Восприятие и цифровое представление речевой информации

Введение

- Виды аппаратной и программной реализации систем цифровой обработки речевой и звуковой информации определяются их исходными характеристиками, особенностями слухового восприятия и требованиями к качеству воспроизведения
- Речевая информация, образующая свойственные используемому языку фонетические комбинации и формирующая те или иные смысловые элементы, по своим физическим параметрам принципиально отличается от звуковой информации, содержащей сочетание голосовых данных с музыкальным сопровождением, особенности и отличия друг от друга речевой и звуковой информации используются при их цифровой обработке и сжатии
- Основную информацию о звуковых колебаниях человек получает в области частот примерно до 4 кГц, именно эти частоты задают разборчивость и ясность аудиоинформации
- Спектральный состав речи занимает полосу частот примерно от 50 до 7000-10000 Гц
- В аналоговой телефонии используется полоса частот 0,3-3,4 кГц, что ухудшает восприятие ряда звуков (например, шипящих), но практически не отражается на разборчивости речи

- В цифровой телефонии отсчеты аналоговой речи приходится брать согласно теореме Котельникова с частотой 8 кГц
- Разрядность аналого-цифрового преобразования для речи – 8 или 16 бит на отсчет
- Идея преобразовывать в цифровой вид не сам речевой сигнал, а его параметры (количество переходов через ноль, спектральные характеристики и др.), чтобы затем по этим параметрам выбирать модель голосового тракта и синтезировать исходный сигнал, лежит в основе синтезирующих кодеков или вокодеров
- Принцип работы гибридных кодеков основан на модели кодирования с использованием линейного предсказания и алгебраической кодовой книги, при этом производится анализ речевого сигнала и выделяются параметры модели (коэффициенты системы линейного предсказания, индексы и коэффициенты усиления в адаптивной и фиксированной кодовых книгах), далее эти параметры кодируются и передаются в канал
- Слуховой аппарат человека различает частотные составляющие звука приблизительно в пределах от 30 Гц до 20 кГц; верхняя граница может несколько отличаться в зависимости от возраста человека, условий воспроизведения информации и др.
- Объективно качество речи оценивается методикой PESQ, доступной по ссылке:

  https://www.itu.int/rec/T-REC-P.862-200102-I/en

Сигналограммы фрагментов музыкальной записи и речи

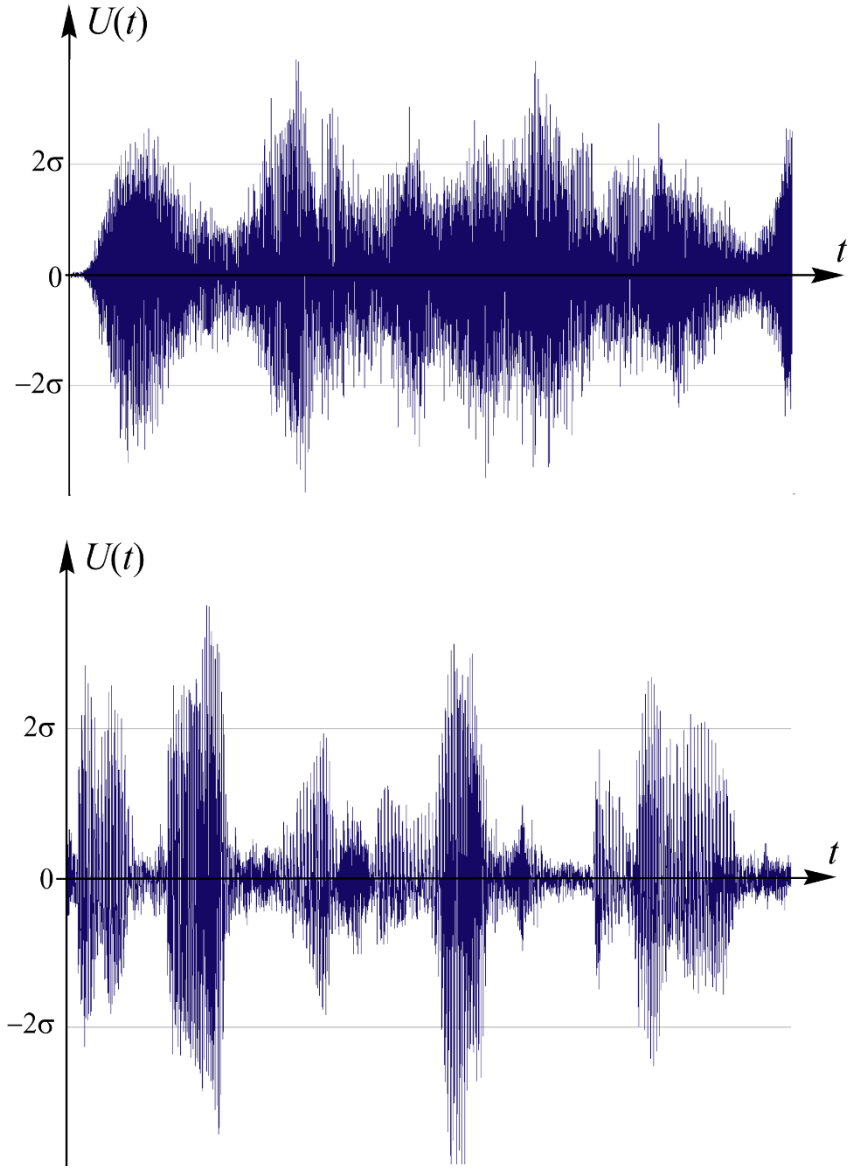
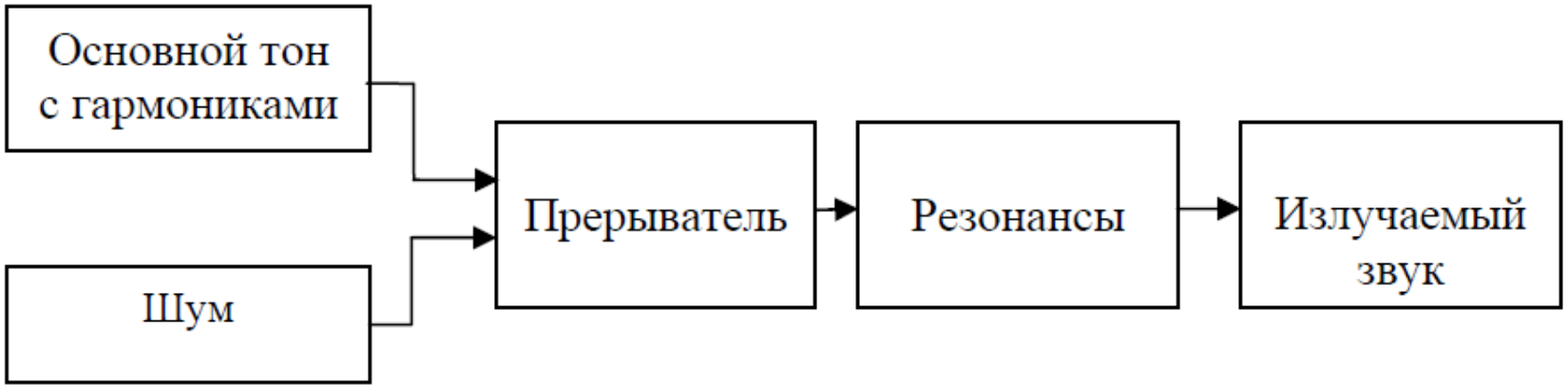
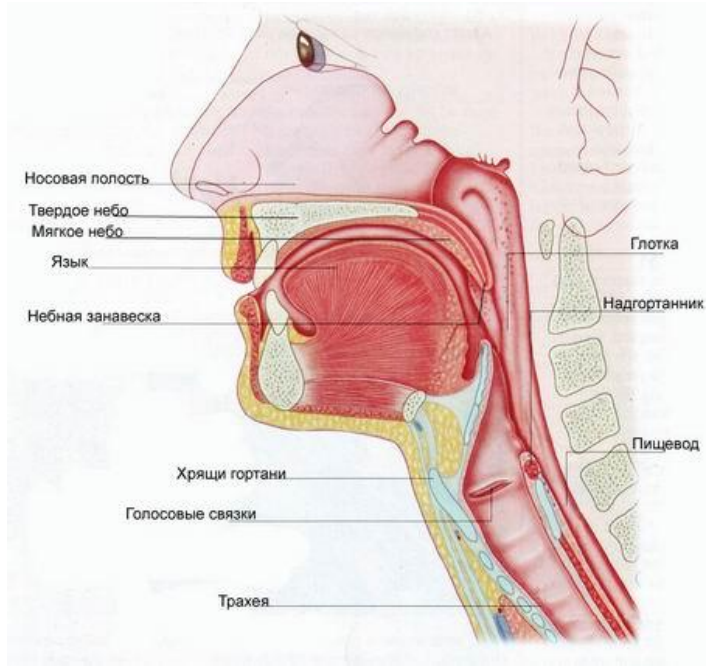


