Processing,* vol. 39, no. 2, pp. 330-335, February 1991.
10. Y.T. Lee and D. Kahn, "Information-Theoretic Distortion Measures for Speech Recognition: Theoretical Considerations and Experimental Results", in *Proceedings IEEE International Conference on Acoustic, Speech, and Signal Processing,* pp. 785-788, Albuquerque, New Mexico, USA, April 1990.
11. S. Furui, "On the Use of Hierarchical Spectral Dynamics in Speech Recognition", in *Proceedings IEEE International Conference on Acoustic, Speech, and Signal Processing,* pp. 789-792, Albuquerque, New Mexico, USA, April 1990.
12. D. Mansour and B.H. Juang, "A Family of Distortion Measures Based Upon Projection Operation for Robust Speech Recognition", *IEEE Transactions on Acoustic, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1659-1671, November 1989.
13. Y. Tohkura, "A Weighted Cepstral Distance Measure For Speech Recognition", *IEEE Transactions on Acoustic, Speech, and Signal Processing*, vol. 35, no. 10, pp. 1414-1422, October 1987.

14. G.R. Doddington, "Phonetically Sensitive Discriminants for Improved Speech Recognition", in *Proceedings IEEE International Conference on Acoustic, Speech, and Signal Processing,* pp. 556-559, Glasgow, Scotland, May 1989.

15. M.J. Hunt and C. Lefebre, "A Comparison of Several Acoustic Representations for Speech recognition with Degraded and Undegraded Speech", in *Proceedings IEEE International Conference on Acoustic, Speech, and Signal Processing,* pp. 262-265, Glasgow, Scotland, May 1989.

16. B.H. Juang, L.R. Rabiner, and J.G. Wilpon, "On the Use of Bandpass Liftering in Speech Recognition", *IEEE Transactions on Acoustic, Speech, and Signal Processing*, vol. 35, no. 7, pp. 947-954, July 1987.

17. V.N. Gupta, M. Lennig, and P. Mermelstein, "Integration Of Acoustic Information In A Large Vocabulary Word Recognizer", in *Proceedings IEEE International Conference on Acoustic, Speech, and Signal Processing,* pp. 697-700, Dallas, Texas, USA, April 1987.

18. E.L. Bocchieri and G.R. Doddington, "Frame Specific Statistical Features for Speaker-Independent Speech Recognition", *IEEE Transactions on Acoustic, Speech, and Signal Processing*, vol. 34, no. 4, pp. 755-764, August 1986.

19. P.K. Rajasekaran, G.R. Doddington, J. Picone, "Recognition of Speech Under Stress and in Noise", in *Proceedings IEEE International Conference on Acoustic, Speech, and Signal Processing,* pp. 733-736, Tokyo, Japan, April 1986.

20. K. Shikano, "Evaluation of LPC Spectral Matching Measures for Phonetic Unit Recognition", TM No. CMU-CS-86-108, Computer Science Department, Carnegie-Mellon, University, Pittsburgh, PA, US, 15213, February 3, 1986.

21. S. Furui, "Speaker-Independent Isolated Word Recognition Using Dynamic Features of the Speech Spectrum", *IEEE Transactions on Acoustic, Speech, and Signal Processing*, vol. 34, no. 1, pp. 52-59, February 1986.

22. J. Makhoul, S. Raucos, and H. Gish, "Vector Quantization In Speech Coding", in *Proceedings of the IEEE,* vol. 73, no. 11, pp. 1551-1588, November 1985.

23. N. Nocerino, F.K. Soong, L.R. Rabiner, and D.H. Klatt, "Comparative Study Of Several Distortion Measures For Speech Recognition", in *Proceedings IEEE International Conference on Acoustic, Speech, and Signal Processing,* pp. 25-28, Tampa, Florida, USA, March 1985.

24. L.R. Rabiner, K.C. Pan, and F.K. Soong, "On the Performance of Isolated Word Speech Recognizers Using Vector Quantization and Temporal Energy Contours", *AT&T Bell Labs Technical Journal*, vol. 63, no. 7, pp. 1245-1260, September 1984.

25. L.R. Rabiner, S.E. Levinson, M.M. Sondhi, "On the Application of Vector Quantization and Hidden Markov Models to Speaker-Independent, Isolated Word Recognition", *Bell System Technical Journal*, vol. 62, no. 4, pp. 1075-1105, April 1983.

26. S.B. Davis and P. Mermelstein, "Comparison of Parametric Representation of Monosyllabic Word Recognition in Continuously Spoken Sentences", *IEEE Transactions on Acoustic, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357-366, August 1980.

27. A.V. Oppenheim and R.W. Schafer, *Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1975.

28. J.R. Deller, J.G. Proakis, J.H.L. Hansen, *Discrete Time Processing of Speech Signals*, MacMillian Publishing Co., New York, NY, USA, 1993.

29. L. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1993.

30. J.G. Proakis, *Digital Communications*, McGraw-Hill, 2nd Ed., New York, NY, USA, 1989.

31. J. Markel and A.H. Gray, Jr., *Linear Prediction of Speech*, Springer-Verlag, New York, NY, USA, 1980.

32. L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1978.

33. A.M. Noll, "Problems of Speech Recognition in Mobile Environments", in *Proceedings if the International Conference on Spoken Language Processing*, pp. 1133-1136, Kobe, Japan, November 1990.

34. Y. Nakadai and N. Sugamura, "A Speech recognition Method For Noise Environments Using Dual Inputs", in *Proceedings if the International Conference on Spoken Language Processing*, pp. 1141-1144, Kobe, Japan, November 1990.

35. S. Tamura, "An Analysis of a Noise Reduction Neural Network", in *Proceedings IEEE International Conference on Acoustic, Speech, and Signal Processing,* pp. 2001-2004, Glasgow, Scotland, May 1989.

36. J.L. Hieronumus, D. McKelvie, and F.R.McInnes, "Use of Acoustic Sentence Level and Lexical Stress in HSMM Speech Recognition", in *Proceedings IEEE International Conference on Acoustic, Speech, and Signal Processing,* pp. 225-227, San Francisco, California, USA, March 1992.

37. T. Matsui and S. Furui, "A Text-Independent Speaker Recognition Method Robust Against Utterance Variations", in *Proceedings IEEE International Conference on Acoustic, Speech, and Signal Processing,* pp. 377-380, Toronto, Ontario, Canada, April 1991.

38. W. Hess, *Pitch Determination Of Speech Signals*, Springer-Verlag, New York, NY, USA, 1983.

39. P. Papamichalis, *Practical Approaches To Speech Coding*, Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1987.

40. D. Gold and L.R. Rabiner, "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain", *Journal of the Acoustical Society of America*, vol. 46, no. 2, pt. 2, pp. 442-448, August 1969.

41. R.S. Sukkar, J.L. LoCicero, and J. Picone, "Design and Implementation of a Parallel Processing Based Pitch Detector", *IEEE Journal on Selected Areas of Communications*, vol. 6, no. 2, pp. 441-451, February 1988.

42. J, Campbell and T.E. Tremain, "Voiced/Unvoiced Classification of Speech with Applications to the U.S. Government LPC-10E Algorithm", in *Proceedings IEEE International Conference on Acoustic, Speech, and Signal Processing,* pp. 473-476, Tokyo, Japan, April 1986.

43. V.C. Welch, T.E. Tremain, and J.P. Campbell, Jr., "A Comparison of U.S. Government Standard Voice Coders", *IEEE Military Communications Conference Record*, pp. 269-273, September 1989.

44. J. Picone, G.R. Doddington, and B.G. Secrest, "Robust Pitch Detection in a Noisy Telephone Environment", in *Proceedings IEEE International Conference on Acoustic, Speech, and Signal Processing,* pp. 1442-1445, Dallas, Texas, USA, April 1987.

45. A.M. Noll, "Cepstrum Pitch Determination", *Journal of the Acoustical Society of America*, vol. 41, no. 2, pp. 293-309, February 1967.

46. G. von Békésy, *Experiments in Hearing*, McGraw-Hill Book Company, New York, NY, 1960.

47. J. Picone, "Analytic Signal Processing", Ph.D. Dissertation, Illinois Institute of Technology, Chicago, Illinois, USA, December 1983.

48. K. Ogata, *Modern Control Engineering*, Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1970.

49. J.O. Pickles, *An Introduction to the Physiology of Hearing*, Academic Press, New York, NY, USA, 1983.

50. A.R. Møller, *Auditory Physiology*, Academic Press, New York, NY, USA, 1983.

51. D. O'Shaughnessy, *Speech Communication: Human and Machine*, Addison Wesley, New York, NY, USA, 1987.

52. E. Zwicker and E. Terhardt, "Analytical expression for critical-band rate and critical bandwidth as a function of frequency", *Journal of the Acoustical Society of America*, vol. 68, no. 5, pp. 1523-1525, December 1980.

53. J.B. Allen, "Cochlear Modeling", *IEEE ASSP Magazine*, vol. 3, no. 3, pp. 3-29, September 1985.

54. S. Seneff, "A Joint Synchrony/Mean-Rate Model of Auditory Speech Processing", *Journal of Phonetics*, vol. 16, no. 1, pp. 55-76, January 1988.

55. O.E. Brigham, *The Fast Fourier Transform*, Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1974.

56. B.S. Atal and S.L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave", *Journal of the Acoustical Society of America*, vol. 50, no. 2, pp. 637-655, March 1971.

57. S.L. Marple, Jr., *Digital Spectral Analysis With Applications*, Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1987.

58. V.R. Viswanathan and J. Makhoul, "Quantization Properties of Transmission Parameters in Linear Predictive Systems," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 3, pp. 309-321, June 1975.

59. B.S. Atal, "Predictive Coding of Speech at Low Bit Rates," *IEEE Transactions on Communications*, vol. 30, no. 4, pp. 600-614, April 1982.

60. B.S. Atal and J.R. Remde, "A New Model of LPC Excitation for Producing Natural Sounding Speech at Low Bit Rates," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 614-617, Paris, France, 1982.

61. G.R. Doddington, J. Picone, and J.J. Godfrey, "The LPC trace as an HMM development tool," *Journal of the Acoustical Society of America*, Vol. 84, pg. S1-561A, Fall 1988.

62. H.F. Silverman and N.R. Dixon, “A Parametrically Controlled Spectral Analysis System for Speech,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 22, no. 5, pp. 362-381, October 1974 (not the original reference by any means, but a good reference on digital computation of the spectrogram — original references on analog techniques date back to the 1940s).

63. R.V. Churchill, J.W. Brown, and R.F. Verhey, Complex Variables and Applications, McGraw-Hill, New York, New York, USA, 1976.

64. B.S. Atal, “Linear Prediction For Speaker Identification,” *Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304-1311, June 1974.

65. A.V. Oppenheim and D.H. Johnson, “Discrete Representation of Signals,” in *Proceedings of the IEEE*, vol. 60, no. 6, pp. 681-691, June 1972.

66. K.F. Lee, *Automatic Speech Recognition: the Development of the SPHINX System*, Kluwer Academic Publishers, Boston, Massachusetts, USA, 1989.

67. J.G. Wilpon, C.H. Lee, and L.R. Rabiner, “Application of Hidden Markov Models for Recognition of a Limited Set of Words in Unconstrained Speech,” in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 254-257, Glasgow, Scotland, 1989.

68. R.W. Hamming, *Digital Filters*, Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1989 (2nd Edition).

69. H. Ney, “Experiments on Mixture-Density Phoneme Modeling For the Speaker-Independent 1000-Word Speech Recognition DARPA Task,” in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 713-716, Albuquerque, New Mexico, USA, April 1990.

70. D. Paul, “A Speaker-Stress Resistant Isolated Word Recognizer,” in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 713-716, Dallas, Texas, USA, April 1990.

71. B.H. Juang, L.R. Rabiner, and J.G. Wilpon, “On the Use of Bandpass Liftering in Speech Recognition,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 7, pp. 947-954, July 1987.

72. K. Fukunaga, *Introduction To Statistical Pattern Recognition*, Academic Press, New York, New York, USA, 1972.

73. J. Picone, “Continuous Speech Recognition Using Hidden Markov Models,” *IEEE ASSP Magazine*, vol. 7, no. 3, pp. 26-41, July 1990.

74. L.R. Rabiner, “A Tutorial on Hidden Markov Models And Selected Applications In Speech Recognition,” in *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257-285, February 1989.

75. W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vettering, *Numerical Recipes in C: The Art of Scientific Programming*, Cambridge University Press, New York, New York, USA, 1988.

76. M.R. Anderberg, *Cluster Analysis for Applications*, Academic Press, New York, USA, 1973.

77. Y. Linde, A. Buzo, and R.M. Gray, “An Algorithm for Vector Quantizer Design,” *IEEE Transactions on Communications*, vol. 28, no. 1, pp. 84-95, January 1980.

