ФГБОУ ВО
«Поволжский государственный университет телекоммуникаций и информатики»

Раздел 5
Технологии цифрового кодирования речи для передачи по сетям связи

Лектор :

проф. кафедры ССС ПГУТИ,

д.т.н. Гребешков А.Ю.

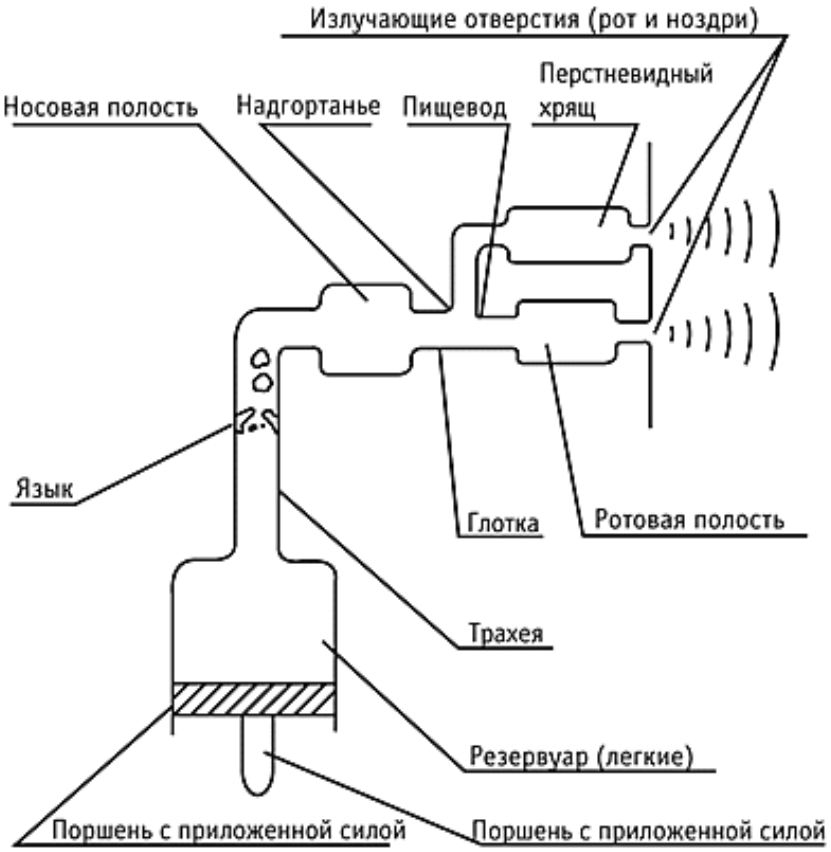
Самара
2021 год

ПГУТИ

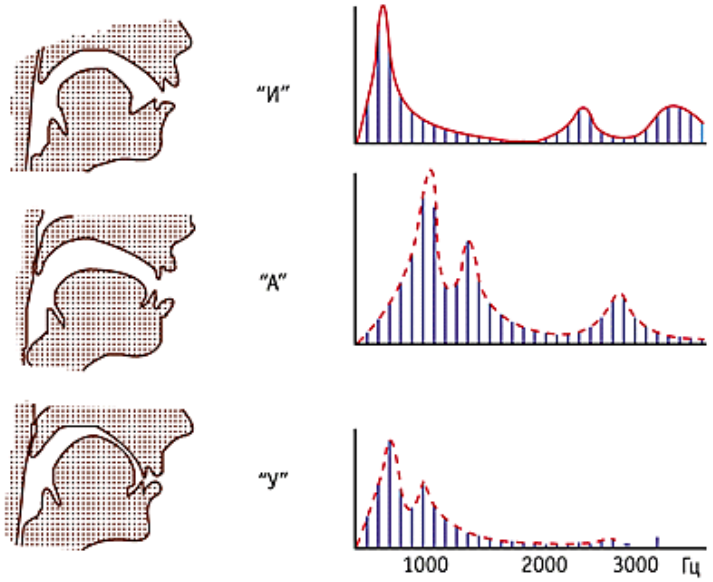
ССС
СЕТИ И СИСТЕМЫ СВЯЗИ

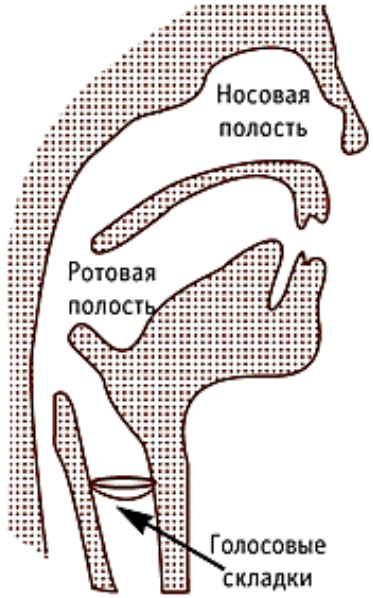
5.1 Формирование речи и её описание

Модель речевого тракта

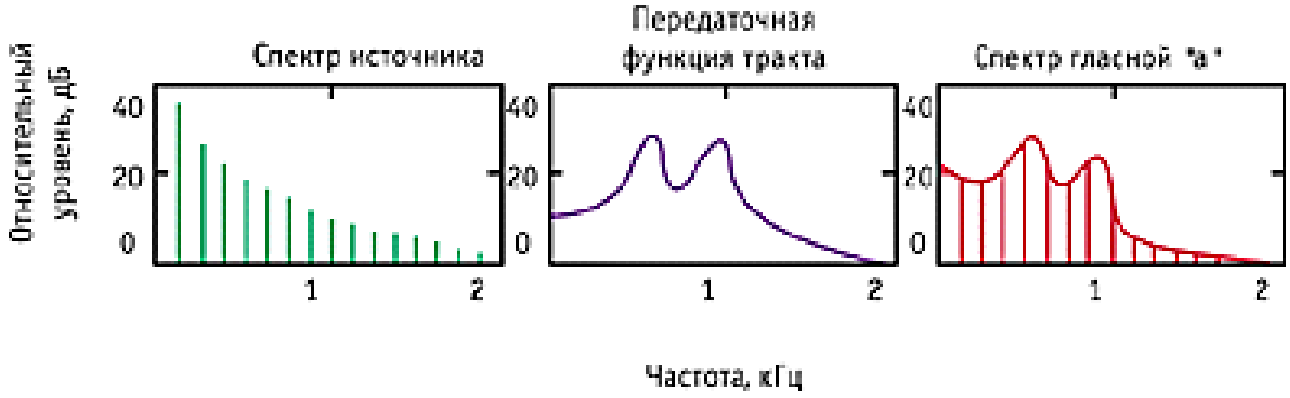


Формирование речевого сигнала





Форманты (резонансное усиление частот)



Формантный состав языка

Под формантами понимаются частотные резонансы (полюса передаточной функции) речевой акустической системы.

Параметры формант (частота, ширина, уровень) опередеяются акустическими свойствами системы. Наиболее важный параметр - частота форманты, тесно связан с геометрической конфигурацией речевого тракта

Форманты	Частотный диапазон формант, Гц		Ширина формант, Гц	Басы	Баритоны	Тенора
	Мужской	Женский				
F1	200-800	250-1000	40-70	380–450	450–540	540–640
F2	600-2800	700-3300	50-90	760–1100	1100	1300
F3	1300-3400	1500-4000	60-180	2100–2500	2500	2500–3000

Модель вокального тракта

Простейшей моделью вокального тракта можно считать цилиндрическую трубу длиной 17 см, закрытую на одном конце.

Собственные моды (формы) колебаний такой трубы и их частоты определяются из соотношений: $L = \lambda/4$; $L = 3\lambda/4$; $L = 5\lambda/4$ и т.д., таким образом частоты равны