Схема речеобразования





Звуковые символы, из которых составлен речевой сигнал, называются фонемами.
В русском языке насчитывается 42 основные и 3 неопределенные фонемы, 5 гласных и 37 согласных фонем.
При произнесении звуков речи через речевой тракт проходит или тональный импульсный сигнал, или шумовой, или тот и другой вместе. Речевой тракт представляет собой сложный акустический фильтр с рядом резонансов, создаваемых полостями рта, носа и носоглотки, т. е. с помощью артикуляционных органов речи. Вследствие этого равномерный тональный или шумовой спектр превращается в спектр с рядом максимумов и минимумов. Максимумы спектра называют формантами, а нулевые провалы – антиформантами.

Виды фонем

Согласные по способу образования делятся на сонорные (л, ль, р, рь, м, мь, н, нь, й), щелевые (ж, з, зь, в, вь, ш, с, сь, ф, фь, х, хь), взрывные (б, бь, д, дь, г, гь, п, пь, т, ть, к, кь) и аффрикаты (ц, ч – комбинация глухих взрывных и щелевых). Гласных фонем всего шесть: а, о, у, э, и, ы (гласные е, я, ё, ю – составные из и краткого или мягкого знака и гласных э, а, о, у).

По месту образования фонемы делятся на губные, зубные, небные, гортанные, передние и задние.

Звонкие звуки речи, особенно гласные, имеют высокий уровень интенсивности, глухие – самый низкий. Громкость речи непрерывно изменяется, особенно резко при произнесении взрывных звуков. Динамический диапазон уровней речи находится в пределах 35 – 45 дБ. Гласные звуки речи имеют в среднем длительность около 0,15 с, согласные – около 0,08 с (звук п – около 0,03 с).

Для записи речи необходимо отводить не менее 8 бит на каждый отсчет.

Основной тон

Звуки речи делятся на звонкие и глухие. Звонкие звуки образуются с участием голосовых связок, в этом случае находящихся в напряженном состоянии. Под напором воздуха, идущего из легких, они периодически раздвигаются, в результате чего создается прерывистый поток воздуха. Импульсы потока воздуха, создаваемые голосовыми связками, с достаточной точностью могут считаться периодическими.

Соответствующий период повторения импульсов называют периодом основного тона голоса.

Частота основного тона для всех голосов лежит в пределах 70 – 450 Гц.

Импульсы основного тона имеют пилообразную форму, и поэтому при их периодическом повторении получается дискретный спектр с большим числом гармоник (до 40), частоты которых кратны частоте основного тона.

Огибающая спектра основного тона имеет спад в сторону высоких частот с крутизной около 6 дБ/октаву, поэтому для мужского голоса уровень составляющих на частоте 3000 Гц ниже их уровня на частоте 100 Гц примерно на 30 дБ. При произнесении глухих звуков связки находятся в расслабленном состоянии, поток воздуха из легких свободно проходит в полость рта. Встречая на своем пути различные преграды в виде языка, зубов, губ, он образует завихрения, создающие шум со сплошным спектром.