78. S.E. Levinson, L.R. Rabiner, A.E. Rosenberg, and J.G. Wilpon, “Interactive Clustering Techniques for Selecting Speaker-Independent Reference Templates for Isolated Word Recognition,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 134-141, April 1979.

79. J. Picone and G.R. Doddington, “Low Rate Speech Coding Using Contour Quantization,” in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1652-1655, Dallas, Texas, April 1987.

80. A. Gersho, “On the Structure of Vector Quantizers,” *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 157-166, March 1982.

81. T. Kohonen, *Self-Organization and Associative Memory*, Third Ed., Springer-Verlag, New York, New York, USA, 1989.

82. R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, Academic Press, New York, New York, USA, 1973.

83. F. Itakura, “Minimum Prediction Residual Principle Applied To Speech Recognition,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 1, pp. 67-72, February 1975.

84. B. Dautrich, L.R. Rabiner, T.B. Martin, “On the Effects of Varying Filter Bank Parameters on Isolated Word Recognition,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 31, no. 4, pp. 793-807, August 1983.

85. J. Picone, “Duration In Context Clustering for Speech Recognition,” *Speech Communication*, vol. 9, No. 2, pp. 119-128, April 1990.

86. K. Kita, T. Kawabata, H. Saito, “HMM Continuous Speech Recognition Using Predictive LR Parsing,” in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 703-706, Glasgow, Scotland, May 1989.

87. A. Waibel, T. Hanazawa, G. Hintom, K. Shikano, and K.J. Lang, “Phoneme Recognition Using Time-Delay Neural Networks,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 328-339, March 1989.

88. Y.L. Chow, M.O. Dunham, O.A. Kimball, M.A. Krasner, G.F. Kubala, J. Makhoul, P.J. Price, S. Roucos, and R.M. Schwartz, “BYBLOS: The BBN Continuous Speech Recognition System,” in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 89-92, Dallas, Texas, USA, April 1987.

89. F. Kubala, M. Feng, J. Makhoul, and R. Schwartz, “Speaker Adaptation from Limited Training in the BBN BY-BLOS Speech Recognition System,” in *Proceedings of the DARPA Speech and Natural Language Workshop*, Morgan Kaufmann Publishers, Inc., Palo Alto, California, USA, pp. 100-105, February 1989.

90. M.M. Hochberg, L.T. Niles, J.T. Foote, and H.F. Silverman, “Hidden Markov Model/Neural Network Training Techniques for Connected Alphadigit Speech Recognition,” in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 109-112, Toronto, Ontario, Canada, April 1991.

91. T.D. Harrison, and F. Fallside, “A Con-nectionist Model for Phoneme Recognition in Continuous Speech,” in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 417-420, Glasgow, Scotland, May 1989.

92. P. Haffner, M. Franzini, and A. Waibel, “Integrating Time Alignment and Neural Networks for High Performance Continuous Speech Recognition,” in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 105-109, Toronto, Ontario, Canada, April 1991.

93. L. Fissore, P. Laface, G. Micca, “Comparison of Discrete and Continuous HMMs in a CSR Task Over the Telephone,” in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 253-256, Toronto, Ontario, Canada, April 1991.

94. L. Fissore, P. Laface, G. Micca, and R. Pieraccini, “Lexical Access to Large Vocabularies for Speech Recognition,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 8, pp. 1197-1213, August 1989.

95. S. Kimura, “100,000-Word Recognition Using Acoustic-Segment Networks,” in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 61-64, Albuquerque, New Mexico, USA, April 1990.

96. A. Averbuch, L. Bahl, R. Bakis, P. Brown, A. Cole, G. Daggett, S. Das, K. Davies, S. De Gennaro, P. de Souza, E. Epstein, D. Fraleigh, F. Jelinek, S. Katz, B. Lewis, R.

Mercer, A. Nadas, D. Nahamoo, M. Picheny, G. Shichman, and P. Spinelli, “An IBM-PC Based Large-Vocabulary Isolated Utterance Speech Recognizer,” in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 53-56, Tokyo, Japan, April 1986.

97. F. Jelinek, “The Development of an Experimental Discrete Dictation Recognizer,” in *Proceedings of the IEEE*, vol. 73, no. 11, pp. 1616-1624, November 1985.

98. L. Deng, V. Gupta, M. Lennig, P. Kenny, and P. Mermelstein, “Acoustic Recognition Component of an 86,000-word Speech Recognizer,” in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 741-784, Albuquerque, New Mexico, USA, April 1990.

99. J. Koo, C.K. Un, H.S. Lee, H.R. Kim, and M.W. Koo, “A Recognition Time Reduction Algorithm for Large-Vocabulary Speech Recognition,” in *Proceedings of the International Conference on Spoken Language Processing*, pp. 253-256, Kobe, Japan, November 1990.

100. V. Zue, J. Glass, M. Phillips, and S. Seneff, “Acoustic Segmentation and Phonetic Classification in the SUM-MIT System,” in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 389-392, Glasgow, Scotland, May 1989.

101. S. Mizuta and K. Kakajima, “An Optimal Discriminative Training Method for Continuous Mixture Density HMMs,” in Proceedings of the International Conference on Spoken Language Processing, pp. 245-248, Kobe, Japan, November 1990.

102. K. Yoshida, T. Watanabe, and S. Koga, “Large Vocabulary Word Recognition Based on Demisyllable Hidden Mark-ov Model Using Small Amount of Training Data,” in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1-4, Glasgow, Scotland, 1989.

103. D. Lubensky, “Word Recognition Using Neural Nets, Multi-State Gaussian and K-Nearest Neighbor Classifiers,” in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 141-144, Toronto, Ontario, Canada, April 1991.

104. S. Matsunaga, S. Sagayama, S. Homma, and S. Furui, “A Continuous Speech Recognition System Based on A Two-Level Grammar Approach,” in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 589-592, Albuquerque, New Mexico, USA, April 1990.

105. T. Matsuoka and K. Shikano, “Robust HMM Phoneme Modeling for Different Speaking Styles,” in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 265-268, Toronto, Ontario, Canada, April 1991.

106. Y. Zhao and H. Wakita, “Experiments With A Speaker-Independent Continuous Speech Recognition System on the TIMIT Database,” in *Proceedings of the International Conference on Spoken Language Processing*, pp. 697-700, Kobe, Japan, November 1990.

107. V. Steinbiss, A. Noll, A Paeseler, H. Ney, H. Bergmann, C. Dugast, H.H. Hamer, H. Piotrowski, H. Tomaschewski, A. Zielinski, “A 10,000-Word Continuous Speech Recognition System,” in *Proceedings IEEE International Confer-ence on Acoustics, Speech, and Signal Processing*, pp. 57-60, Albuquerque, New Mexico, USA, April 1990.

108. M.J. Russel, K.M. Ponting, S.M. Peeling, S.R. Browning, J.S. Bridle, R.K. Moore, I. Galiano, P. Howell, “The ARM Continuous Speech Recognition System,” in *Proceed-*

ings IEEE Inter-national Conference on Acoustics, Speech, and Signal Processing, pp. 69- 72, Albuquerque, New Mexico, USA, April 1990.

109. M. Weintraub, H. Murveit, M. Cohen, P. Price, J. Bernstein, G. Baldwin, and D. Bell, “Linguistic Constraints in Hidden Markov Model Based Speech Recognition,” in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 699-702, Glasgow, Scotland, 1989.

110. H. Murveit, and M. Weintraub, “1000- Word Speaker-Independent Continuous-Speech Recognition Using Hidden Markov Models,” in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 115-118, New York, NY, USA, April 1988.

111. W.S. Meisel, M.T. Anikst, S.S. Pirzadeh, J.E. Schumacher, M.C. Soares, and D.J. Trawick, “The SSI Large Vocabulary Speaker-Independent Continuous Speech Recognition System,” in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 273-276, Toronto, Ontario, Canada, April 1991.

112. S. Makino, A. Ito, M. Endo, and K. Kido, “A Japanese Text Dictation System Based on Phoneme Recognition and a Dependency Grammar,” in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 273-276, Toronto, Ontario, Canada, April 1991.

113. K. Shirai, N. Hosaka, E. Kitagawa, and T. Endou, “Speaker Adaptable Phoneme Recognition Selecting Reliable Acoustic Features based on Mutual In-formation,” in *Proceedings of the International Conference on Spoken Language Processing*, pp. 353-356, Kobe, Japan, November 1990.

114. H.M. Meng and V.W. Zue, “A Compara-tive Study of Acoustic Representations of Speech for Vowel Classification Using Multi-Layer Perceptrons,” in *Proceedings of the International Conference on Spoken Language Processing*, pp. 1053-1056, Kobe, Japan, November 1990.

115. R.G. Leonard, “A Database for Speaker-Independent Digit Recognition,” in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 42.11.1- 42.11.4, San Diego, California, USA, April 1984.

116. H. Hermansky and N. Morgan, “Towards Handling the Acoustic Environment in Spoken Language Processing,” in *Proceedings of the International Conference on Spoken Language Processing*, pp. 85-88, Banff, Alberta, Canada, October 1992.

IX. Описание рисунков

1. Показан обзор процесса моделирования сигнала. Успехи в технологии распознавания речи сгладили различие между моделированием сигнала и статистическим моделированием с помощью введения контекстно-зависимых статистических моделей в параметризацию сигнала.
2. Показана последовательность операций для преобразования аналогового сигнала в цифровой сигнал, который пригоден для спектрального анализа. Некоторые компоненты, такие как низкокачественный АЦП или нелинейный микрофон, могут внести не желаемые изменения в сигнал.
3. Показан частотный отклик типичного телефонного АЦП.
4. Показан частотный отклик взвешивающего фильтра используемого в системах распознавания речи. Предназначение таких фильтров в том, чтобы спектрально выровнять речевой сигнал и усилить важные области спектра. Обычные значения близки к 1.0.
5. Показаны отклики временных и частотных областей окна Ханнинга для диапазона значений. В районе границ окна отклик должен быть близок к нулю. a) Временной отклик; b) Частотный отклик.
6. Показан покадровый перекрывающий анализ. Здесь показаны 33% перекрытия. Третья часть данных используемых при анализе кадра совпадает с предыдущем кадром. Обратите внимание на то, что только треть данных уникальна для каждого кадра — оставшиеся две трети перекрываются соседними кадрами.
7. Показаны различные вычисления энергии для речевого сигнала (слово “tea”) на рис. (g). На рис. (a) показаны прямоугольные окна с размерами 5 мс и 10 мс. На рис. (b) показаны прямоугольные окна с размерами 10 мс и 20 мс. На рис. (c) показаны прямоугольные окна с размерами 20 мс и 30 мс. На рис. (d) показаны окна Хамминга с параметрами 20 мс и 30 мс. На рис. (e) показано сильное сглаживание с помощью окна размером 60 мс, в то время как размер кадр фиксирован на 20 мс. Наконец, на рис. (f) энергия подсчитывается с помощью рекурсивного метода применения 50 Гц фильтра низких частот второго порядка. Стрелки указывают на точки быстрого изменения энергии. Обратите внимание на то, что применение окна Хамминга на рис. (d) помогает сгладить такие пики.
8. Показаны шесть основных алгоритмов спектрального анализа. Кэпстральные параметры, полученные либо с помощью Фурье-преобразования, либо с помощью линейного преобразования, являются наиболее популярными из них. Методы, использующие Фурье-преобразования, традиционно считались устойчивыми к различному фоновому шуму и популярны за их схожесть с начальными шагами человеческой слуховой системы.
9. Bark- и Mel-шкалы показаны как функции акустической частоты на рис. (a) и (b) соответственно. На рис. (c) показан критический диапазон как функция частоты.
10. На рис. (a) показан выходной сигнал банка цифровых фильтров для речи, содержащей слово "speech". На рис. (b) показан выходной сигнал фильтра с центральной частотой 250 Гц и диапазоном 100 Гц. На рис. (c) показан выходной сигнал фильтра центрованного на частоту 2500 Гц. Обратите внима-

ние на то, что амплитуда выходного сигнала каждого фильтра варьируется в зависимости природы сигнала. Завершающий звук "ch", например, в основном состоит из высокочастотной информации.

11. Упрощенный пример преимущества перегруженного спектра. Показан спектр сигнала (подсчитанный с помощью дискретного Фурье-преобразования, если быть точным) по значениям частот, на которых он был оцифрован с помощью банка фильтров содержащим 5 наборов на секцию (показан индекс 14 из таб. 1). Если спектр оцифрован точно на центральной частоте критического диапазона, то выходное значение будет равно 0 dB. Если использовалось среднее число спектра на критическом диапазоне, то значение будет равно –1 dB. Перегрузка спектра часто дает результаты при более стабильных, или сглаженных, оценок амплитуд.