$$f_n = (2n-1)c/4L,$$

где
n-целое число;
L-длина трубы;
c-скорость звука в воздухе

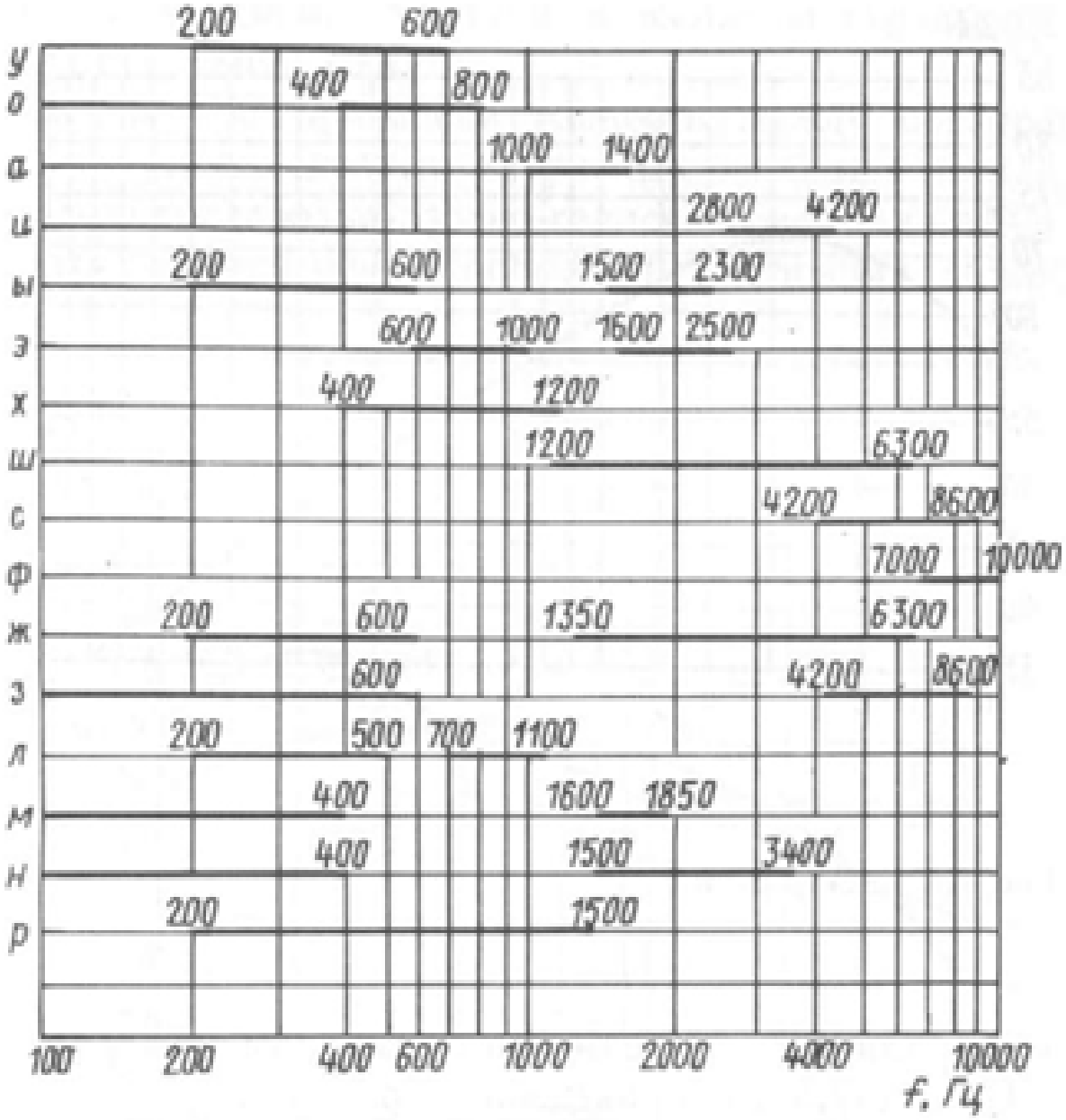
Процессу образования звуков речи с помощью фонации в терминах передаточных функций, может быть описан следующим образом:

$P(\omega)=S(\omega)T(\omega)R(\omega)$,
где
$S(\omega)$ – передаточная функция входного сигнала,
$T(\omega)$ – передаточная функция тракта,
$R(\omega)$ – активная составляющая сопротивления излучения.

Под передаточной функцией тракта понимается отношение комплексных амплитуд
объемной скорости на губах U_0 к объемной скорости у голосовой щели $U_г$:
$T(\omega)=U_0/U_г$.

Для цилиндрической трубы с одним закрытым концом она вычисляется по формуле: $T(\omega)=1/\cos(2\pi fLc)$.

Формантный состав русского языка



Восприятие музыки органами слуха

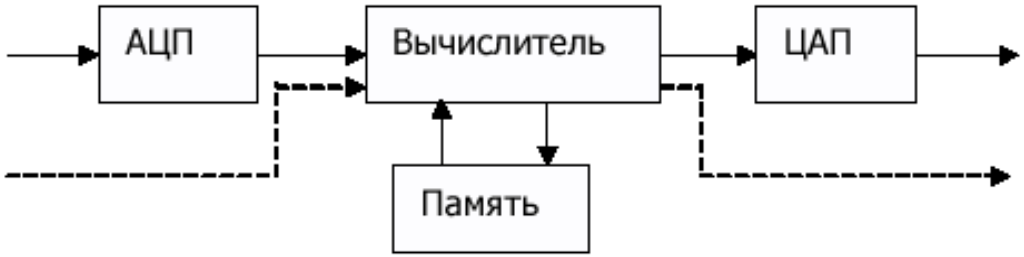
Диапазон	Описание	Эффект слышимости
Суббас 16–80 Гц	Ощущение мощи звука, чувствуется больше чем слышится	**Даёт ощущение насыщенности и глубины звука**
Верхние басы, басы 80–250 Гц	Основные ноты ритм–секции (танцевальной музыки)	**Самые низкие ноты таки инструментов, как гитара. Дает ощущение силы звука.**
Нижняя середина, диапазон средних частот 250–2000 Гц	Нижние гармоники музыкальных инструментов	**Чрезмерное усиление полосы 500-1000Гц дает эффект «медных духовых» инструментов, усиление полосы 1000–2000 Гц делает звучание «жестяным», что утомляет слух.**
Верхняя середина, диапазон средних частот 2000 – 4000 Гц	Распознавание звуков речи, таких как «м», «б», «в». Подъем в области 2,5–3кГц дает тембру звонкость.	**При сильном усилении данной полосы речь становится шепелявой, неразличимые фонемы вокала на фоне инструментов. Излишнее усиление 3 КГц утомляет слух.**
Присутствие, 4000–6000 Гц	Чистота и определенность звучания голосов и инструментов	**При подъеме данного участки музыка становится субъективно «ближе» к слушателю. Убавление полосы 5 кГц делает звучание более удаленным и прозрачным.**
Бриллианс, 6000 – 16 000 Гц	**Управляет блеском и прозрачностью звука**	**Слишком большой подъем этой области может преувеличить шипящие фонемы у певцов**

ПГУТИ

СЕТИ И СИСТЕМЫ СВЯЗИ

5.2 Цифровое кодирование речи

Схема цифрового кодирования речи



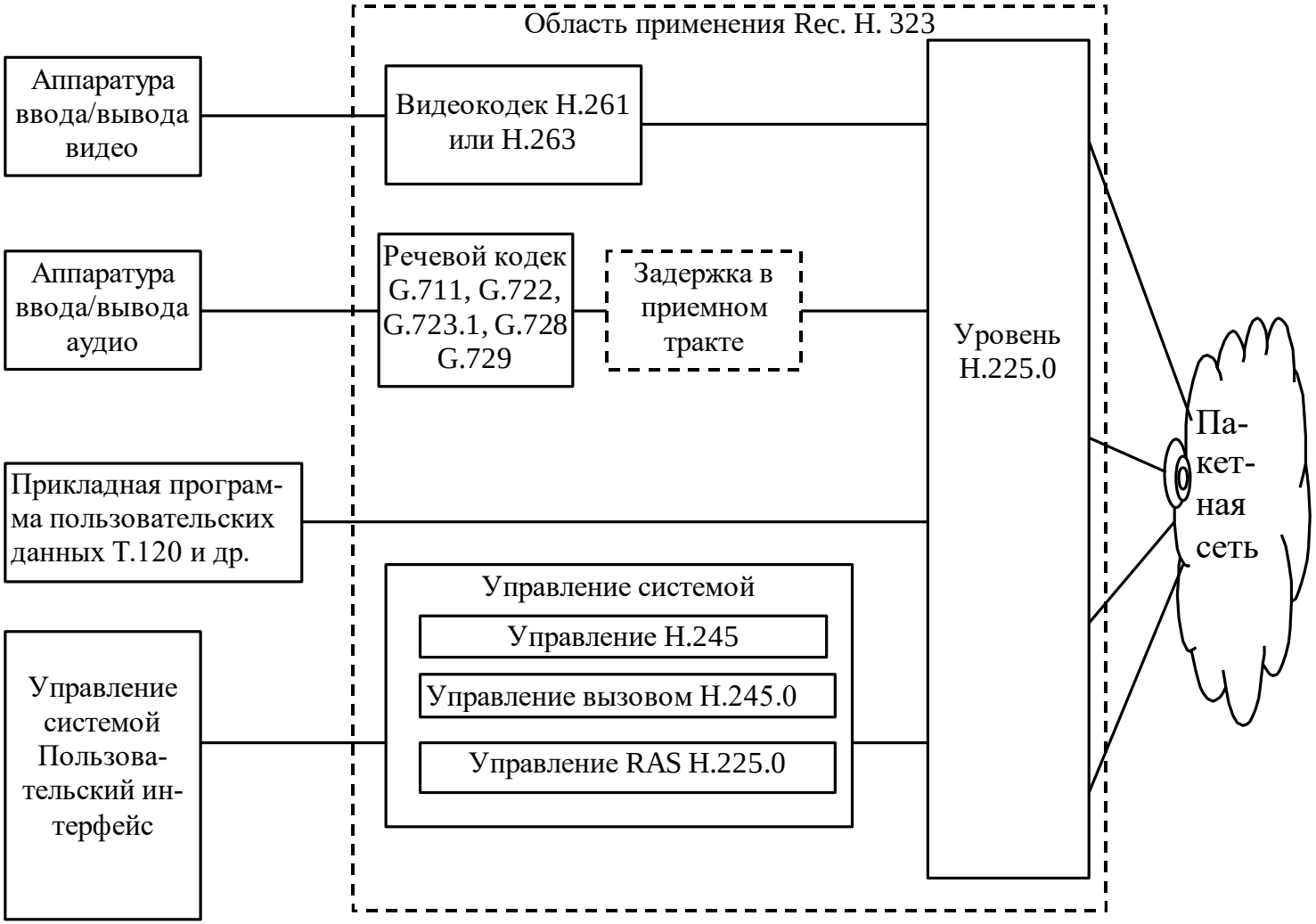
АЦП – аналогового цифровой преобразователь
ЦАП –цифро-аналоговый преобразователь