Информативность речевого сигнала

Звуки речи неодинаково информативны. Так, гласные звуки содержат малую информацию о смысле речи, а глухие согласные наиболее информативны.

Разборчивость речи снижается при действии шумов, в первую очередь из-за маскировки глухих звуков.

Полоса равна 7 000 Гц, динамический диапазон 42 дБ, т.е. требуется семизначный код, откуда имеем требование на пропускную способность линии связи: $2 \cdot 7000 \cdot 7 = 98\,000$ бит/с.

При этом образование звуков речи происходит путем подачи команд к мускулам артикуляционных органов речи от речевого центра мозга. Общий поток сообщений от него составляет в среднем не более 100 бит/с. Вся остальная информация в речевом сигнале называется сопутствующей.

Речевой сигнал представляет собой своего рода модулированную несущую. Его спектр $p(\omega)=E(\omega)\cdot F(\omega)$, где $E(\omega)$ – спектр генераторной функции, т. е. импульсов основного тона или шума; $F(\omega)$ – модулирующая кривая фильтровой функции речевого тракта. Несущая имеет широкополосный спектр, но почти вся информация о звуках речи заключена в спектральной огибающей речи и ее временном изменении, частично – в переходах от тонального спектра к шумовому и обратно, по которым узнают о смене звонких звуков на глухие и обратно. Все эти изменения происходят медленно.

Для воспроизведения речи достаточно передавать сведения о форме огибающей спектра речи и ее временном изменении в темпе смены звуков речи, а также об изменении основного тона речи и переходов тон-шум.

Вокодеры

На передающей станции установлен спектроанализатор, который вычисляет кратковременное преобразование Фурье. В результате для каждого временного сегмента получаем $N/2$ комплексных коэффициентов, или N вещественных чисел. Если передавать на расстояние все эти числа, а затем на приемной станции из них восстанавливать сегмент сигнала, тогда сигнал восстановится без потерь.

Можно передавать на расстояние не все спектральные коэффициенты, а лишь «большие». Разумеется, при этом нужно еще указать центральные частоты соответствующих полосовых фильтров (или, что то же, номера коэффициентов).

Основная информация о гласных звуках содержится всего в трех формантах. Это означает, что одну гласную можно представить набором из девяти чисел, тогда как при обычном способе передачи для этого требуется примерно 100 чисел (интервал 10 мс при частоте дискретизации 10 кГц). Для согласных звуков эта разница значительно меньше.

Дальнейшего сжатия можно добиться, например, уменьшив разрядность передаваемых чисел. В конечном счете на практике удается понизить скорость передачи данных с 64 до 9,6 кбит/с без существенной потери качества синтезированной речи, и даже до 2,4 кбит/с, но уже с заметной потерей качества.

Современные вокодеры обеспечивают хорошее качество речи при скорости передачи 800 – 2 400 бит/с и качество речи, пригодное для ведения служебных переговоров, при скорости передачи 1 200 бит/с.

Виды вокодеров

В полосных вокодерах спектр речи делится на 7 – 20 полос (каналов) аналоговыми или цифровыми полосовыми фильтрами. Большее число каналов в вокодере дает большую натуральность и разборчивость. С каждого полосового фильтра сигнал поступает на детектор и фильтр низких частот с частотой среза 25 Гц. Таким образом, сигналы на выходе каждого канала изменяются с частотой менее 25 Гц. Их передача возможна в аналоговом или цифровом виде.