12. Области низких энергий спектра часто отсекается для взвешивания высокочастотных порций спектра в модели сигнала. Отсечение обычно выполняется после взвешивания, т.о. высокочастотные компоненты речевого спектра не сильно страдают.

13. В линейной акустической модели речевоспроизведения [32], речевой сигнал воспроизводится с помощью наложения на возбуждающий сигнал (создаваемый голосовой щелью) вариативный линейный фильтр (голосовой тракт). Голосовой тракт может быть отделен от возбуждающего сигнала с помощью методов гомоморфичной обработки сигналов. Необходимо отметить, что эта модель не верна для всех классов речевых звуков, таких как фрикативный звук, где возбуждение происходит выше голосовой щели.

14. Показан пример вычисления кэпстра. На рис. (a) показан участок сигнала без речи. На рис. (b) подсчитан кэпстр 1000 точек. На рис. (c) показан сигнал с речью. Наконец, на рис. (d) показан соответствующий кэпстр. Обратите внимание, что кэпстр на рис. (d) отражает периодичность в сигнале, представленную двумя локальными максимумами. Величины малого порядка в кэпстре отражает сглаженную спектральную структуру речевого сигнала (информация голосового тракта).

15. Показан речевой спектр для моделей линейного предсказания четвертого и восьмого порядка. Обратите внимание, что модель четвертого порядка недостаточна для детализации спектра. Модели десятого и двенадцатого порядка часто используются в системах распознавания речи.

16. Показана стабилизация модели линейного предсказания. Речевой спектр показан для модели линейного предсказания десятого порядка и та же модель с –10 dB стабилизирующим фактором. Обратите внимание, что пока увеличивается низ спектральной модели, производительность модели около спектральных пиков также значительно сглажена. В случаях, когда два спектральных резонанса близки по частоте, стабилизация часто объединяет их в один широкий спектральный пик.

17. Показан спектральный анализ речевого сигнала в виде широкополосной спектрограммы речевого сигнала и трех соответствующих моделей линейного предсказания. Обычно, анализ показанный на рис. (e) достаточен для захвата ключевых аспектов определенных звуков. Тем не менее, с увеличением вычислительной мощности и с увеличением технологии фонетического распознавания, кадры размеров 10 мс становятся стандартом из-за необходимости лучшей характеризации динамических звуков, таких как согласные. На рис. (a) показан речевой сигнал (слово "speech") и его график энергий. На рис. (b) показана широкополосная спектрограмма (окно 6 мс). На рис. (c) по-

казана спектрограмма модели линейного предсказания (кадр – 5 мс, окно – 10 мс). На рис. (d) показана спектрограмма модели линейного предсказания (кадр – 10 мс, окно – 20 мс). На рис. (e) показана спектрограмма модели линейного предсказания (кадр – 20 мс, окно – 30 мс).

18. Билинейное преобразование сравнима с Mel-шкалой для диапазона значений. Билинейное преобразование подсчитывается при оцифровке с частотой 16 кГц. Обратите внимание, что положительные значения α_{bt} дают сжатие частотной шкалы, а отрицательные — расширение.

19. Показан речевой спектр и его модель линейного предсказания. Дополнительно, показаны логарифм спектра линейно предсказанных кэпстральных коэффициентов и логарифм спектра соответствующих преобразованных кэпстральных коэффициентов ($\alpha_{bt} = 0.25$). Подобные результаты могут быть получены для линейно предсказанного спектра обработкой либо коэффициентов линейного предсказания, либо автокорреляционной функции.

20. Преобразование измерений сигналов в вектор параметров сигнала обычно состоит из двух шагов: дифференцирование (опционально) и сопоставление. Большинство систем распознавания речи в настоящее время используют абсолютные измерения и их первую производную. Позднее, в рассмотрение включили и вторые производные. Результатом этой стадии обработки является один вектор параметров, который объединяет все параметры.

21. Показан частотный отклик трех различных реализаций дифференциатора. На рис. (a) N_d=1. На рис. (b) N_d=3. На рис. (c) N_d=5. Обратите внимание, что идеальный дифференциатор обладает частотным откликом пропорциональным логарифму частоты. Те м не менее, желательно затухание высоких частот (важно не усиливать эти компоненты), т.к. высокочастотные компоненты слишком шумят. Следовательно, каждая реализация не усиливает высокие частоты.

22. Статистические модели в распознавании речи в общем делятся на две категории: параметрические модели (непрерывное распределение) и непараметрические модели (дискретное распределение). Тип моделей изменяется от прямой оценки модели линейного предсказания до умной вероятностной модели базирующейся на декорреляционных преобразованиях.

23. a) Эллиптическая область показывает диапазон допустимых значений пар (x,y) (предполагается, что все точки в этой области равновероятны). Какое расстояние больше: (a) расстояние от точки **a** до точки **b** или (b) расстояние от точки **c** до точки **d**? Правильный ответ (a). Так как расстояние от **a** до **b** имеет больший процент изменения в вертикальном направлении, то мы должны поверить, что это расстояние "ощутимо" больше, чем расстояние между **c** и **d**. Обратите внимание, что расстояния измеряются в одинаковых единицах измерения.
    b) Важное изменение задачи на рис. 23(a): какому распределению принадлежит точка данных? Расстояния между центром каждого распределения и точкой данных одинаковы. Тем не менее, так как формы распределений различны, по какой шкале мы должны сравнивать эти два расстояния? [72]

24. Пример преобразования речевых данных полученных с помощью телефона. Данные были оцифрованы с частотой 8 кГц. Первая кривая соответствует собственному вектору, который взвешивает амплитуды низкочастотного банка фильтров и понижает высокочастотную часть. Вторая кривая показы-

вает измерение, которое нацелено на 1 кГц область спектра. Оба этих измерения пытаются проследить первую форманту для гласных. Третья кривая представляет измерение, которое отображает высоки частоты звуков, такие как свистящие звуки. Использование таких типов преобразований по отношению к моделям специфичных классов звуков речи является областью продолжающихся исследований в распознавании речи.

25. Показан K-MEANS алгоритм для двухмерной задачи группировки. Входные векторы группируются в кластеры в соответствии с правилом ближайшего соседа. Центры этих кластеров подсчитываются на основе содержащихся в них данных. Данные затем переклассифицируются с помощью новых кластеров. Процедура повторяется до тех пор, пока не достигается приемлемое качество кластеризации. Центры кластеров затем становятся векторами словаря.

26. Показано искажение словаря как функция его размера. Размер словаря для систем распознавания речи обычно лежит между 32 и 256. Обратите внимание, что искажение словаря в лучшем случае отрицательно влияет на производительность системы распознавания речи.

X. Описания таблиц

1. Показаны два банка фильтров критического диапазона. Первый банк фильтров (столбцы 2 и 3) базируется на Bark-шкале. Второй (столбцы 4 и 5) базируется на Mel-шкале. Затененные клетки приведены для сравнения. Обычно их не включают. Для банка фильтров на Bark-шкале, деградация сигнала после телефона обрабатывается банком состоящим из 16 диапазонов (индексы 2-17).
2. Приведены все общие методы моделирования сигнала в распознавании речи. Аббревиатуры расшифровываются после таблицы.

XI. Рисунки

Рисунок 1.

Рисунок 2.

Рисунок 3.

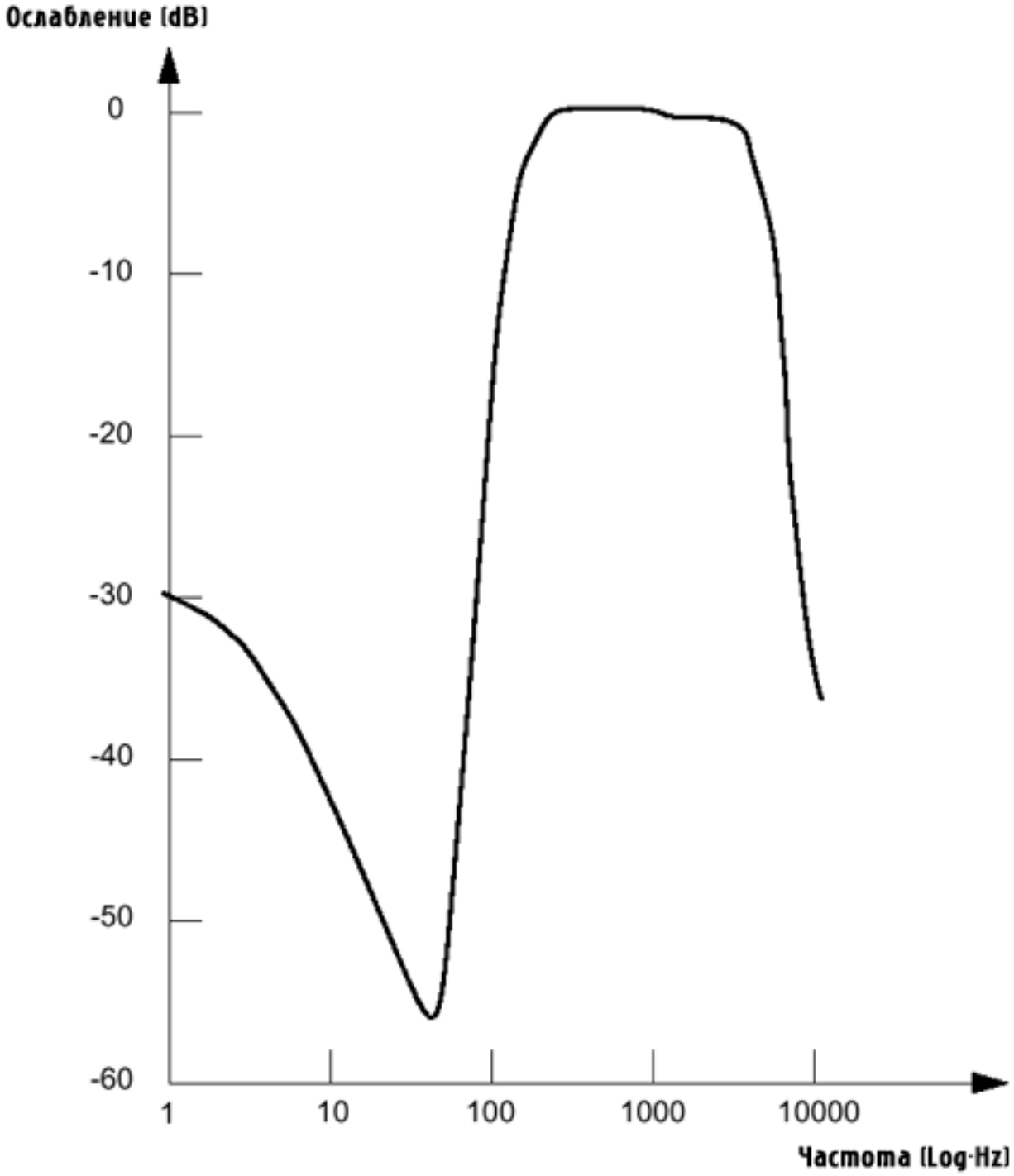

Ослабление (dB)
0
-10
-20
-30
-40
-50
-60
1
10
100
1000
10000
Частота (Log-Hz)

Рисунок 4.