Mode	Signal bandwidth (Hz)	Sampling rate (kHz)	Bit-rate (kb/s)	Examples
Narrowband (NB)	300–3400	8	2.4–64	G.711, G.729, G.723.1, AMR, LPC-10
Wideband (WB)	50–7000	16	6.6–96	G.711.1, G.722, G.722.1, G.722.2
Super-wideband (SWB)	50–14000	32	24–48	G.722.1 (Annex C)
Fullband (FB)	20–20000	48	32–128	G.719

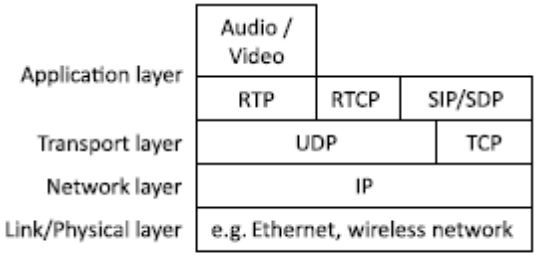
Типовые кодеры речи

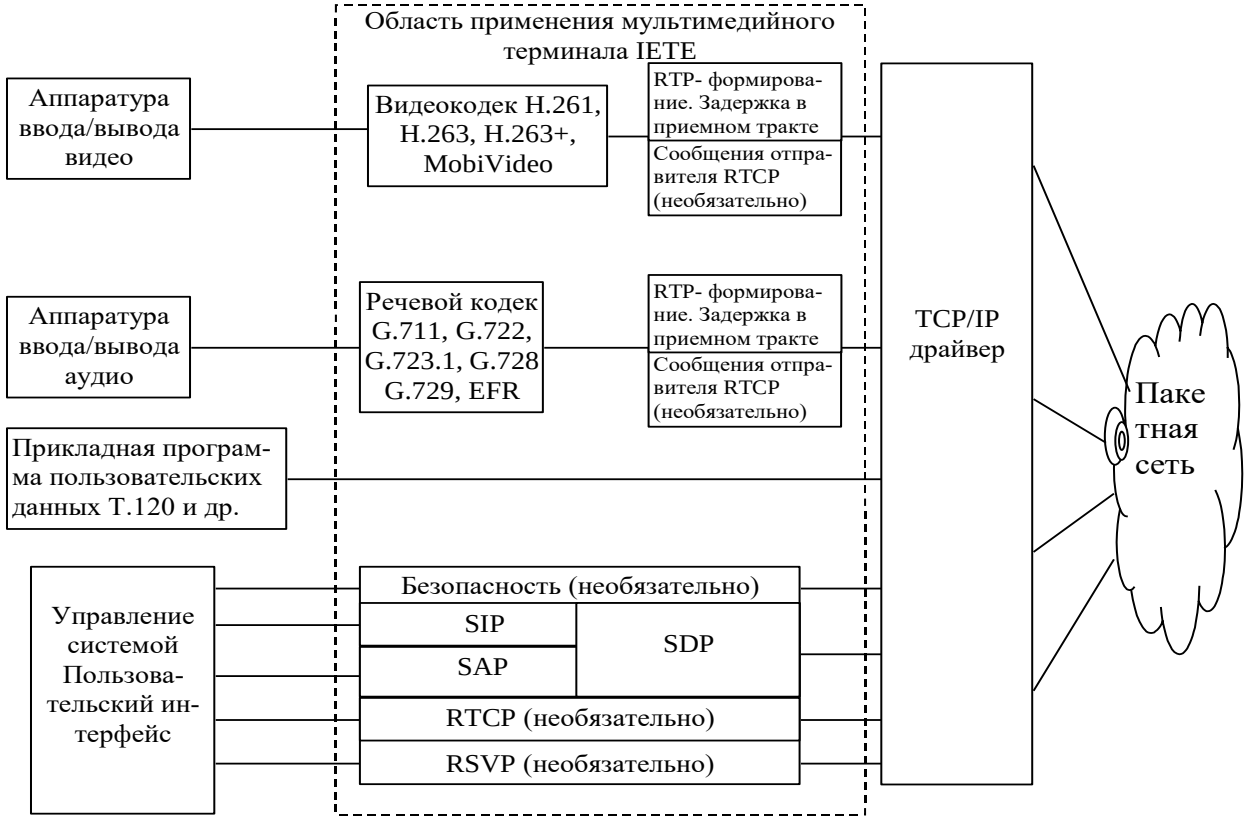
Year Finalized	Standard Name	Bit-Rate (kbps)	Applications
1972[a]	ITU-T G.711 PCM	64	General purpose
1984[b]	FS 1015 LPC	2.4	Secure communication
1987[b]	ETSI GSM 6.10 RPE-LTP	13	Digital mobile radio
1990[c]	ITU-T G.726 ADPCM	16, 24, 32, 40	General purpose
1990[b]	TIA IS54 VSELP	7.95	North American TDMA digital cellular telephony
1990[c]	ETSI GSM 6.20 VSELP	5.6	GSM cellular system
1990[c]	RCR STD-27B VSELP	6.7	Japanese cellular system
1991[b]	FS1016 CELP	4.8	Secure communication
1992[b]	ITU-T G.728 LD-CELP	16	General purpose
1993[b]	TIA IS96 VBR-CELP	8.5, 4, 2, 0.8	North American CDMA digital cellular telephony
1995[a]	ITU-T G.723.1 MP-MLQ / ACELP	5.3, 6.3	Multimedia communications, videophones
1995[b]	ITU-T G.729 CS-ACELP	8	General purpose
1996[a]	ETSI GSM EFR ACELP	12.2	General purpose
1996[a]	TIA IS641 ACELP	7.4	North American TDMA digital cellular telephony
1997[b]	FS MELP	2.4	Secure communication
1999[a]	ETSI AMR-ACELP	12.2, 10.2, 7.95, 7.40, 6.70, 5.90, 5.15, 4.75	General purpose telecommunication