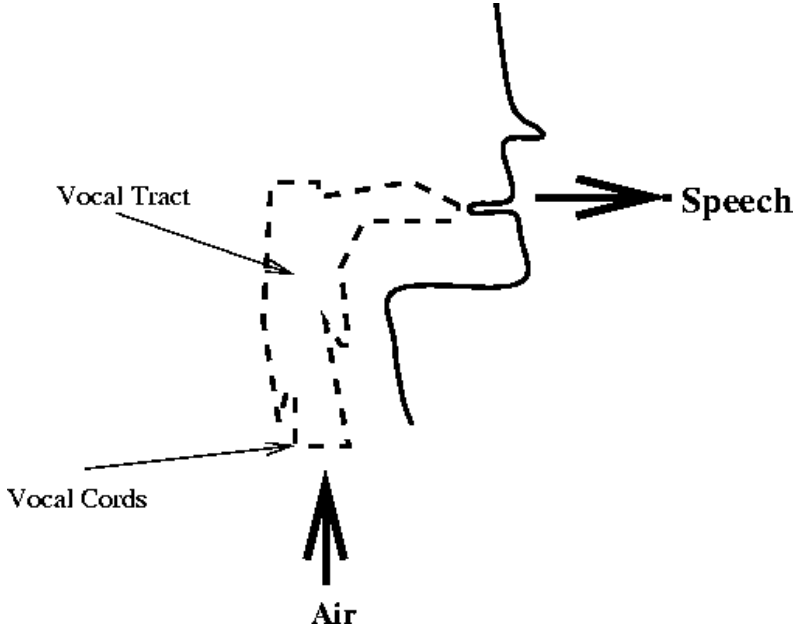
В формантных вокодерах огибающая спектра речи описывается комбинацией формант (резонансных частот голосового тракта). Основные параметры формант – центральная частота, амплитуда и ширина.

В ортогональных вокодерах огибающая мгновенного спектра раскладывается в ряд по выбранной системе ортогональных базисных функций. Вычисленные коэффициенты этого разложения передаются на приемную сторону. Распространение получили гармонические вокодеры, использующие разложение в ряд Фурье.

Вокодеры с линейным предсказанием (Linear Prediction Coding, LPC), или липредеры, основаны на оригинальном математическом аппарате: передаются не параметры речевого сигнала, как такового, а параметры некоторого фильтра, в известном смысле эквивалентного голосовому тракту, и параметры сигнала возбуждения этого фильтра.

LPC

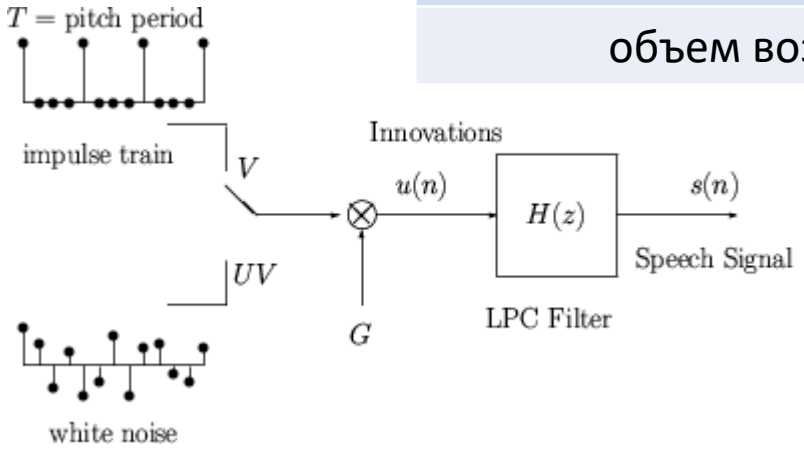
- Множество современных звуковых кодеков основано на кодировании с линейным предсказанием (LPC, linear predictive coding)
- При разговоре
  - Воздух из легких проходит голосовой тракт
  - При звонких звуках голосовые связки вибрируют, скорость этих вибраций определяет тембр голоса; у женщин и детей он выше (более быстрые колебания), чем у взрослых мужчин
  - При фрикативных и глухих звуках связки не вибрируют, а остаются приоткрытыми
  - Форма голосового тракта определяет произносимые звуки
  - При разговоре форма тракта меняется, давая различные звуки
  - Форма звукового тракта изменяется сравнительно медленно (10 – 100 мсек)
  - Количество выходящего из легких воздуха определяет громкость



Математическая модель LPC

- Цифровой голосовой сигнал – это выход цифрового фильтра LPC, на вход которого поступают либо последовательности импульсов, либо участки белого шума
- Связь между физической и математической моделями:

Физическая модель	Математическая модель
голосовой тракт	$H(z)$ (фильтр LPC)
воздух	$u(n)$ (обновления)
колебания голосовых связок	V (огласованный звук)
период колебаний связок	T (период основного тона)
неогласованные звуки	UV (неогласованный звук)
объем воздуха	G (усиление)



Математическая модель LPC

- Фильтр LPC определяется формулой: что эквивалентно следующей связи входа и выхода фильтра:

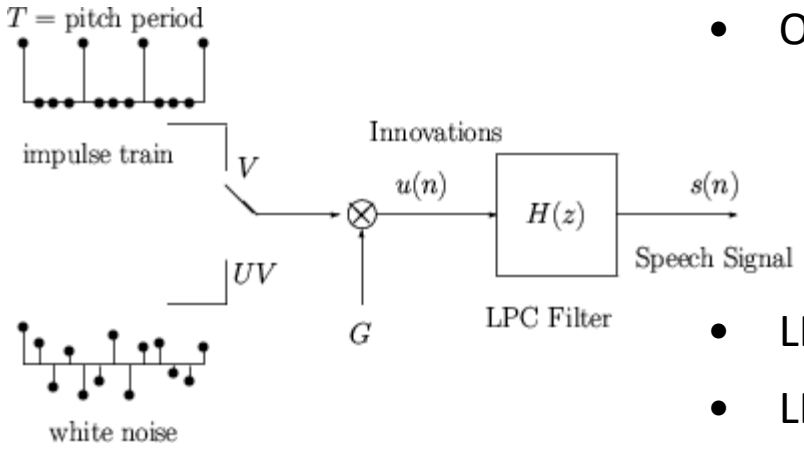
$$H(z) = \frac{1}{1 + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_{10} z^{-10}}$$

$$s(n) + \sum_{i=1}^{10} a_i s(n-i) = u(n)$$

- Модель LPC представляется в виде вектора

$$\mathbf{A} = (a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9, a_{10}, G, V/UV, T)$$

- Вектор $\mathbf{A}$ изменяется примерно каждые 20 мсек, что при частоте дискретизации 8 кГц соответствует 160 отсчетам
- Цифровой голосовой сигнал делится на кадры по 20 мсек (50 кадров/сек)
- Таким образом, согласно модели, 160 отсчетов сигнала $\mathbf{S}$ компактно представляются 13-ю значениями вектора $\mathbf{A}$



- Особенности восприятия:
  - Для звонких звуков - сдвиг импульсов (нечувствительность к фазе)
  - Для неогласованных звуков – используются различные шумовые последовательности
- LPC синтез: получение $\mathbf{S}$ из $\mathbf{A}$ (фильтрация)
- LPC анализ: оценка $\mathbf{A}$ по $\mathbf{S}$

LPC анализ

- Рассмотрим один кадр голосового сигнала: **S = (*s*(0), *s*(1), …, *s*(159))**
- Сигнал ***s(n)*** связан с обновлением ***u(n)*** линейным уравнением:

$$s(n) + \sum_{i=1}^{10} a_i s(n-i) = u(n)$$

- Десять параметров LPC $(a_1, a_2, \ldots, a_{10})$ выбираются так, чтобы минимизировать энергию обновлений:

$$f = \sum_{n=0}^{159} u^2(n)$$

- Стандартный подход - производные f по a_i равны нулю:

$$\begin{aligned} df/da_1 &= 0 \\ df/da_2 &= 0 \\ &\cdots \\ df/da_{10} &= 0 \end{aligned}$$