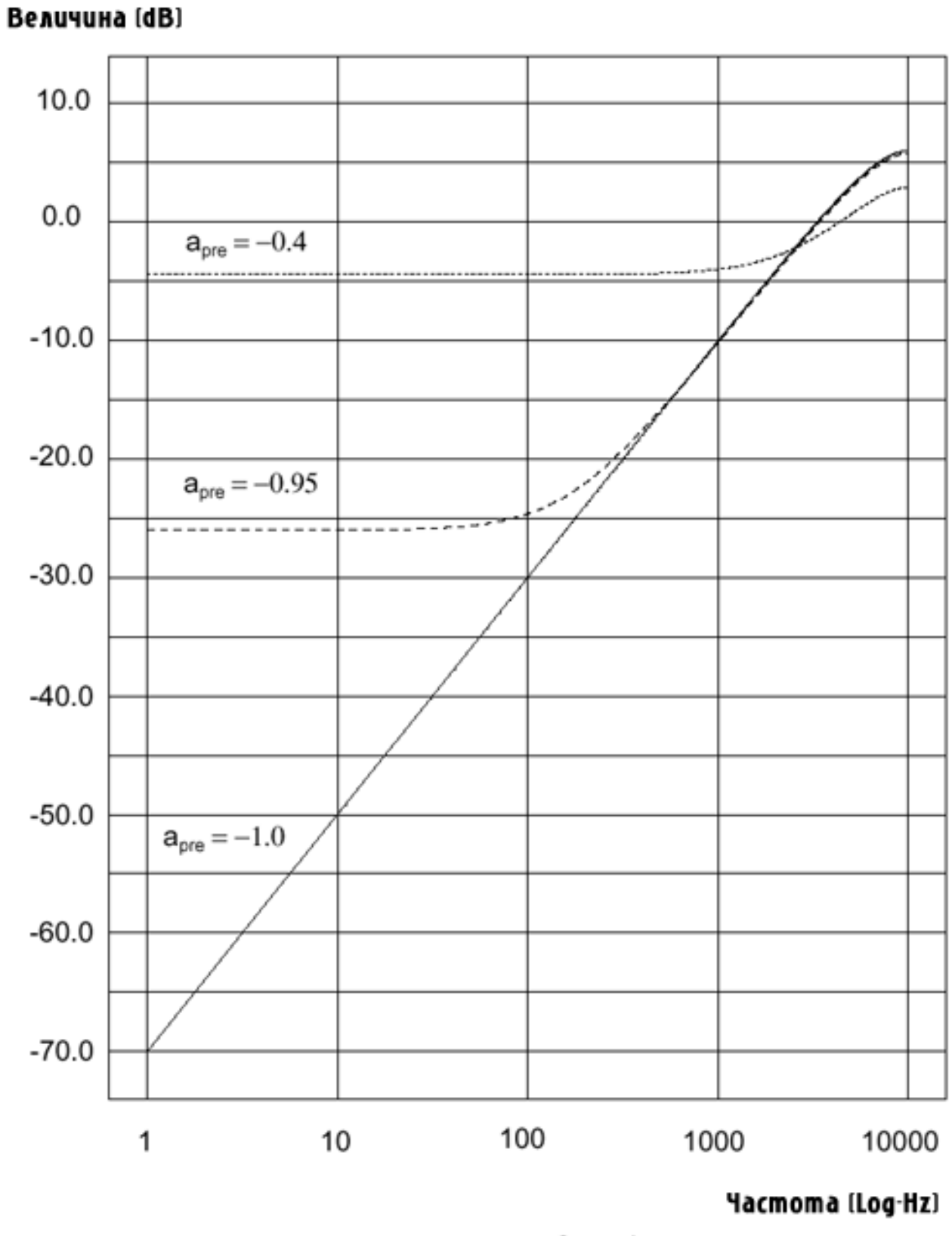

Величина [dB]
10.0
0.0
-10.0
-20.0
-30.0
-40.0
-50.0
-60.0
-70.0
$a_{pre} = -0.4$
$a_{pre} = -0.95$
$a_{pre} = -1.0$
1
10
100
1000
10000
Частота [Log-Hz]

Рисунок 5(a).

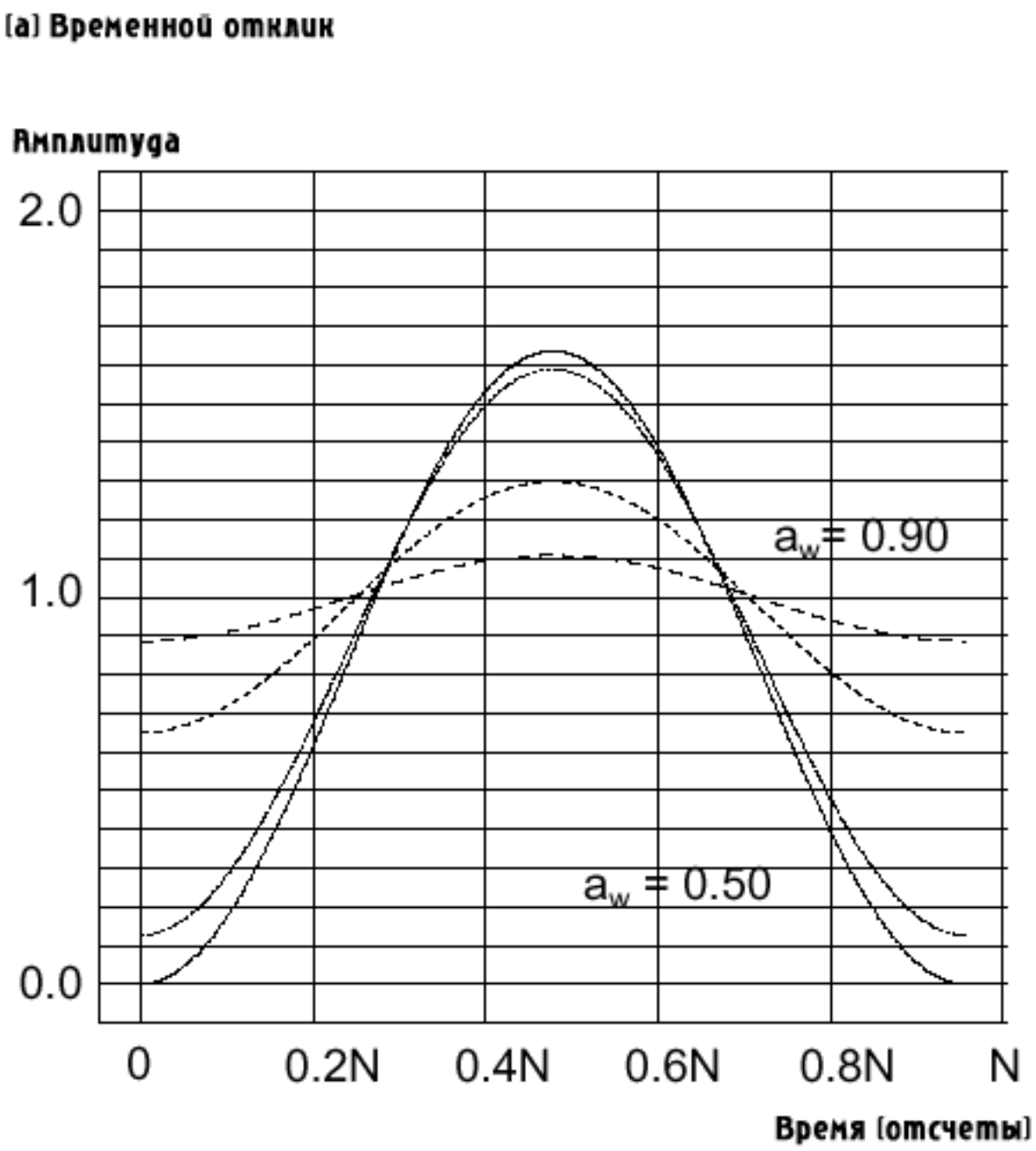

(a) Временной отклик
Амплитуда
2.0
1.0
0.0
a_w = 0.90
a_w = 0.50
0
0.2N
0.4N
0.6N
0.8N
N
Время (отсчеты)

Рисунок 5(b).

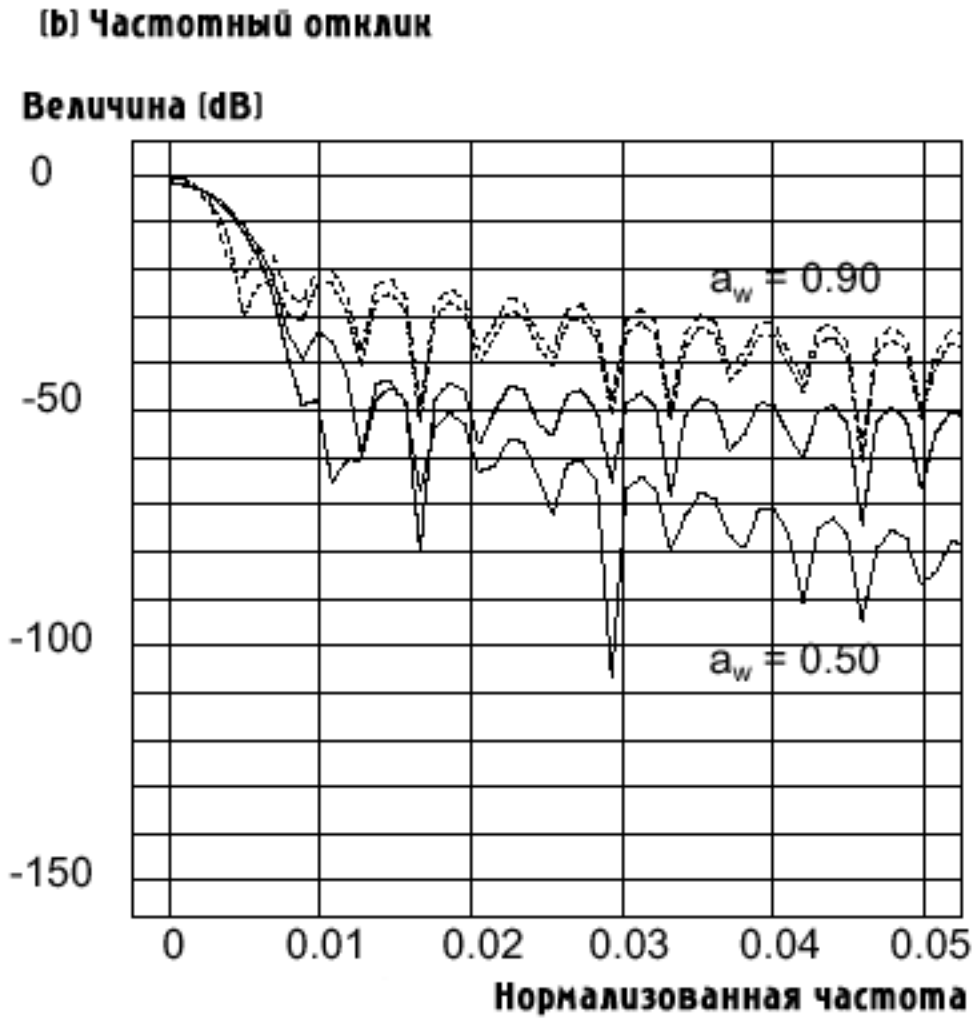

(b) Частотный отклик
Величина (dB)
0
-50
-100
-150
a_w = 0.90
a_w = 0.50
0
0.01
0.02
0.03
0.04
0.05
Нормализованная частота

Рисунок 6.

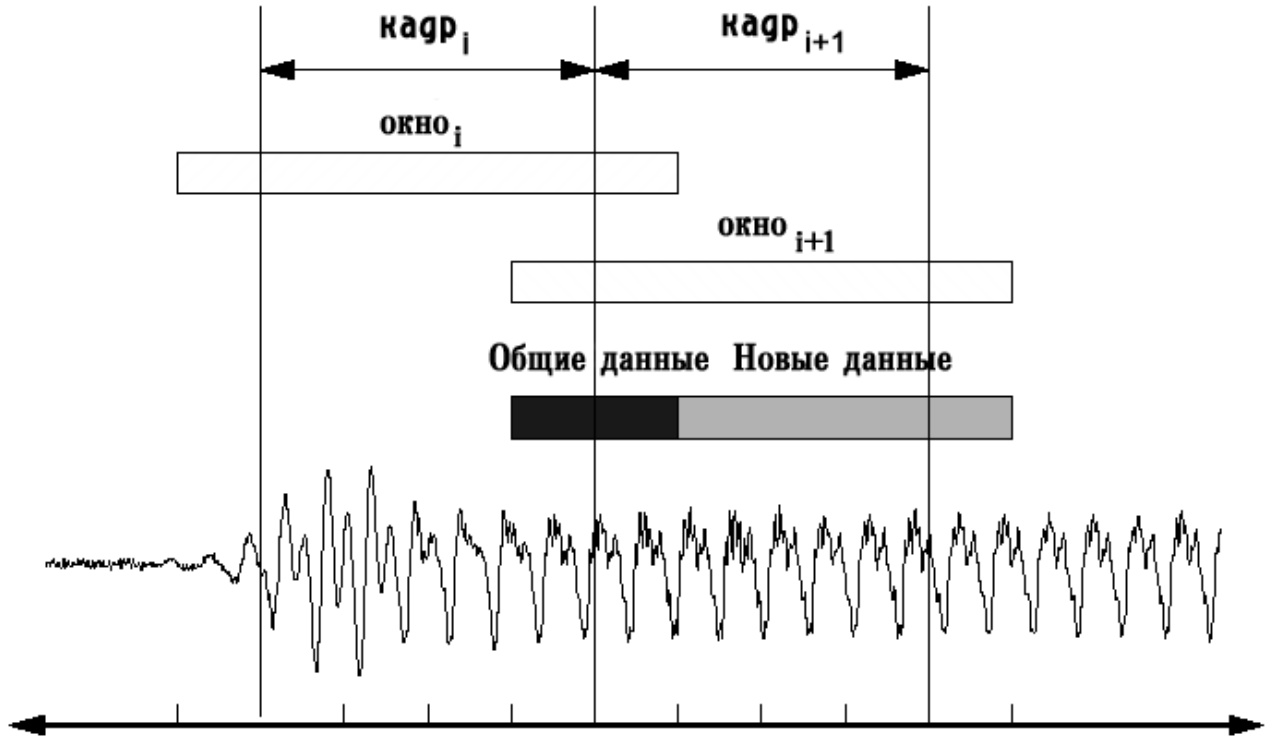


Рисунок 7.