Кодирование речи в сетях IP-телефонии

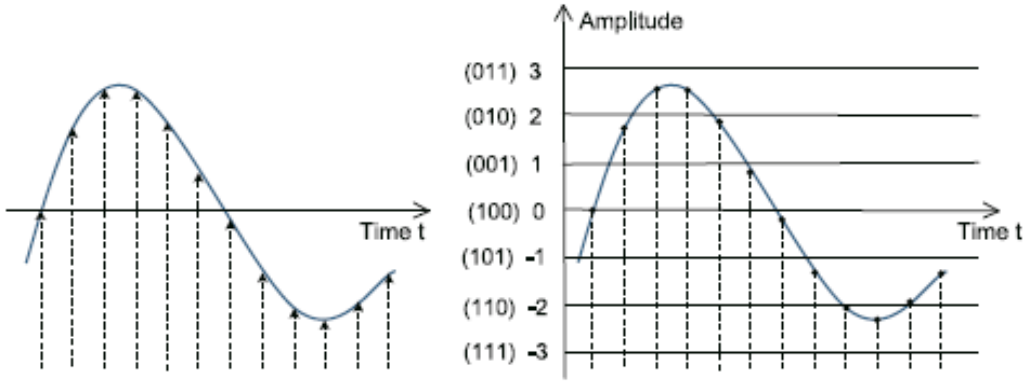


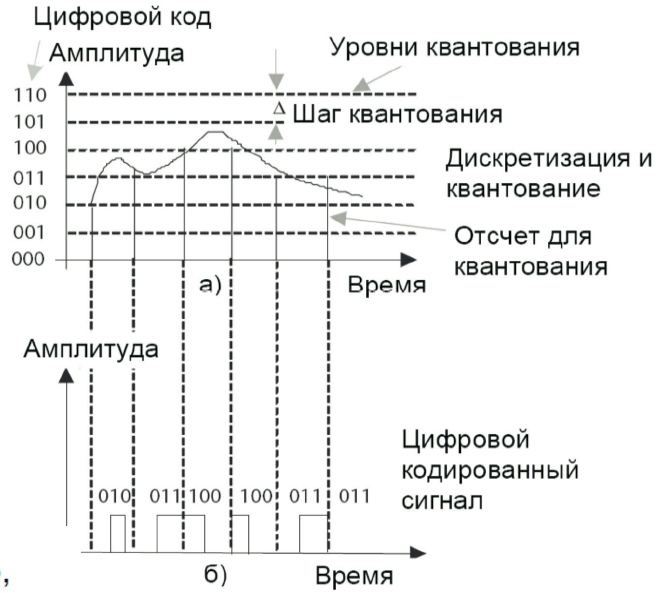
Кодирование речи в сетях IP/IETF





Дискретизация и квантование сигнала





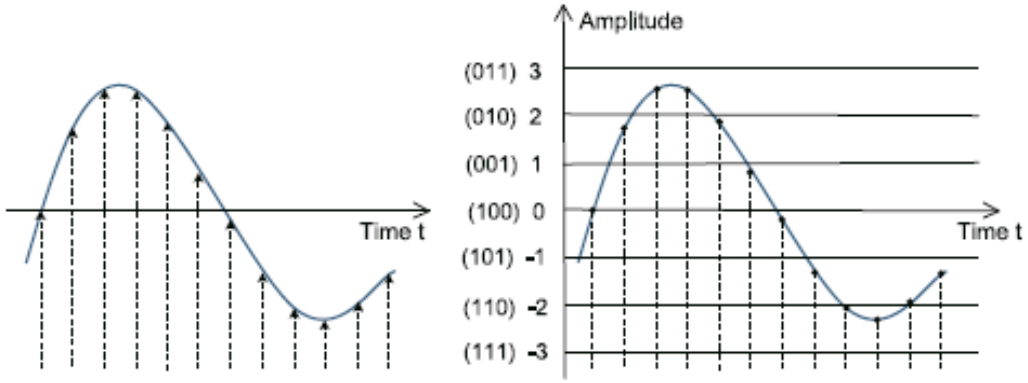
Если исходный аналоговый сигнал *x(t)* имеет ограниченный спектр, то этот сигнал может быть восстановлен по своим дискретным отсчетам, взятым с частотой, более удвоенной максимальной частоты спектра F_{max} – верхней частоты спектра исходного аналогового сигнала:

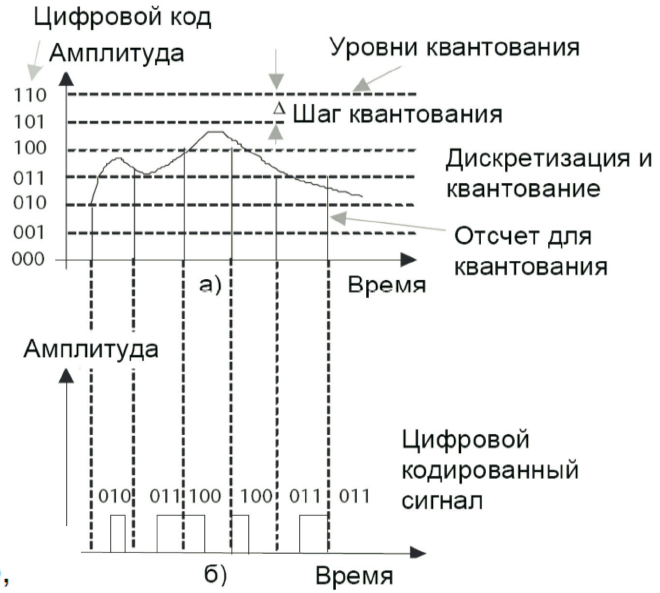
$$f_{ДИСКР} \geq 2 \times \mathrm{F}_{\max}$$

Тогда период дискретизации $T_{ДИСКР}$ аналогового сигнала, т.е. периоды времени, через которые формируются дискретные отсчеты, рассчитывается по формуле:

$$T_{ДИСКР} \leq \frac{1}{2\mathrm{F}_{\max}}$$

Дискретизация и квантование сигнала





Если исходный аналоговый сигнал *x(t)* имеет ограниченный спектр, то этот сигнал может быть восстановлен по своим дискретным отсчетам, взятым с частотой, более удвоенной максимальной частоты спектра F_{max} – верхней частоты спектра исходного аналогового сигнала:

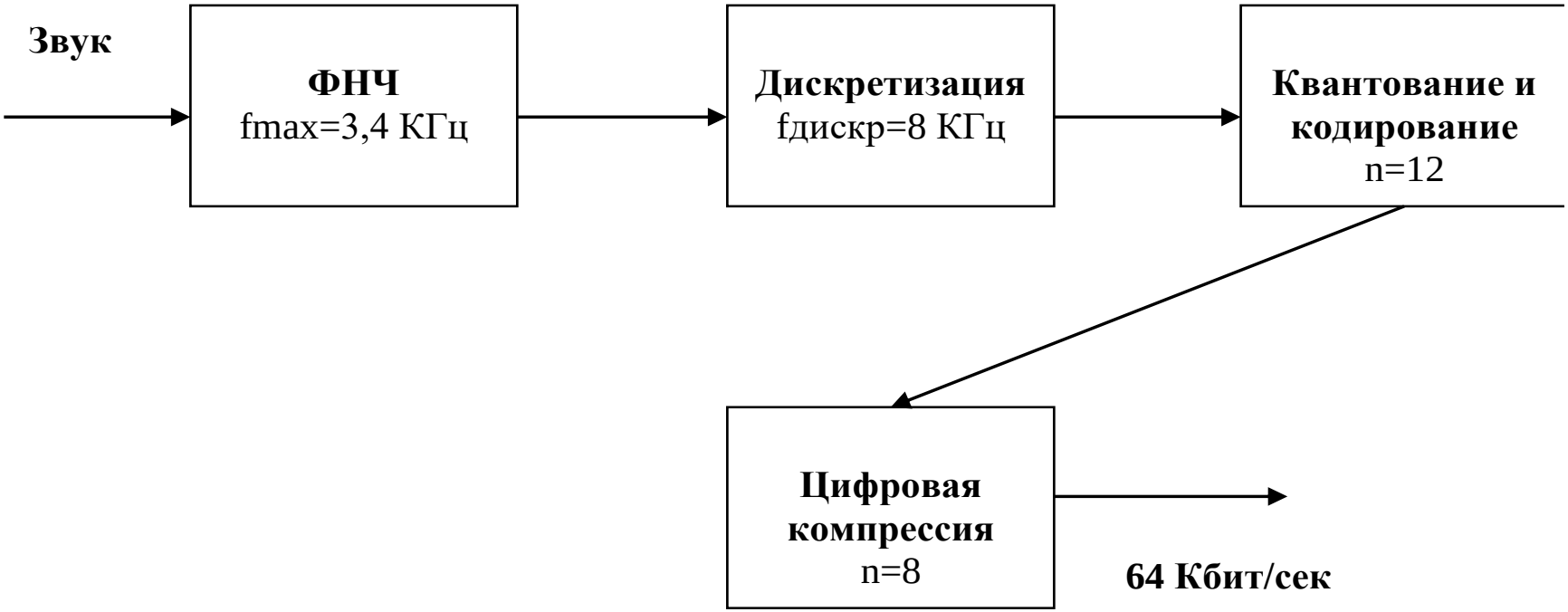
$$f_{ДИСКР} \geq 2 \times \mathrm{F}_{\max}$$

Тогда период дискретизации $T_{ДИСКР}$ аналогового сигнала, т.е. периоды времени, через которые формируются дискретные отсчеты, рассчитывается по формуле:

$$T_{ДИСКР} \leq \frac{1}{2\mathrm{F}_{\max}}$$

Кодер формы G.711

Рекомендация МСЭ-Т G.711



Компрессия осуществляется с помощью нелинейного безинерционного преобразования вида:

$$v = F_{комп}(U),$$

где U – значение сигнала на входе **компандера**;
v - соответствующее значение на выходе.

В современных системах связи с подвижными объектами наиболее часто используются два закона компандирования: **А-закон и μ-закон**.

Сжатие по А и мю-закону

Сжатие по А-закону определяется следующим нелинейным преобразованием:

$$v = \begin{cases} Au & \text{при } 0 \le u \le \dfrac{1}{A}, \\ 1 + \ln A & \\ 1 + \ln(Au) & \text{при } \dfrac{1}{A} \le u \le 1, \\ 1 + \ln A & \end{cases}$$

где А - параметр сжатия с типовыми значениями 86 (Северная Америка) и 87,56 (Европа) для семибитовых речевых преобразователей.

Максимальное значение и здесь должно быть нормировано к 1.

μ -закон сжатия имеет вид

$$v = \frac{\ln(1+\mu u)}{\ln(1+\mu)}, \quad 0 \le u \le 1.$$

где **μ** - положительная постоянная с возможными значениями из интервала от 50 до до 300. Максимальное значение и здесь также должно быть нормировано к 1

Общим для обоих *законов сжатия* является то, что функция преобразования является линейной или близка к линейной и имеет наибольшую производную при малых значениях аргумента и (речевого сигнала) и является примерно логарифмической или логарифмической при больших значениях.