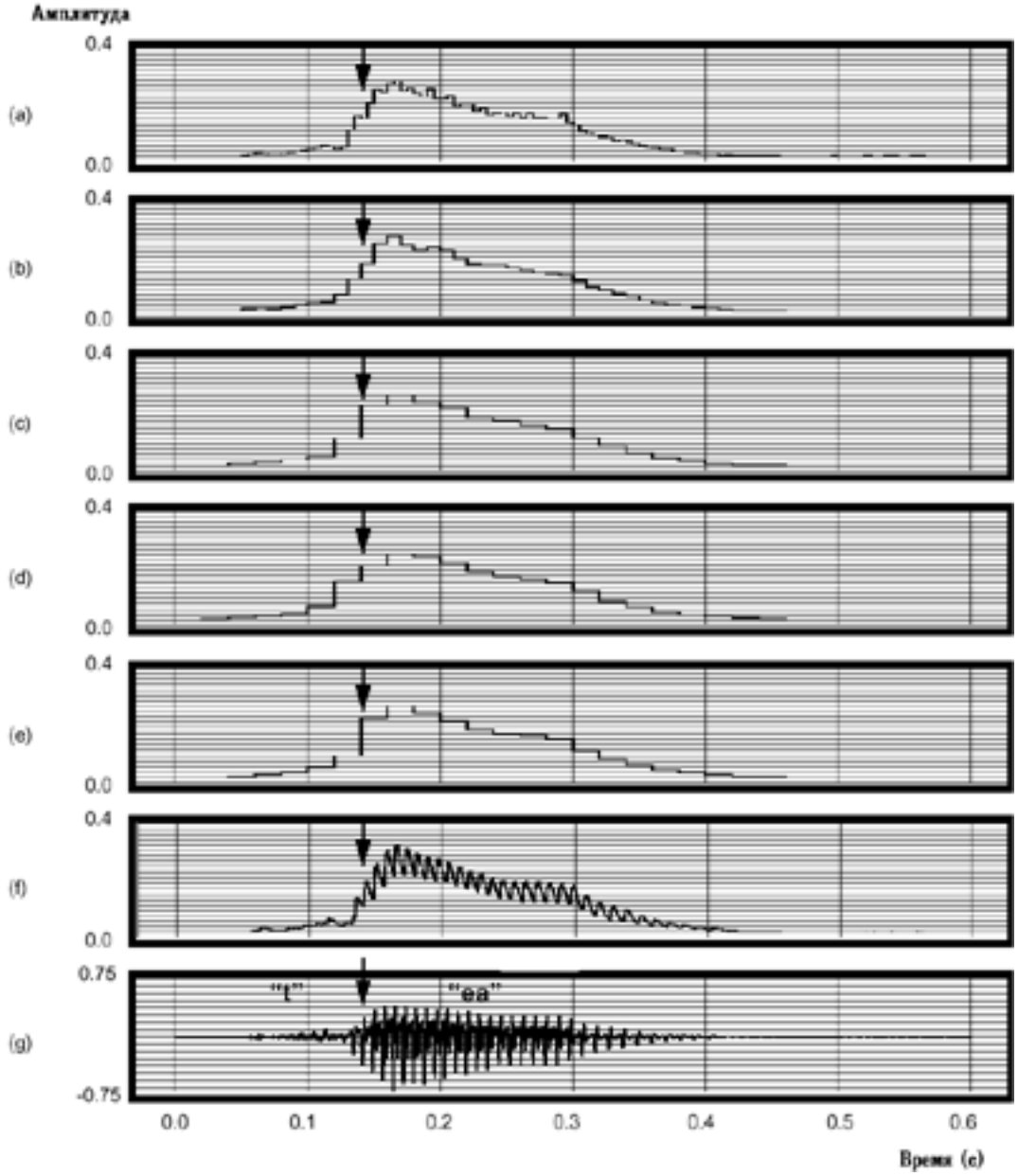

Амплитуда
0.4
(a)
0.0
0.4
(b)
0.0
0.4
(c)
0.0
0.4
(d)
0.0
0.4
(e)
0.0
0.4
(f)
0.0
0.75
"t"
"ea"
(g)
-0.75
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Время (с)

Рисунок 8.

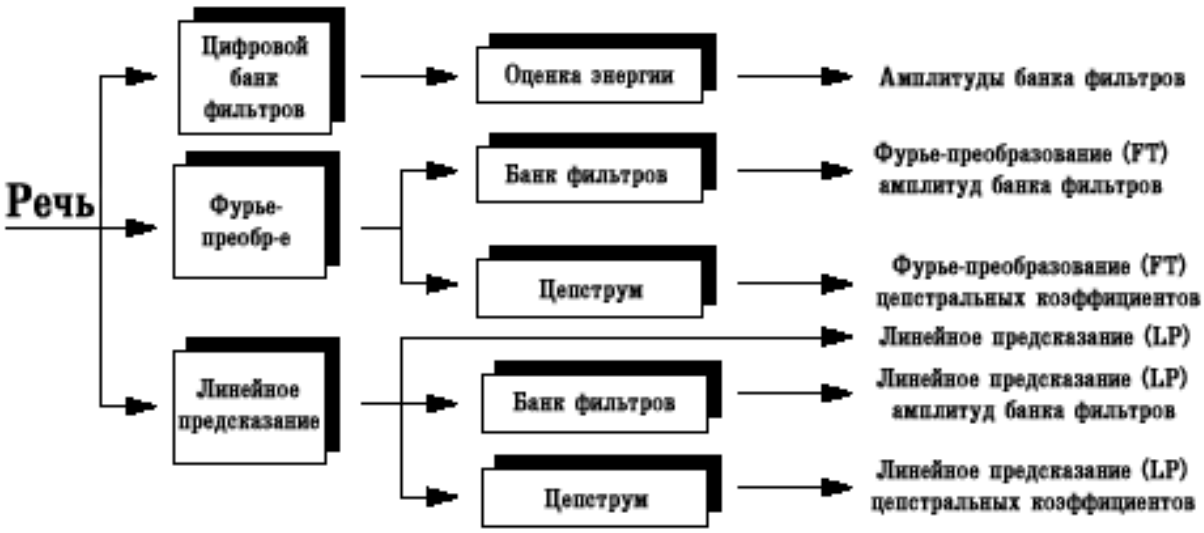

Речь
Цифровой банк фильтров
Фурье-преобр-е
Линейное предсказание
Оценка энергии
Банк фильтров
Цепструм
Банк фильтров
Цепструм
Амплитуды банка фильтров
Фурье-преобразование (FT) амплитуд банка фильтров
Фурье-преобразование (FT) цепстральных коэффициентов
Линейное предсказание (LP)
Линейное предсказание (LP) амплитуд банка фильтров
Линейное предсказание (LP) цепстральных коэффициентов

Рисунок 9.

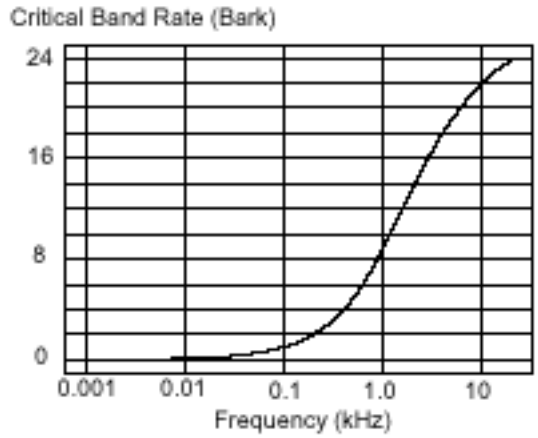


Figure 9(a). The Bark scale transformation.

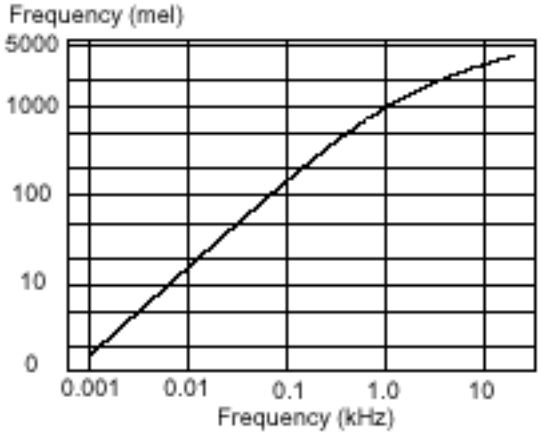


Figure 9(b). The mel scale transformation.

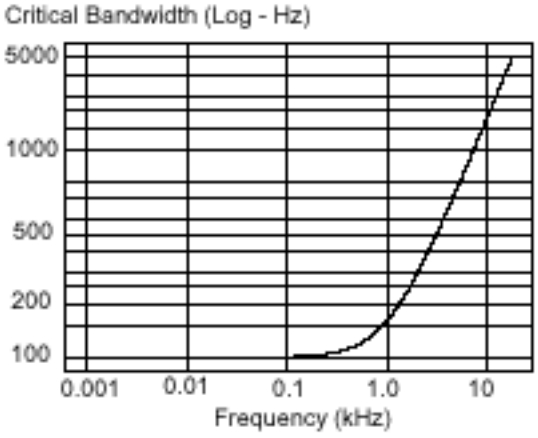


Figure 9(c). The critical bandwidth transformation.

Рисунок 10.

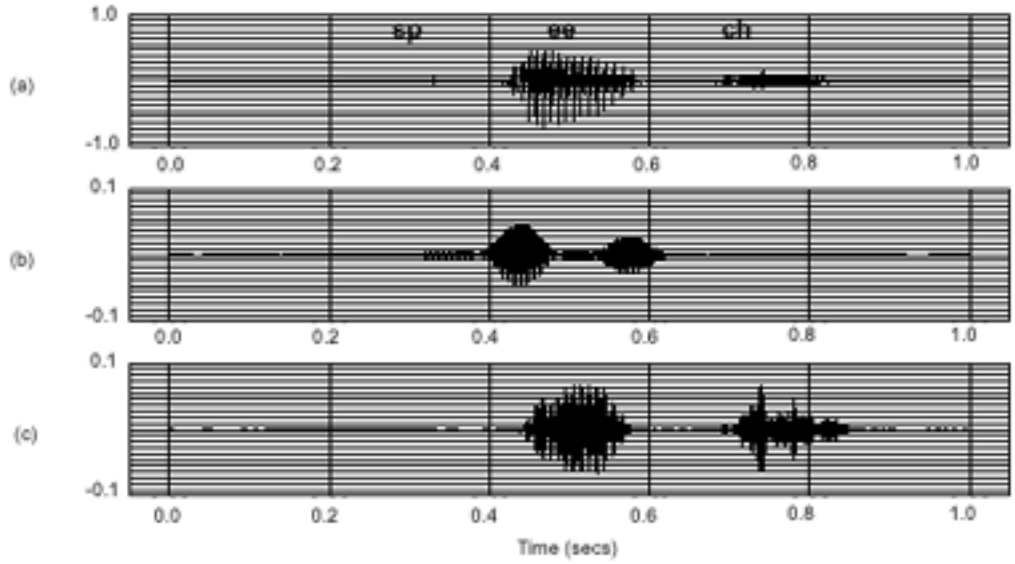


Рисунок 11.

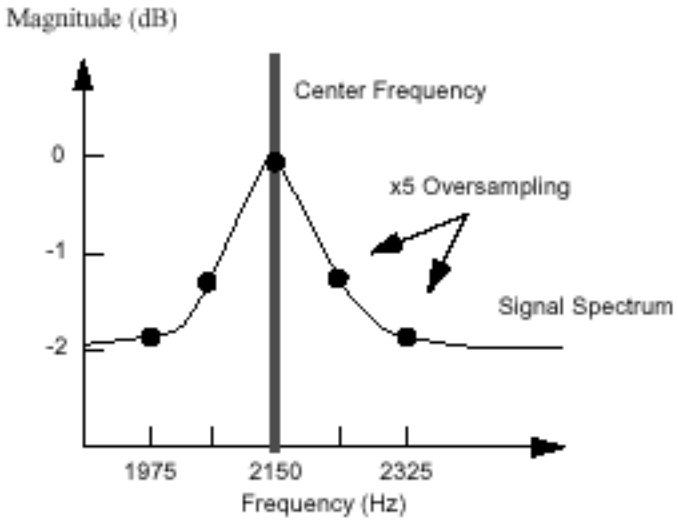


Рисунок 12.