Адаптивная импульсно-кодовая модуляция АДИКМ

Между соседними отсчетами речевого сигнала обычно наблюдается сильная корреляции. Это свойство реализуется при применении **адаптивной дифференциальной импульсно-кодовой модуляции (АДИКМ)**.

При сильной корреляции дисперсия разности $D(t_i) = U(t_i) - U(t_i\text{-}1)$ значений соседних отсчетов намного меньше дисперсии самих отсчетов $U(t_i)$.

При применении **АДИКМ** в цифровую форму переводятся приращения D(ti), i = ... -1,0,+1,..., которые в точке приема используются для восстановления значений отсчетов путем суммирования, т.е.

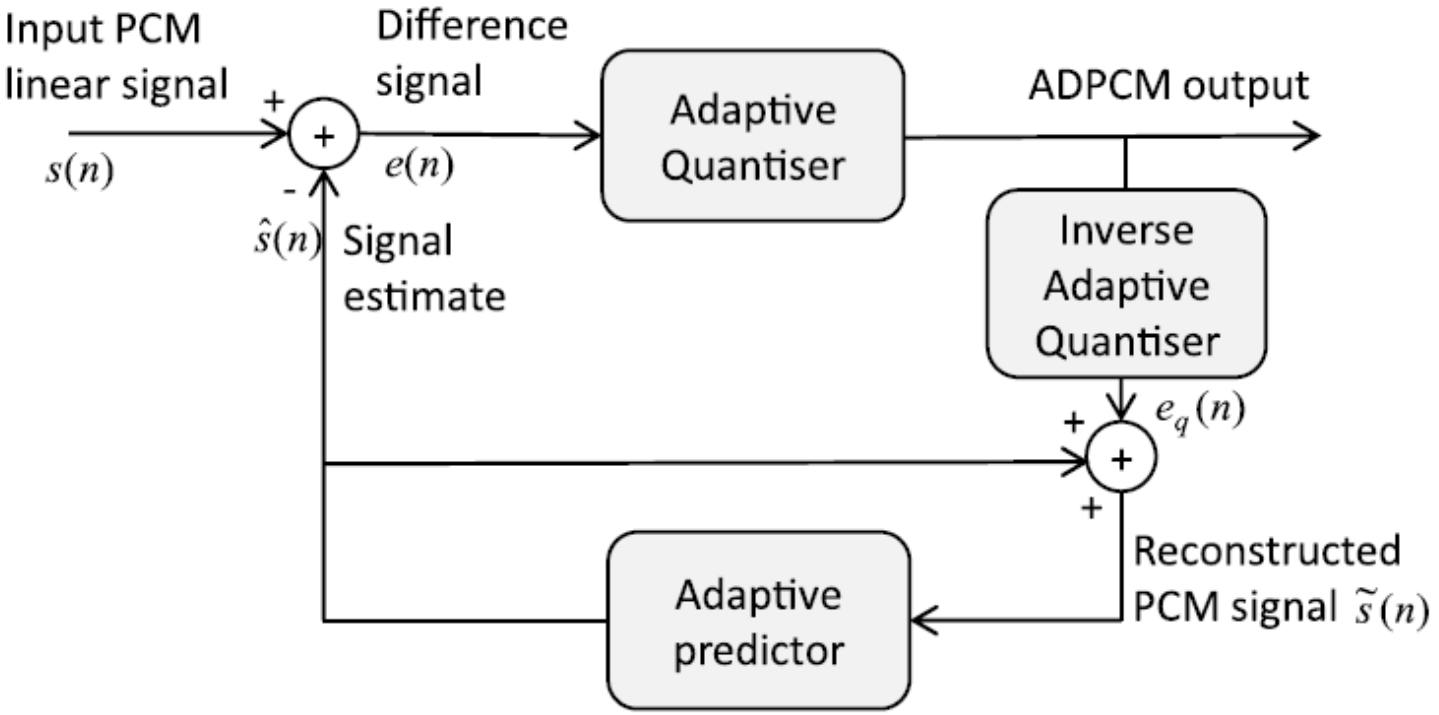
$$\hat{U}(t_i) = \hat{U}(t_{i-1}) + \hat{D}(t_i).$$

Здесь угловая скобка над символом обозначает оценку соответствующей величины, сформированную в приемнике.

На практике **АДИКМ** реализуется с применением различных алгоритмов предсказания: вместо кодирования разности соседних отсчетов кодируется разность между значением очередного отсчета и предсказанным его значением.

Эту разность обычно называют **ошибкой предсказания**, значение которой переводится в цифровую форму и передается по линии связи. Предсказание основывается на знании ковариационной функции речевого сигнала

Компоненты кодера АДИКМ



(a). Encoder

Линейное предсказание сигналов (G.728, G.729)

Линейное предсказание текущего значения сигнала S_n основывается на предположении, что это значение может быть представлено с небольшой ошибкой как линейная комбинация p предшествующих его значений, где p обычно выбирают из интервала от 10 до 15:

$$s_n = \sum_{k=1}^{p} a_k s_{n-k} + \varepsilon_n$$

здесь ε_n - ошибка предсказания.

Уравнение устанавливает, что текущее значение выходного сигнала может быть определено суммированием взвешенного текущего входного значения и взвешенной суммы предыдущих выходных выборок. В LPC даны измерения сигнала, требуется определить параметры передаточной функции системы.

Оценки коэффициентов в этом выражении вычисляются по конечной выборке наблюдаемых значений речевого сигнала в каждом сегменте и выбираются так, чтобы минимизировать значение суммы квадратов ошибок:

$$D = \sum_{n=1}^{N} e_n^2 = \sum_{n=1}^{N} \left(\sum_{k=0}^{p} a_k s_{n-k} \right)^2.$$

Полученные значения **коэффициентов фильтра предсказания** должны передаваться по линии связи.

Схема кодера с линейным предсказанием сигналов

Для обеспечения приемлемой точности *восстановления значений речевого сигнала* в приемнике оказалось необходимым использовать от 8 до 10 бит для цифрового представления каждого коэффициента. Поэтому вместо этих коэффициентов обычно передают так называемые коэффициенты отражения C(k), значения которых имеют меньший динамический диапазон и требуют для кодирования всего 6 бит на каждый коэффициент.

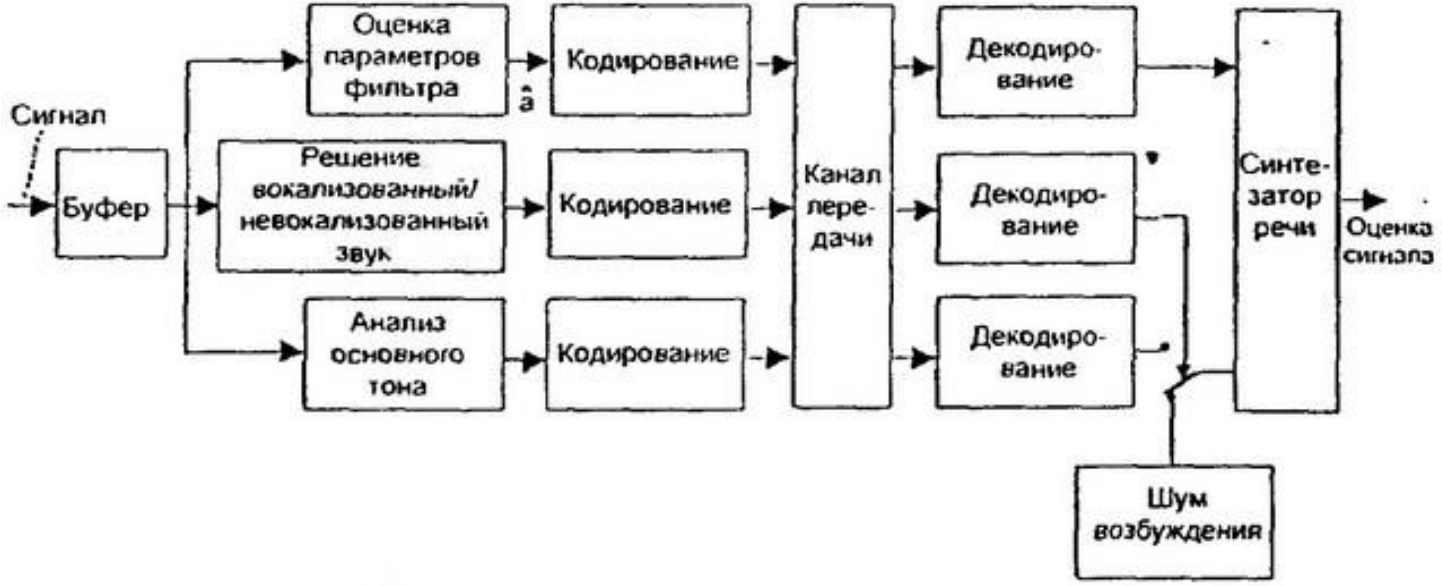
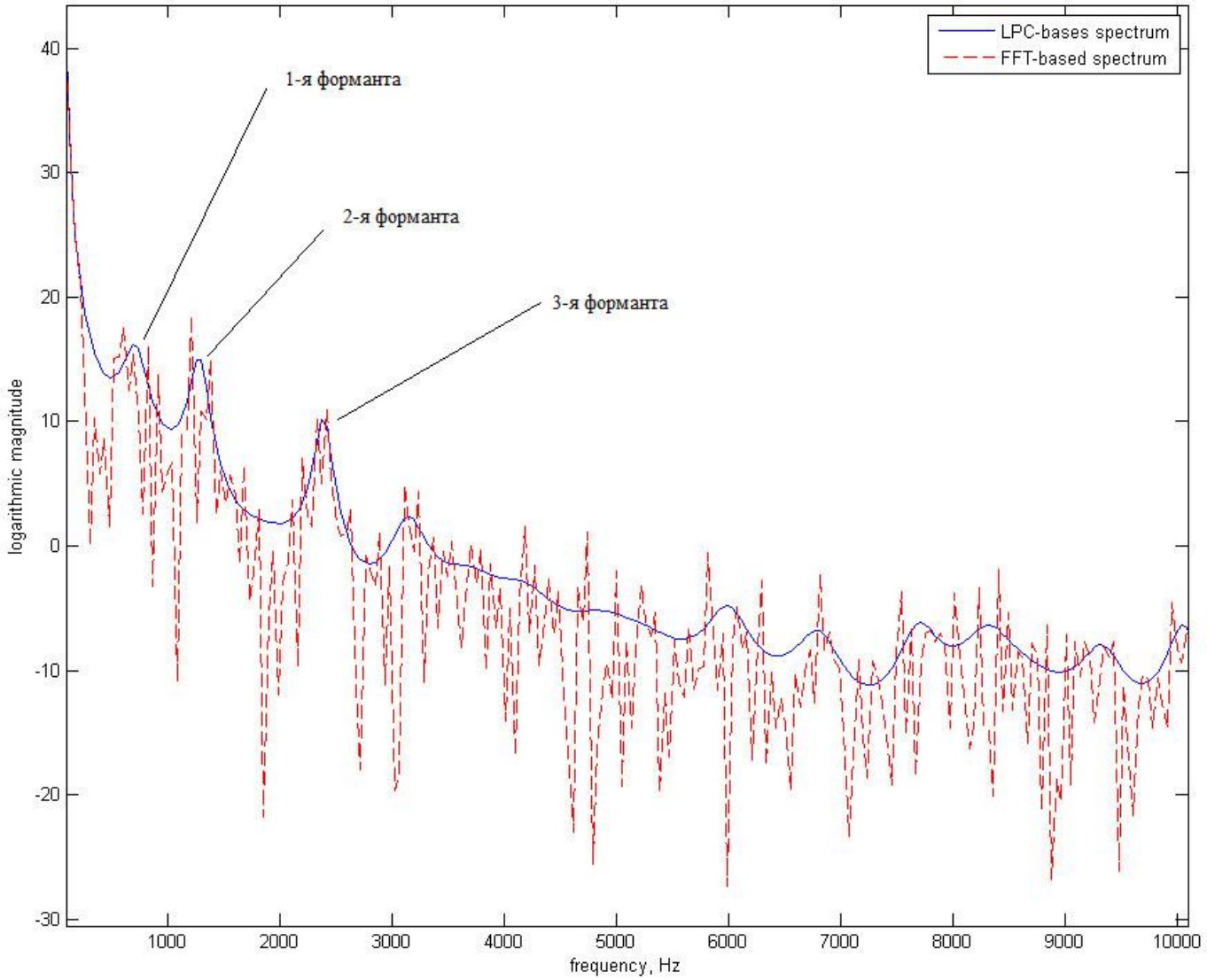


Рис. 2.13. Функциональная схема кодирования речевого сигнала на основе линейного предсказания

Точность предсказания



Гибридные кодеры G.723

CELP – Алгоритм кодирования с кодовым возбуждением и линейным предсказанием (Code Excited Linear Prediction), ориентирован на низкие скорости.

Линейное предсказание при анализе речевых сигналов обычно используется в двух направлениях: проведение кратковременного спектрального анализа речи и построение систем анализа–синтеза.

В результате метод CELP входит в группу методов «анализ через синтез» /analysis by syntesis, AbS/ и занимает промежуточное положение между кодерами формы и параметрическими вокодерами.

MP-MLQ – много импульсное возбуждение с использованием алгоритма максимального правдоподобия (для больших скоростей изменения сигнала)
LSP – пары линейного предсказания

В этих схемах используется процедура оптимизации типа «замкнутая петля» для нахождения возбуждающего сигнала, который при возбуждении моделирующего фильтра создает оптимальный речевой сигнал. Это позволяет схемам AbS более успешно работать на скоростях 4,8.. .9,6 кбит/с.

В методе AbS допускается, что сигнал можно исследовать и представить в какой-либо форме, например в виде временных или частотных доменов. Затем созданная модель сигнала подвергается оптимизации (подгонке). Модель имеет несколько параметров, изменение которых приводит к изменению формы моделируемого сигнала.

Обобщенная и детальная схема «анализ через синтез AbS»