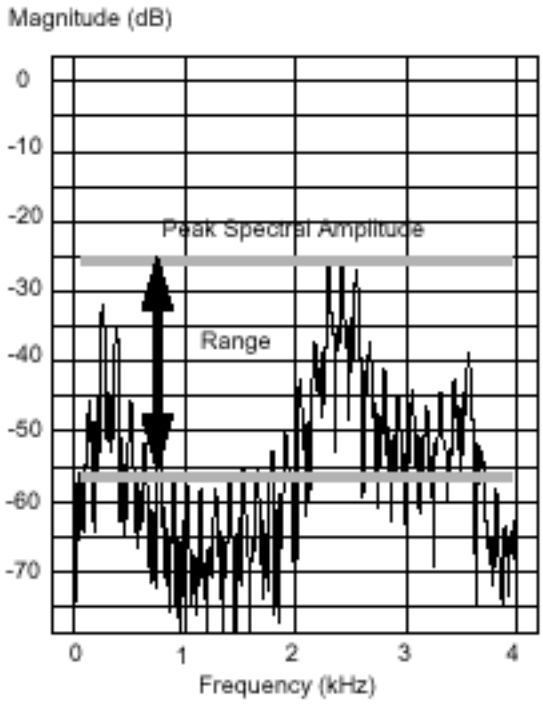


Рисунок 13.

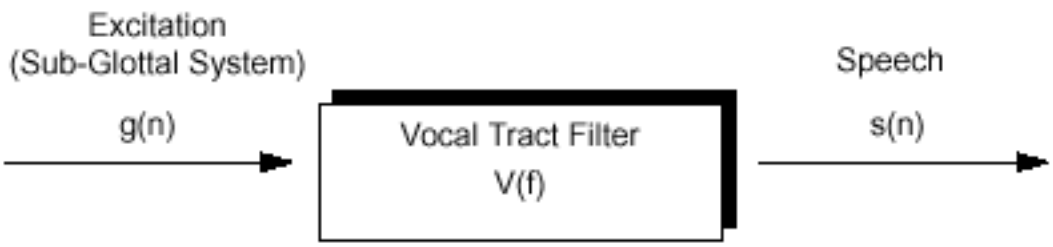


Рисунок 14.

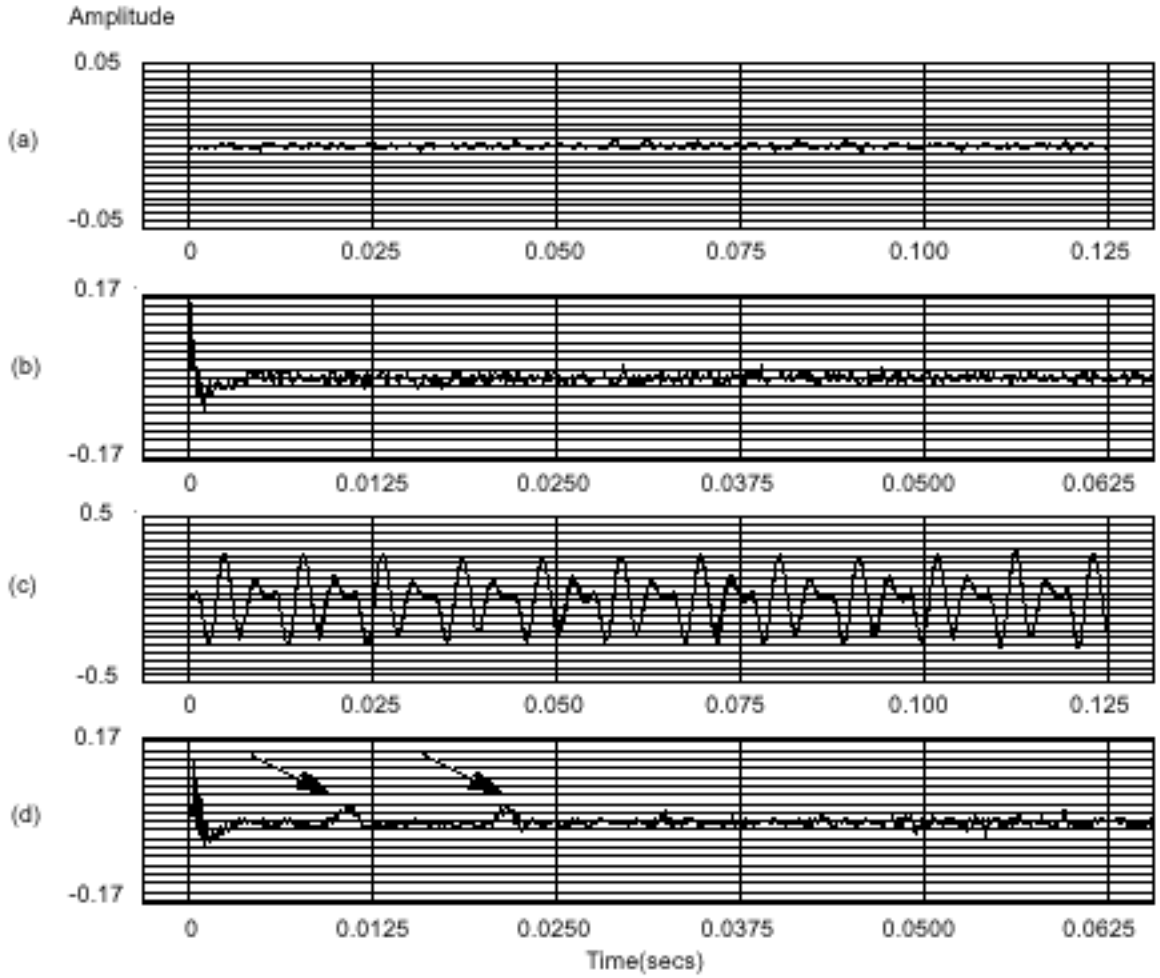

Amplitude
0.05
-0.05
(a)
0
0.025
0.050
0.075
0.100
0.125
0.17
-0.17
(b)
0
0.0125
0.0250
0.0375
0.0500
0.0625
0.5
-0.5
(c)
0
0.025
0.050
0.075
0.100
0.125
0.17
-0.17
(d)
0
0.0125
0.0250
0.0375
0.0500
0.0625
Time(secs)

Рисунок 15.

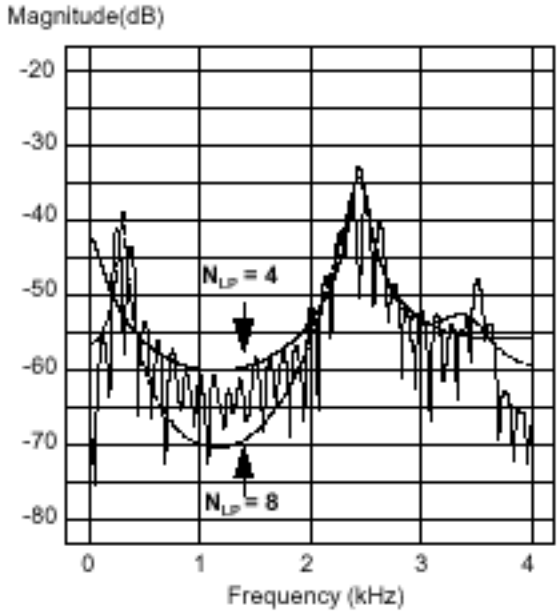


Рисунок 16

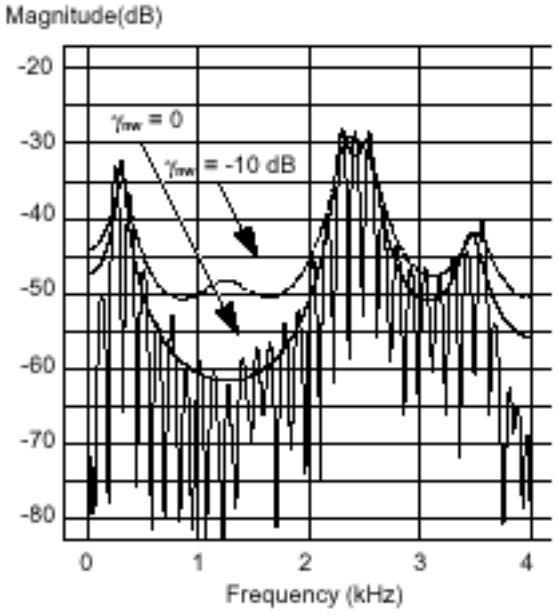
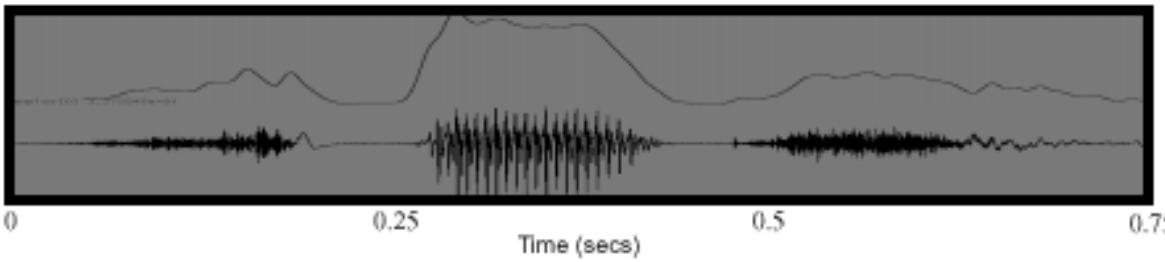
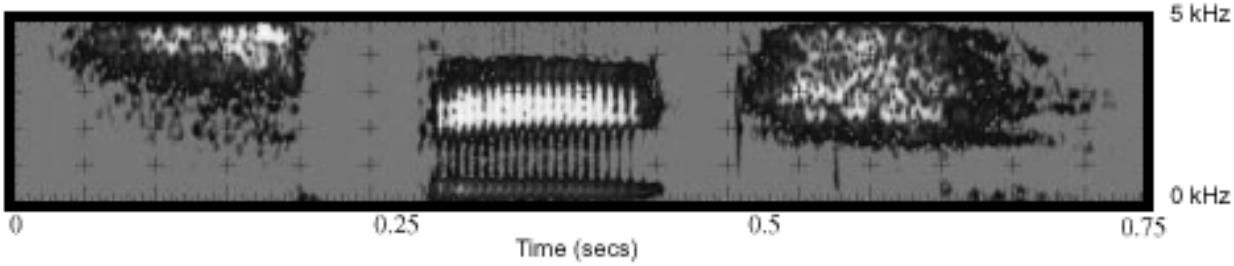


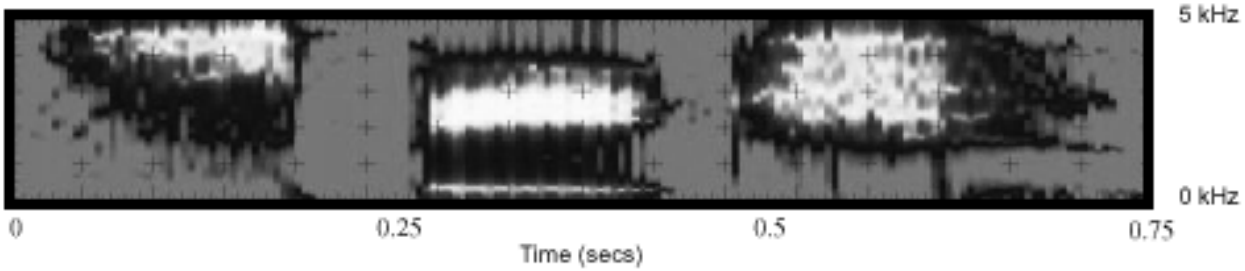
Рисунок 17.



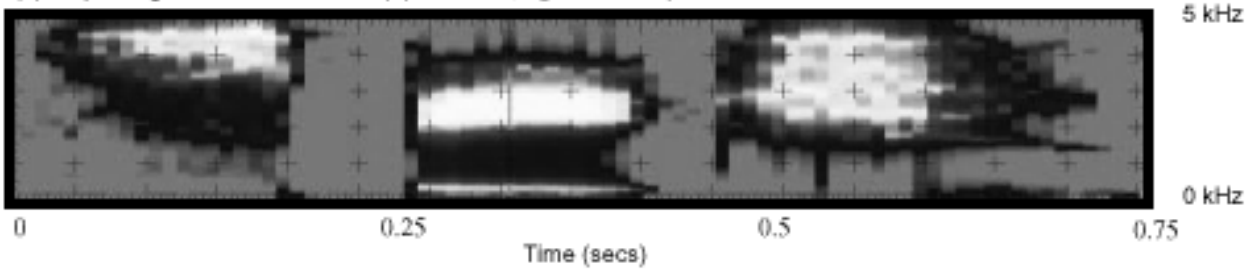
(a) A speech waveform (the word "speech") and its power contour.