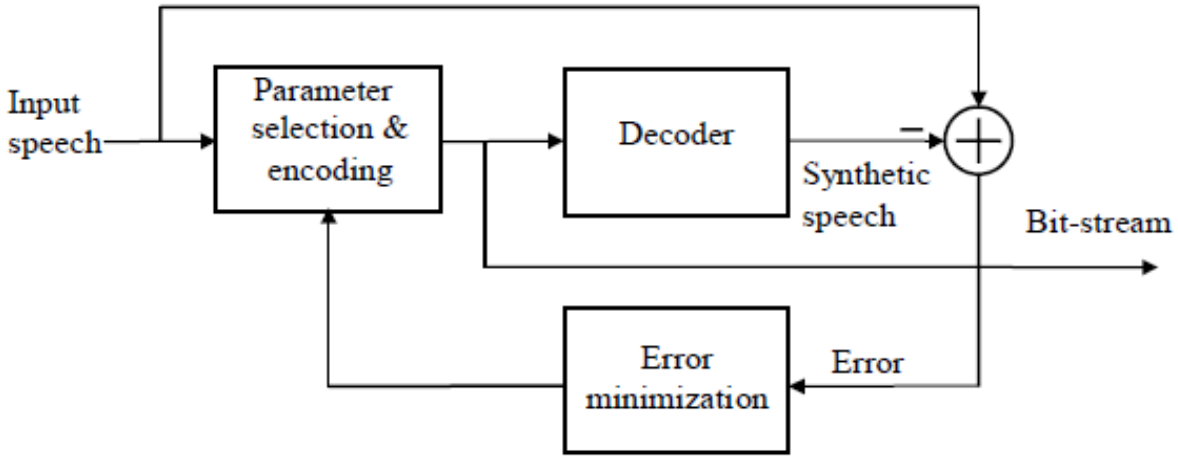


Схема кодера

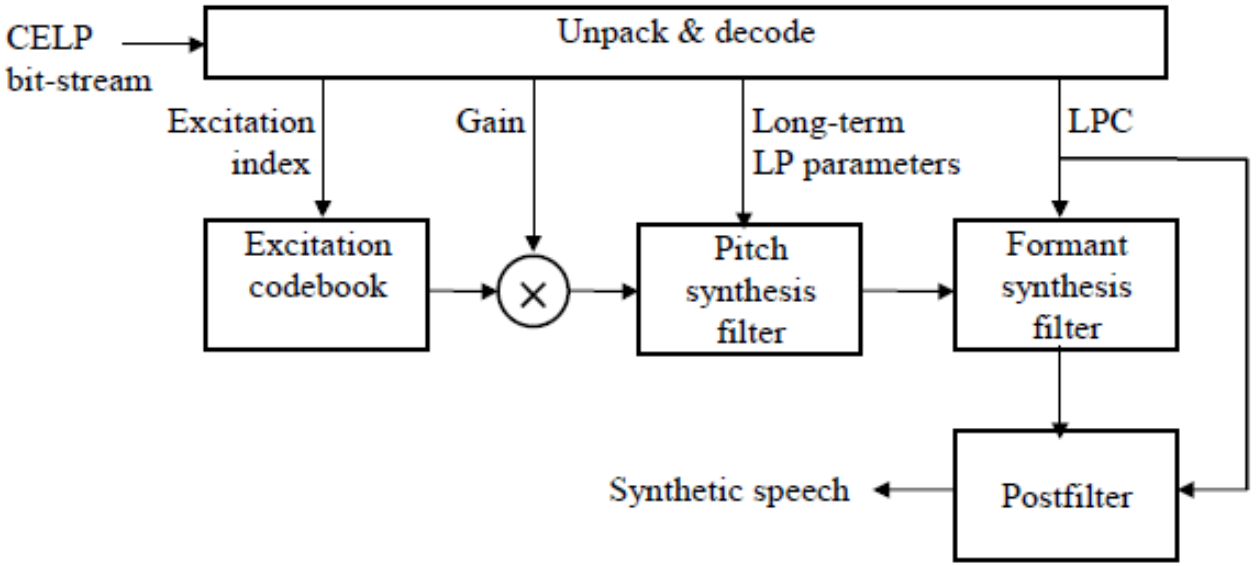
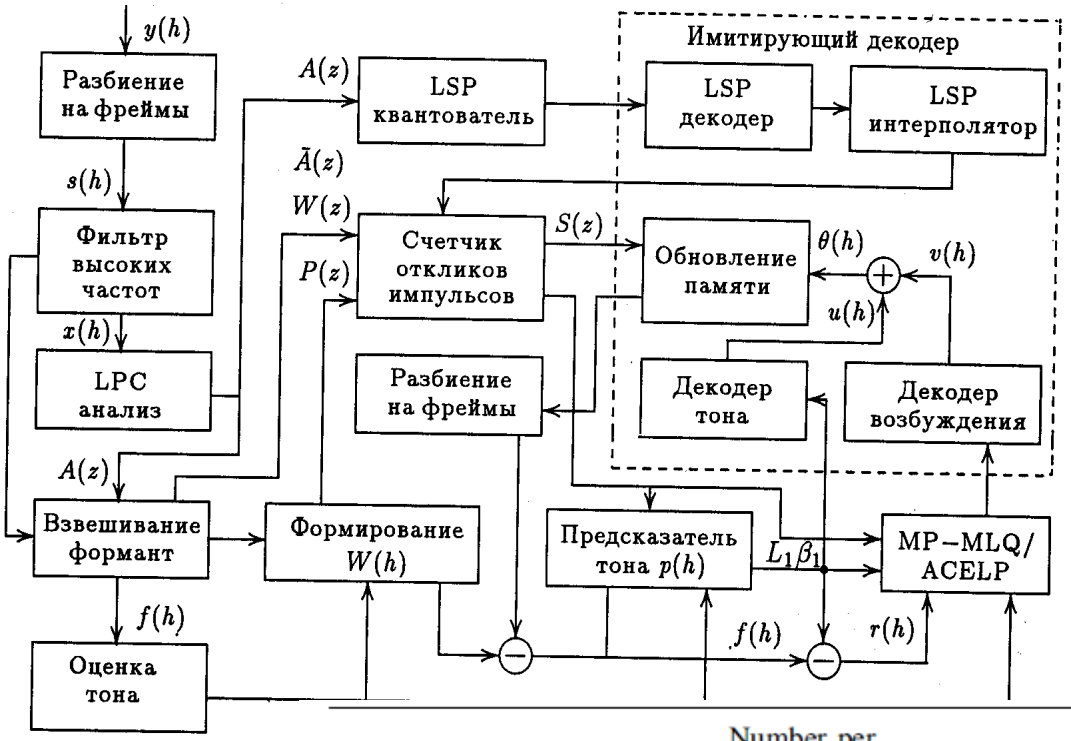


Схема декодера

Обобщенная и детальная схема «анализ через синтез AbS»



Количество бит для кодирования одного фрейма для кодера с процедурой analysis-by-synthesis (анализ через синтез) Скорость 4800 бис, длина фрейма 30 мс

Parameter	Number per Frame	Resolution	Total Bits per Frame
LPC	10	3, 4, 4, 4, 4, 3, 3, 3, 3, 3	34
Pitch period (adaptive codebook index)	4	8, 6, 8, 6	28
Adaptive codebook gain	4	5	20
Stochastic codebook index	4	9	36
Stochastic codebook gain	4	5	20
Synchronization	1	1	1
Error correction	4	1	4
Future expansion	1	1	1
Total			**144**

Многоскоростной кодер AMR

Многоскоростной речевой кодер – это единый интегрированный речевой кодек с восемью исходными скоростями: 12,65 (GSM-EFR); 8,85; 6,60 (PDS-EFR) а также 23,85; 23,05; 19,85; 18,25; 15,85; 14,25 Кбит/с согласно Рек. МСЭ–Т G.722.2–2003. Скорости передачи в AMR-WB управляются сетью радиодоступа и не зависят от речевой активности.

Для облегчения совместимости с существующими сотовыми системами некоторые из режимов выбраны такими же, что и у существующих сотовых сетей. Речевой кодек AMR со скоростью передачи 12,65 Кбит/с соответствует кодеку EFR в GSM, со скоростью передачи 8,85 Кбит/с соответствует речевому кодеку US-TDMA, а со скоростью передачи 6,6 Кбит/с – Японскому кодеку PDS.

Речевой кодер AMR может по команде производить переключение своей скорость передачи в каждом речевом фрейме [кадре] длительностью 20 мс. Для переключения режима AMR выбраны два способа: управление по каналам сети или с использованием выделенного канала.

Кодер AMR работает с речевыми фреймами длительностью 20 мс, что соответствует 160 выборкам при частоте 8000 выборок в секунду. Схема режимов многоскоростного кодирования представляет собой так называемый алгебраический метод кодирования и линейного предсказания (ACELP).

Основные функции AMR

Биты с параметрами речи, переданные кодирующим устройством речи, перераспределяются в соответствии с их субъективной важностью перед тем, как они передаются по сети. Перераспределенные биты затем сортируются с учетом их восприимчивости к ошибкам и делятся на три класса по их важности: A, B и C.

Класс A является наиболее уязвимым, и в воздушном интерфейсе используется самое мощное канальное кодирование для битов класса A.

У AMR имеется три основных функции для эффективного использования прерывистой занятости:

- Детектор речевой активности (VAD) на передающей стороне
- Оценка фонового акустического шума на передающей стороне для того, чтобы передавать характеристические параметры приемной стороне
- Передача комфортного шумового фона на приемную сторону, что достигается посредством фрейма Дескриптора тишины, который посылается через одинаковые промежутки времени
- Генерация (воспроизведение) комфортного шума на приемной стороне в периоды, когда не принимаются нормальные речевые фреймы.

Спецификация AMR также содержит механизмы скрадывания ошибок. Замена фрейма имеет целью скрыть (ослабить) влияние речевых фреймов AMR. Заглушение выходного сигнала при потере нескольких фреймов производится для того, чтобы показать нарушение канала пользователю и избежать возможного возникновения раздражающих звуков в результате процедуры замены фреймов.

ПГУТИ

ССС
СЕТИ И СИСТЕМЫ СВЯЗИ

5.3 Оценка качества передачи речи

E-модель оценки качества речи

E-модель, согласно Рек. МСЭ-Т G.107, 2015 г., основана на методе коэффициентов снижения качества оборудования, который логически вытекает из предшествующих моделей оценки качества передачи. Метод был разработан специальной группой ETSI по качеству передачи речи ото рта к уху (Voice Transmission Quality from Mouth to Ear).

Величины шума помещения и коэффициентов D обрабатывают отдельно для передающей и приемной сторон, могут иметь разные значения.

Параметры показатель громкости передачи (SLR), показатель громкости приема (RLR) и шум цепи Nc сравнивают с определенной точкой 0 дБо.

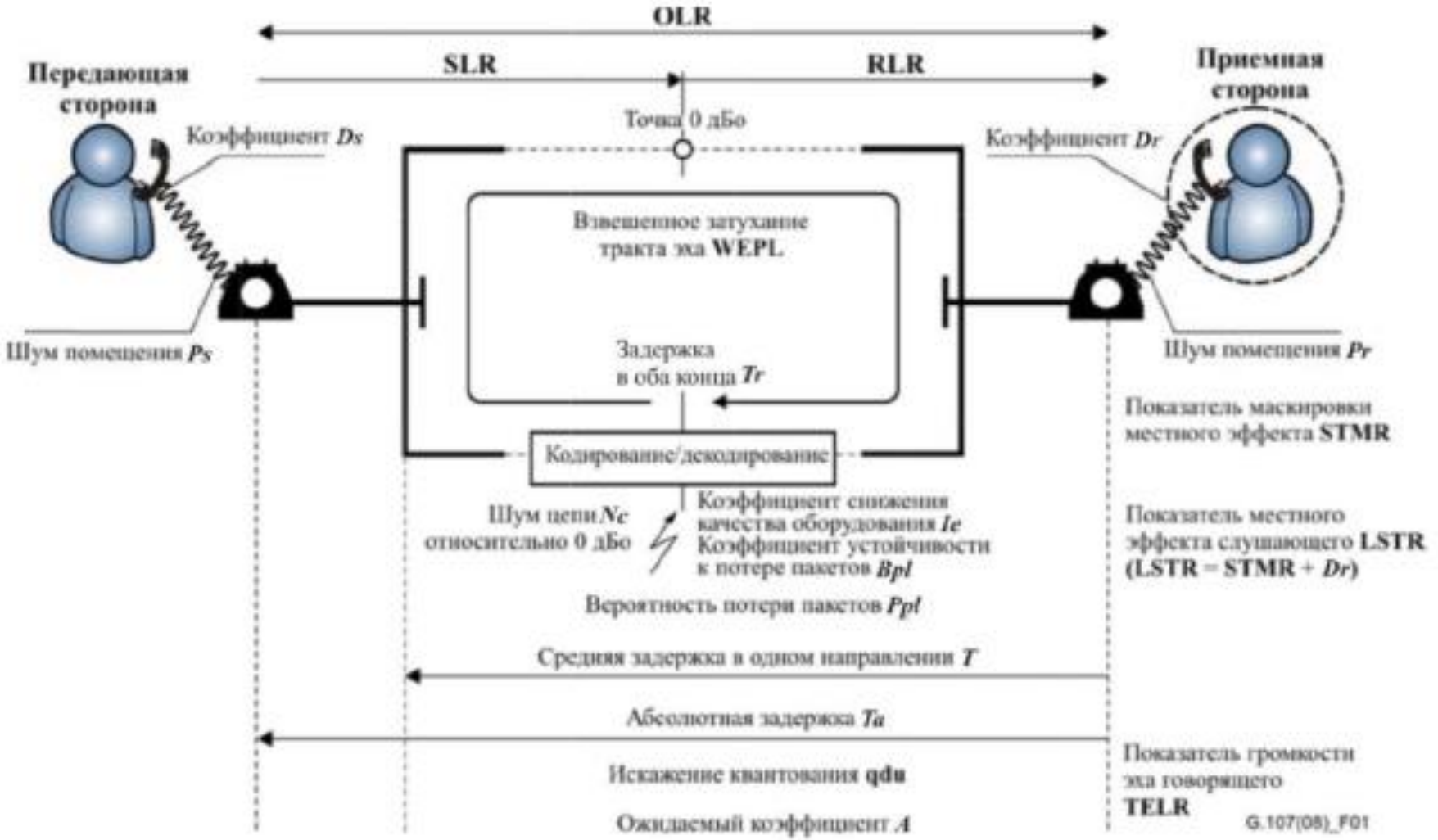
Другие входные параметры рассматривают:

– как значения для общего соединения, например общий показатель громкости (OLR), то есть сумма SLR и RLR, число устройств с qdu, коэффициенты снижения качества оборудования Ie и коэффициент выигрыша A,

либо

– как значения, относящиеся только к приемной стороне, например показатель маскировки местного эффекта (STMR), показатель местного эффекта слушающего (LSTR), взвешенная потеря эха в тракте передачи (WEPL), используемая для вычисления эха слушающего, и показатель громкости эха говорящего (TELR).

Схема E-модели оценки качества речи



Параметры E-модели оценки качества речи

***R* – показатель оценки характеристики передачи речи**

$$R = Ro - Is - Id - Ie\text{-}eff + A.$$

где :

Is – коэффициент представляет комбинацию из всех снижений качества, которые действуют на речевой сигнал более или менее одновременно;

коэффициент *Id* представляет снижение качества, вызываемое задержкой;

а – коэффициент снижения эффективности оборудования *Ie-ef*f представляет снижение качества, вызываемое кодеками с низкой битовой скоростью. Этот коэффициент также включает снижение качества из-за потери пакетов с произвольным распределением;

коэффициент выигрыша *A* позволяет компенсировать коэффициенты снижения качества в тех случаях, когда пользователь получает преимущества от других типов доступа к сети;

член *Ro* и величины *Is* и *Id* подразделяют на дальнейшие специфические значения снижения качества.

Коэффициент выигрыша A

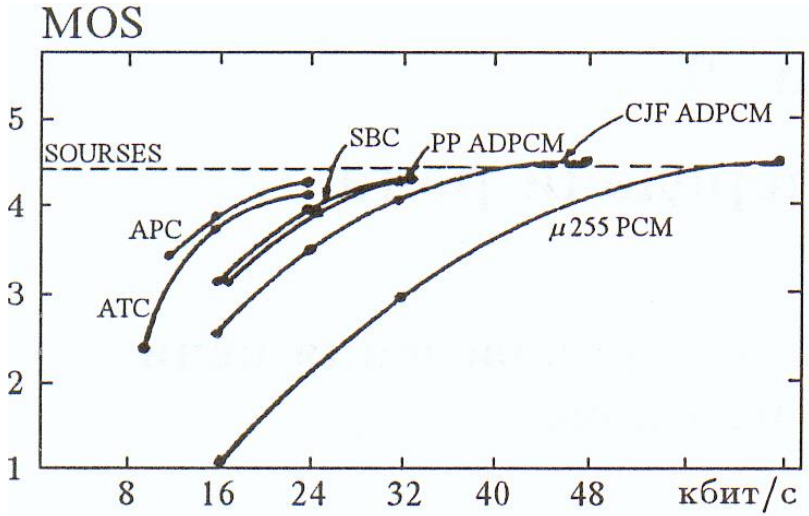
Таблица 2 – Примеры предварительных значений для коэффициента выигрыша *A*

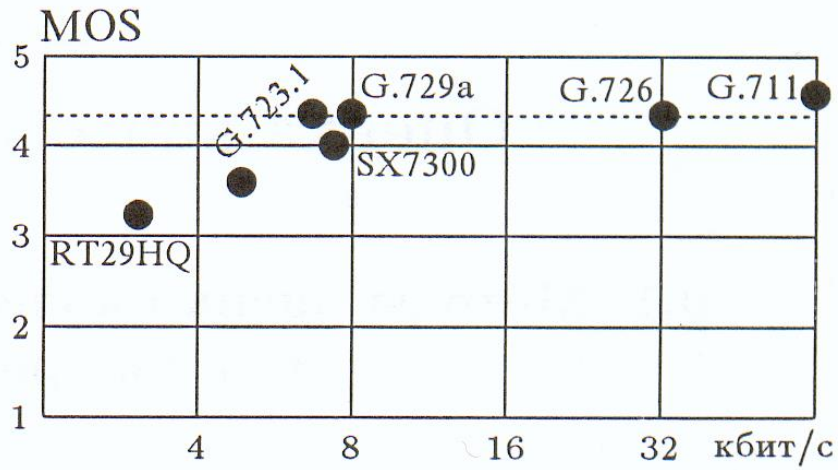
Пример системы связи	Максимальное значение *A*
Обычная (проводная)	0
Мобильная с помощью сотовой сети в здании	5
Мобильная в географической области или подвижная на транспортном средстве	10
Доступ к труднодоступному месту, например с помощью спутников со многими интервалами соединений	20

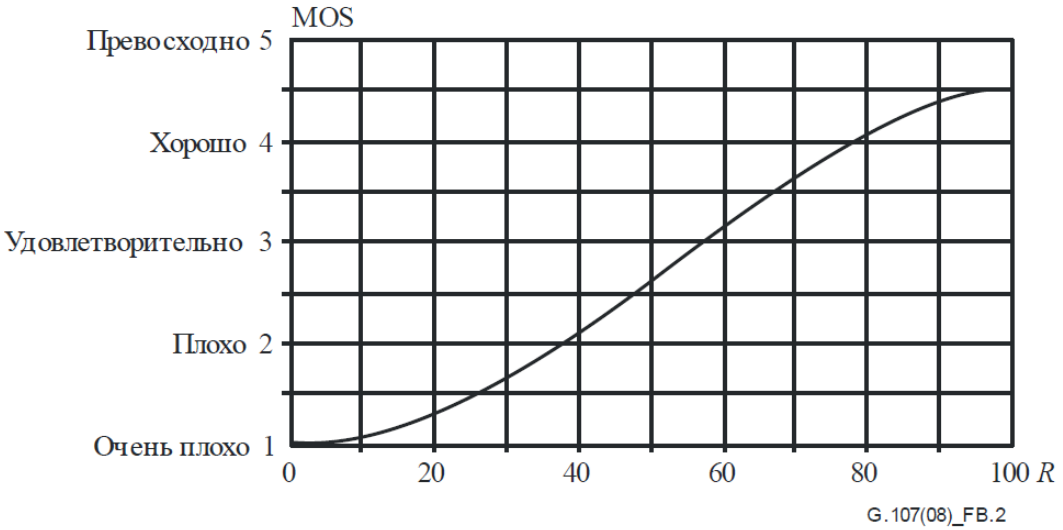
Пересчет оценки R в среднесубъективную оценку качеств передачи речи MOS

$$MOS = \begin{cases} 1 & \text{for } R \le 0 \\ 1 + 0.035R + R(R-60)(100-R)7 \cdot 10^{-6} & \text{for } 0 < R < 100 \\ 4.5 & \text{for } R \ge 100 \end{cases}$$

Экспертные оценки качества передачи речи







Выводы по разделу 4

1. В сетях современных сетях связи используются методы цифрового кодирования и декодирования речевых сигналов. Различают кодеры формы, вокодеры и гибридные кодеры. Цель работы каждого кодера – обеспечить снижение скорости передачи речевых сигналов при сохранении приемлемого качества.

2. Кодеры формы используют преимущественно описание огибающей речевого сигнала. Вокодеры используют обобщенную модель человеческой речи с разделением на вокализованные, невокализованные и шумы. Гибридные кодеры объединяют функции кодеров формы и вокодеров, применяют методы линейного предсказания сигналов.

3. Качество речи при передачи по сетям связи можно оценить аналитически или использовать 5-ти балльную шкалу средне-субъективной оценки.