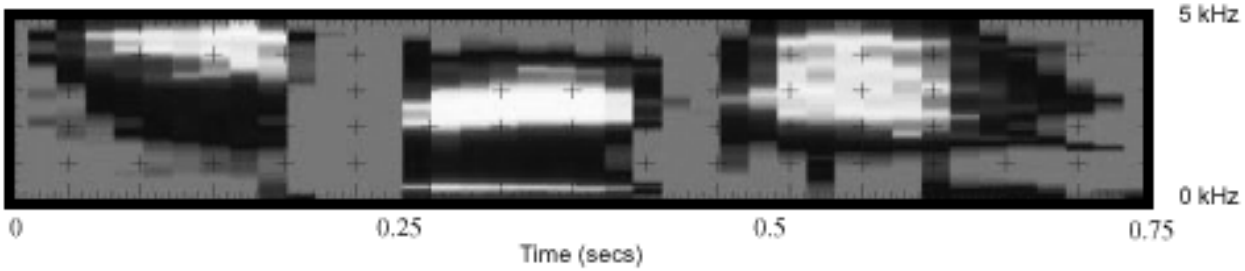
(b) A wideband spectrogram (6 msec window).



(c) A spectrogram of the LP model (T_r = 5 msec, T_w = 10 msec).



(d) A spectrogram of the LP model (T_r = 10 msec, T_w = 20 msec).



(e) A spectrogram of the LP model (T_r = 20 msec, T_w = 30 msec).

Рисунок 18.

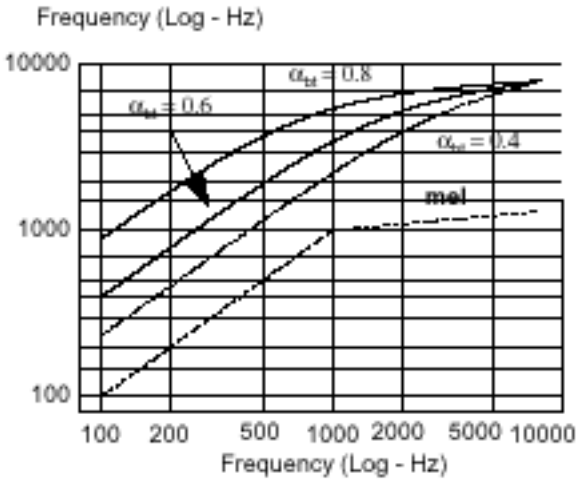


Рисунок 19.

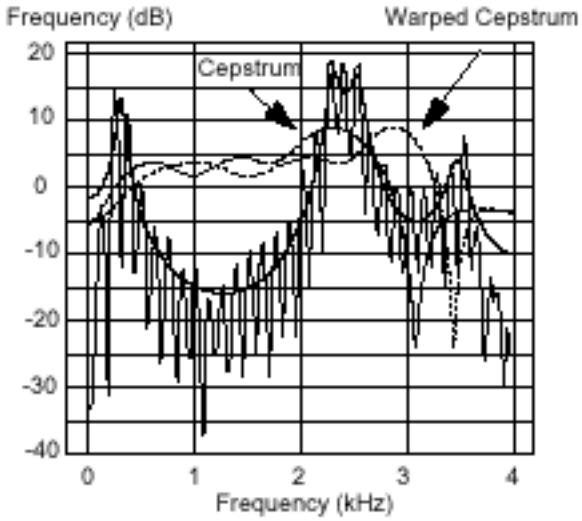


Рисунок 20.

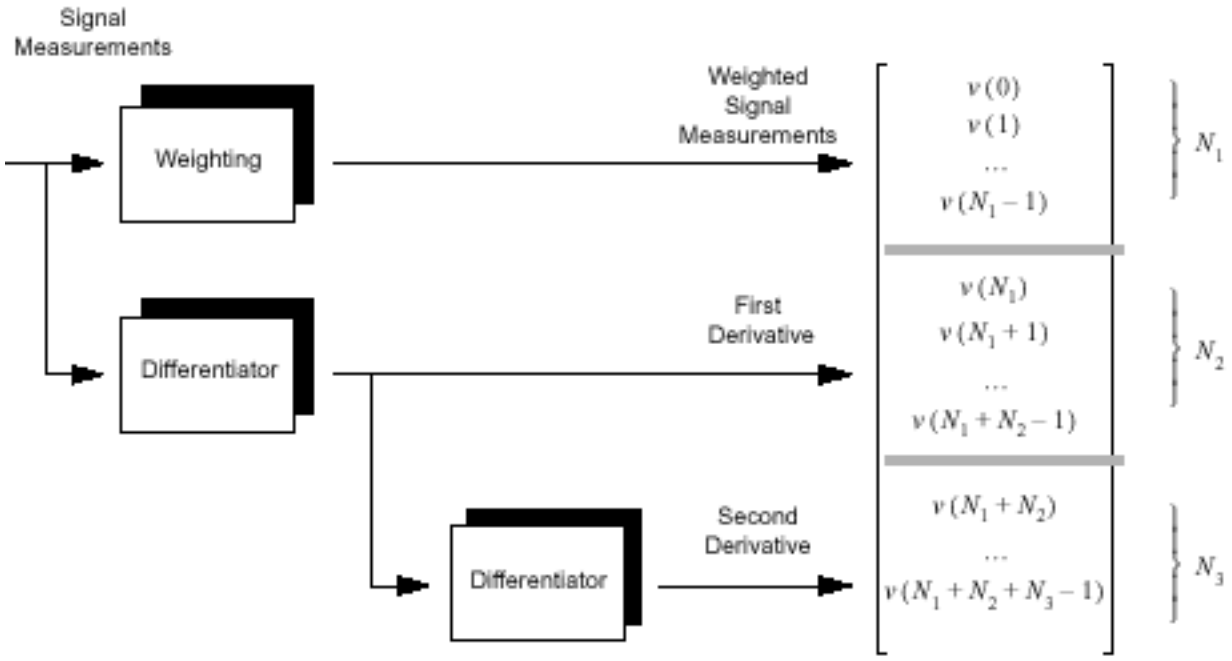


Рисунок 21.

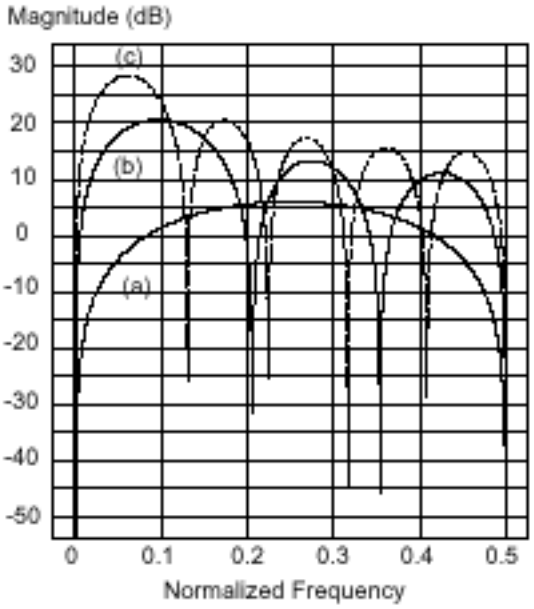


Рисунок 22.

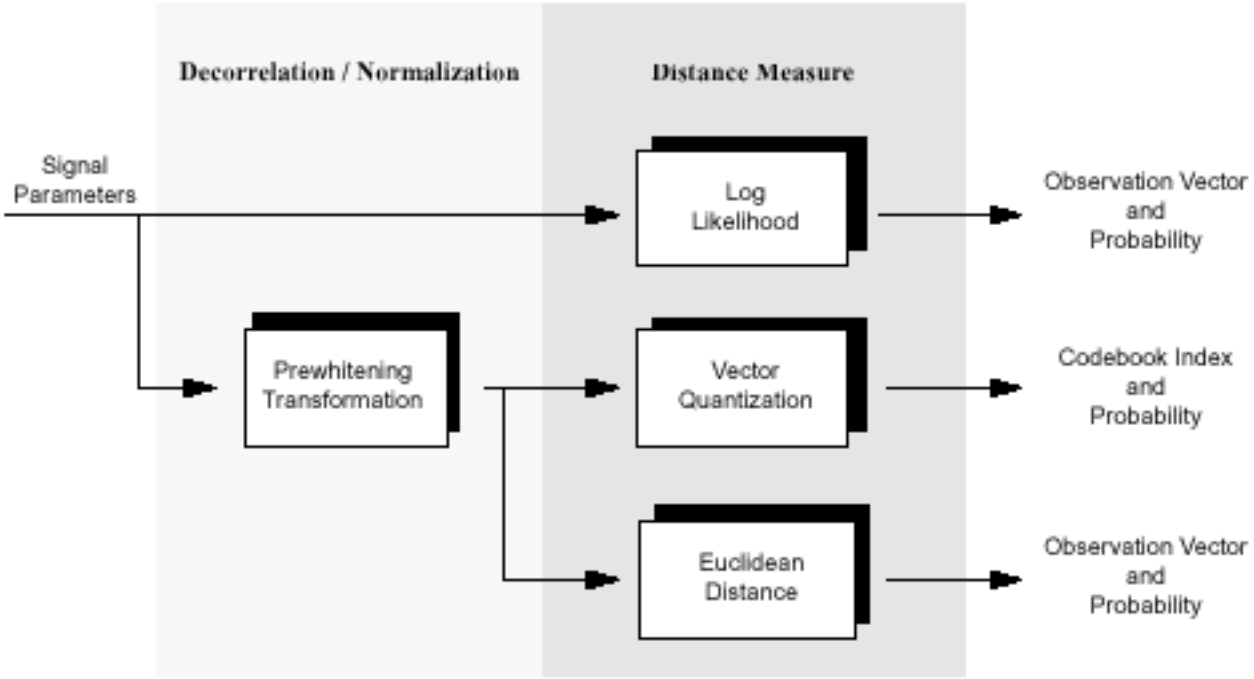

Decorrelation / Normalization
Distance Measure
Signal Parameters
Log Likelihood
Observation Vector and Probability
Prewhitening Transformation
Vector Quantization
Codebook Index and Probability
Euclidean Distance
Observation Vector and Probability

Рисунок 23.

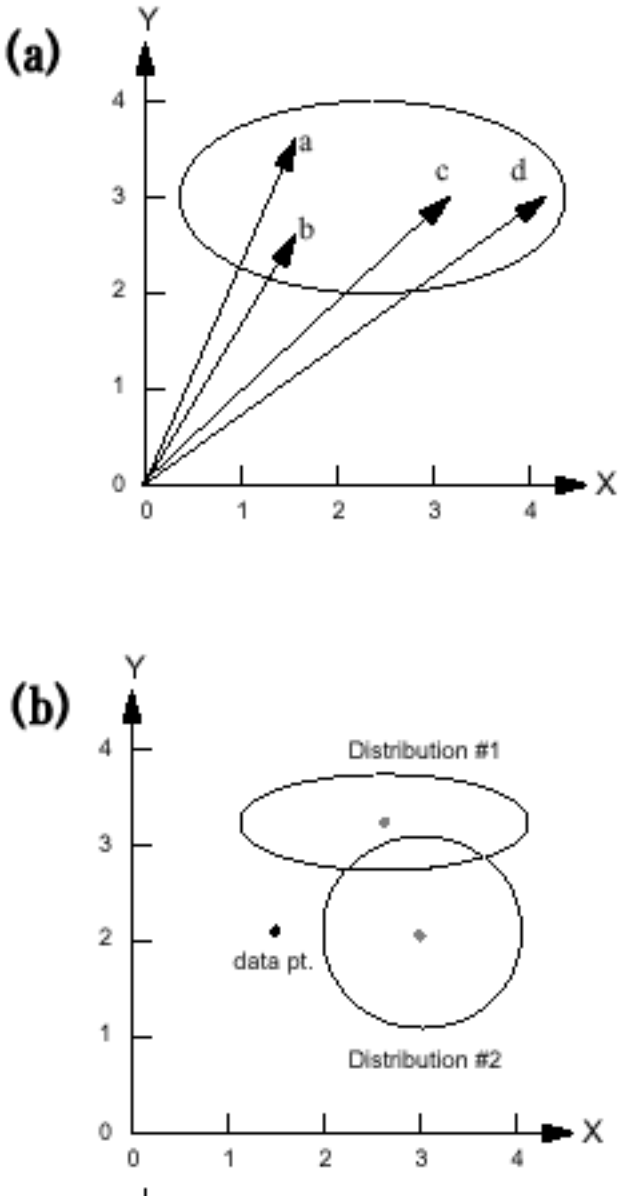

(a)
Y
X
a
b
c
d
0
1
2
3
4
(b)
Y
X
Distribution #1
data pt.
Distribution #2
0
1
2
3
4

Рисунок 24.

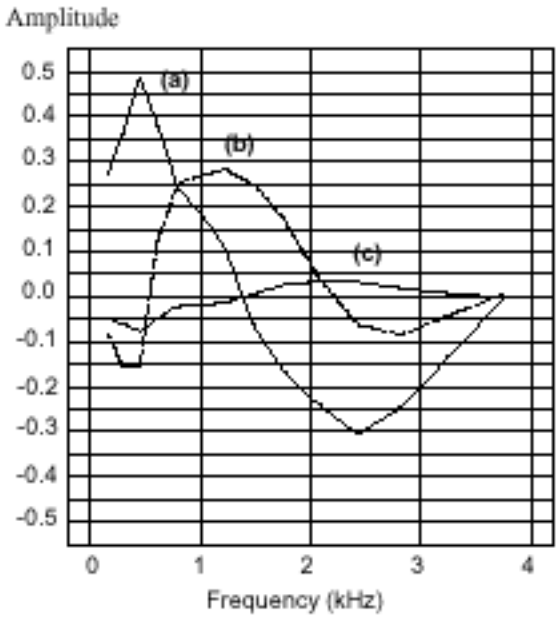


Рисунок 25.

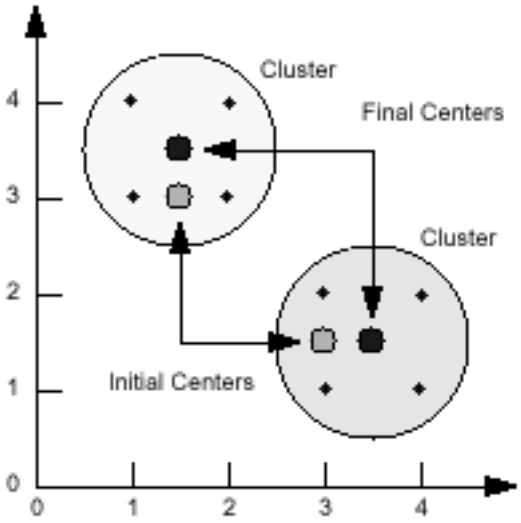
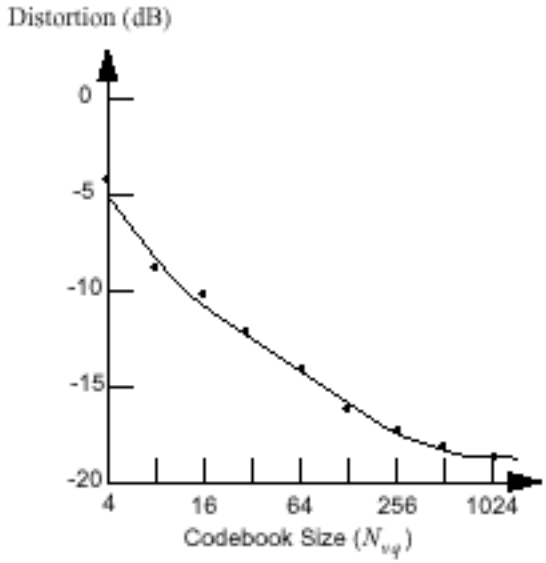


Рисунок 26.



XII. Таблицы

Таблица 1.

Индекс	Bark-шкала		Mel-шкала	
	Центр. частота, Гц	BW, Гц	Центр. частота, Гц	BW, Гц
1	50	100	100	100
2	150	100	200	100
3	250	100	300	100
4	350	100	400	100
5	450	110	500	100
6	570	120	600	100
7	700	140	700	100
8	840	150	800	100
9	1000	160	900	100
10	1170	190	1000	124
11	1370	210	1149	160
12	1600	240	1320	184
13	1850	280	1516	211
14	2150	320	1741	242
15	2500	380	2000	278
16	2900	450	2297	320
17	3400	550	2639	367
18	4000	700	3031	422
19	4800	900	3482	484
20	5800	1100	4000	556
21	7000	1300	4595	639
22	8500	1800	5278	734
23	10500	2500	6063	843
24	13500	3500	6964	969
Индекс	Bark-шкала		Mel-шкала	
	Центр. частота, Гц	BW, Гц	Центр. частота, Гц	BW, Гц
1	50	100	100	100
2	150	100	200	100
3	250	100	300	100
4	350	100	400	100
5	450	110	500	100
6	570	120	600	100
7	700	140	700	100
8	840	150	800	100
9	1000	160	900	100
10	1170	190	1000	124
11	1370	210	1149	160
12	1600	240	1320	184
13	1850	280	1516	211
14	2150	320	1741	242
15	2500	380	2000	278
16	2900	450	2297	320
17	3400	550	2639	367
18	4000	700	3031	422
19	4800	900	3482	484
20	5800	1100	4000	556
21	7000	1300	4595	639
22	8500	1800	5278	734
23	10500	2500	6063	843
24	13500	3500	6964	969

Таблица 2.

Общая информация		Измерения сигналов					Параметры сигналов	Статистическая модель	
Авторство (ссылка)	Размер/Задача приложения	f_s кГц	a_{pre}	Размер кадра, мс	Размер окна, мс	Спектральный анализ	Типы параметров	Модель	Метод распознавания речи
ATR [87]	Large Office	12	-0.97	3	21.3	LP(12)	LP Power	PWLR VQ(256)	DD-HMM
ATR [86]	Large Office	12	0.0	5	21.3	FFT(128)	Mel FB, D-FB, D-D-FB	-	TD-NN
AT&T [7]	Small Telecom	6.67	-0.95	15	45	LP(8) Cep(12)	Liftered-Cep. D-Cep. D-D-Cep. Power D-Power D-D-Power	Variance	CD-HMM
AT&T [69]	Medium Office	8	-0.95	10	30	LP(10) Cep(12)	(same)	(same)	(same)
BBN [88,89]	Large Office	20	0.0	10	20	FFT(512) Cep(14)	Mel-Cep. D-Cep. Power D-Power	VQ	DD-HMM
Brown [90]	Small Office	16	0.0	10.0	40.0	LP(12) Cep(12)	Cep. D-Cep. Power D-Power	MS-VQ (256/stage)	HMM-NN
Cambridge [91]	Large Office	10	FD	10	10	FFT(128)	FB	-	NN
CMU [66]	Large Office	16	-.097	10.0	20.0	LP(14) Cep(12)	BT Cep. D-Cep. Power D-Power	MS-VQ (256/stage)	DD-HMM
CMU [92]	Large Office	16	-0.97	10.0	20.0	FFT(256)	Mel-FB D-FB D-D-FB	MS-VQ (256/stage)	TD-NN
CSELT [93]	Medium Telecom	16	0.0	10	20	FFT(256) Cep(12)	Mel-Cep. D-Cep. Power D-Power	Variance MS-VQ	CD-HMM DD-HMM
CSELT [94]	Large Office	12	0.0	10	20	FFT(256) Cep(18)	Mel-Cep.	VQ (128)	DD-HMM
Fujitsu [95]	Large Office	16	0.0	5	32	FFT(512)	Mel-FB Power	Identity	DP
IBM [96,97]	Large Office	16	0.0	10	20	-	Auditory Model (20)	-	DD-HMM
INRS [17]	Large Office	16	0.0	10	25.6	FFT(256)	Mel-Cep. D-Cep. Power	MS-VQ (64/stage)	CD-HMM
INRS [98]	Large Office	16	0.0	10	25.6	FFT(256)	(same)	Variance	DD-HMM
KAIST [99]	Large office	10	0.0	10	25.6	LP(12) Cep(12)	Cep. D-Cep.	MS-VQ (256/stage)	DD-HMM

продолжение на следующей странице

Общая информация		Измерения сигналов					Параметры сигналов	Статистическая модель	
Авторство (ссылка)	Размер/Задача приложения	f_s кГц	a_{pre}	Размер кадра, мс	Размер окна, мс	Спектральный анализ	Типы параметров	Модель	Метод распознавания речи
LL [2,70]	Med./Lg. Office/Mil.	8	FD	10	20.0	FFT(256) Cep(14)	Mel-Cep. D-Cep	Fixed	CD-HMM
MIT [54, 100, 113]	Large Office	16	FD	5	-	FB(40)	Auditory Model	PT	CD-CFG
Mitsubishi [101]	Large Office	10	-0.95	10	25.6	FFT(256)	Mel-Cep. D-Cep Power D-Power	Variance	CD-HMM
NEC [102]	Large Office	16	0.0	5	32	FFT(512) Cep(10)	Mel-Cep.	Variance	CD-HMM
NYNEX [103]	Small Telecom	8	0.95	5	20	LP(10) Cep(10)	Cep. D-Cep. Power D-Power	PT -	HMM MLP-NN
NTT [11,104]	Large Office	12	0.0	10	30	LP(16) Cep(16)	Cep. D-Cep. D-Power	Variance	DD-FSA
NTT	Large Office	12	0.0	8	32	(same)	(same)	Variance	CD-HMM
Panasonic [106]	Large Office	10.67	0.0	9.3	18.6	PLP(8)	Cep. D-Cep. Power D-Power	Fixed	CD-HMM
Phillips [107]	Large Office	16	0.0	10	25	FFT(512)	Mel-FB D-FB D-D-FB Power D-Power D-D-Power	MS-VQ	DD-HMM
RSRE [108]	Large Office Mil.	20	0.0	10	-	FB(27)	Bark-Cep. Power D-Cep.	Fixed	CD-HMM
SRI [109,110]	Large Office	16	0.0	8	16	FFT(256)	Mel-Cep. D-Cep. Power D-Power	MS-VQ	DD-CSG
SSI [111]	Large Office	16	0.0	6.6	-	FB(20)	Bark FB Max Ampl. DM Ampl Pitch	-	CD-HMM
TI [6,14,18]	Small Telecom	8	-1.0	20	30.0	LP(10)	Mel-FB D-FB Power D-Power	PT	CD-HMM
Tohoku Univ. [112]	Large Office	16	0.0	10	-	FB(29)	Cep. D-Cep	-	LVQ2-NN
Waseda [113]	Large Office	12	0.0	10	20	LP(16)	Mel-Cep D-Cep Power D-Power	MS-VQ (256/stage)	DD-HMM

продолжение на следующей странице

A. Описание аббревиатур:

Small	Малый размер словаря (обычно распознавание цифр)
Medium	Средний размер словаря (обычно менее 5000 слов)
Large	Большой размер словаря (обычно более 5000 слов)
Office	База данных была собрана в тихой или обычной офисной обстановке (уровень шума около 70 dB)
Telecom	Телекоммуникационные данные (данные собраны через стандартные телефонные линии)
Mil	Военные приложения, включающие в себя шумную обстановку и различные стили речи
FD	Взвешивание частотной области (применяется напрямую к спектру)
LP	Линейное предсказание (порядок показан в скобках)
PLP	Мотивированное "ощущениями" линейное предсказание
FFT	Быстрое Фурье-преобразование
FB	Банк фильтров
Cep	Кэпстральные параметры
Power	Энергия сигнала (обычно в dB)
Mel	Mel-шкала параметров
Bark	Bark-шкала параметров
Liftered	Параметры лифтеринга
BT	Параметры Mel-шкалы подсчитанные с помощью билинейного преобразования
D-FB	Первая производная банка фильтров
D-D-FB	Вторая производная банка фильтров
D-Power	Первая производная энергии
D-D-Power	Вторая производная энергии
PWLR	"Ощутимо" взвешенный логарифм вероятности измерения расстояния
VQ	Векторное квантование
MS-VQ	Многоэтапное векторное квантование
PT	Предварительное преобразование
Variance	Случайно взвешенные параметры (диагональные компоненты предварительного преобразования)
Identity	Идентификационная матрица взвешенных параметров (не взвешивать)
Fixed	Зафиксированная (в основном) взвешивающая матрица
HMM	Скрытая Марковская модель
NN	Нейросеть
TD-NN	Нейросеть с временной задержкой
DP	Динамическое программирование
CD-HMM	Непрерывная скрытая Марковская модель
DD-HMM	Дискретная скрытая Марковская модель
FSA	Finite State Automaton
CFG	Context Free Grammar
CSG	Context Sensitive Grammar
LVQ2	Обучающее векторное квантование (нейросетевой подход к векторному квантованию)
MLP	Многослойная нейросеть

B. Описание категорий

Авторство	Компания/Университет, которые провели соответствующее исследование.
Приложение	Краткое описание типа базы данных, описанной в соответствующей публикации.
Измерение сигнала	Частота оцифровки, взвешивание, размеры кадра и окна спектрального анализа. Для некоторых систем, происходит взвешивание определенного частотного диапазона. Под спектральным анализом показана последовательность операций. Где возможно показаны порядки анализа, см. в круглых скобках).
Параметры сигнала	Параметры сигнала используемые в системе. Берутся из производных параметров спектрального анализа.
Статистическая модель	Статистическая модель применяющаяся в системе распознавания речи.