- Таким образом, получаем 10 линейных уравнений с 10-ю неизвестными:

$$\begin{bmatrix} R(0) & R(1) & R(2) & R(3) & R(4) & R(5) & R(6) & R(7) & R(8) & R(9) \\ R(1) & R(0) & R(1) & R(2) & R(3) & R(4) & R(5) & R(6) & R(7) & R(8) \\ R(2) & R(1) & R(0) & R(1) & R(2) & R(3) & R(4) & R(5) & R(6) & R(7) \\ R(3) & R(2) & R(1) & R(0) & R(1) & R(2) & R(3) & R(4) & R(5) & R(6) \\ R(4) & R(3) & R(2) & R(1) & R(0) & R(1) & R(2) & R(3) & R(4) & R(5) \\ R(5) & R(4) & R(3) & R(2) & R(1) & R(0) & R(1) & R(2) & R(3) & R(4) \\ R(6) & R(5) & R(4) & R(3) & R(2) & R(1) & R(0) & R(1) & R(2) & R(3) \\ R(7) & R(6) & R(5) & R(4) & R(3) & R(2) & R(1) & R(0) & R(1) & R(2) \\ R(8) & R(7) & R(6) & R(5) & R(4) & R(3) & R(2) & R(1) & R(0) & R(1) \\ R(9) & R(8) & R(7) & R(6) & R(5) & R(4) & R(3) & R(2) & R(1) & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \\ a_7 \\ a_8 \\ a_9 \\ a_{10} \end{bmatrix} = \begin{bmatrix} -R(1) \\ -R(2) \\ -R(3) \\ -R(4) \\ -R(5) \\ -R(6) \\ -R(7) \\ -R(8) \\ -R(9) \\ -R(10) \end{bmatrix}$$

$$\begin{aligned} R(k) &= \sum_{n=0}^{159-k} s(n)s(n+k) \\ &= \text{autocorrelation of } s(n) \end{aligned}$$

LPC анализ

- Полученную систему уравнений можно решить следующими способами:
  - метод Гаусса
  - любой метод инвертирования матрицы (MATHLAB)
  - рекурсия Левинсона-Дурбина:

$$
\begin{aligned}
E^{(0)} &= R(0) \\
k_i &= [R(i) - \textstyle\sum_{j=1}^{i-1} \alpha_j^{(i-1)} R(i-j)]/E^{(i-1)} \qquad i = 1, 2, \ldots, 10 \\
\alpha_i^{(i)} &= k_i \\
\alpha_j^{(i)} &= \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)} \qquad j = 1, 2, \ldots, i-1 \\
E^{(i)} &= (1 - k_i^2) E^{(i-1)}
\end{aligned}
$$

Уравнения решаются для $i = 1, 2, \ldots, 10$, а затем

Для получения оставшихся трех параметров $(V/UV, G, T_0)$ решается $a_i = \alpha_i^{(10)}$ уравнение для обновлений:

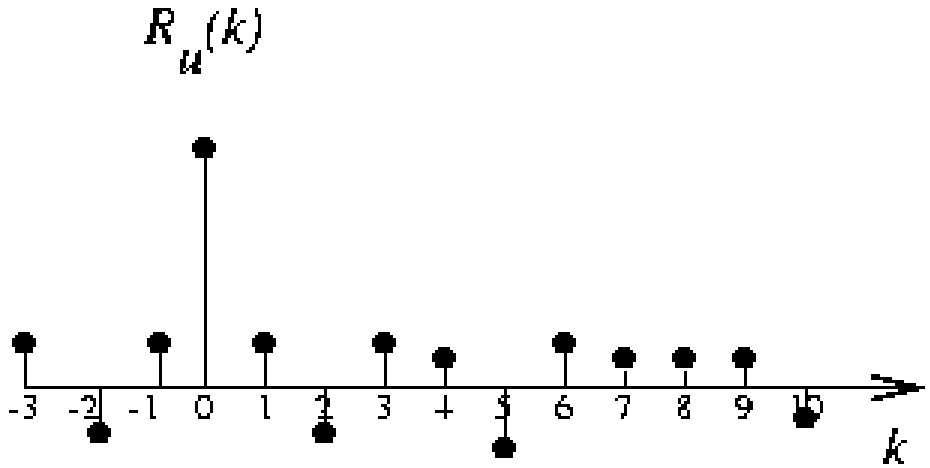
$$u(n) = s(n) + \sum_{i=1}^{10} a_i s(n-i)$$

Затем рассчитывают автокорреляцию $u(n)$:

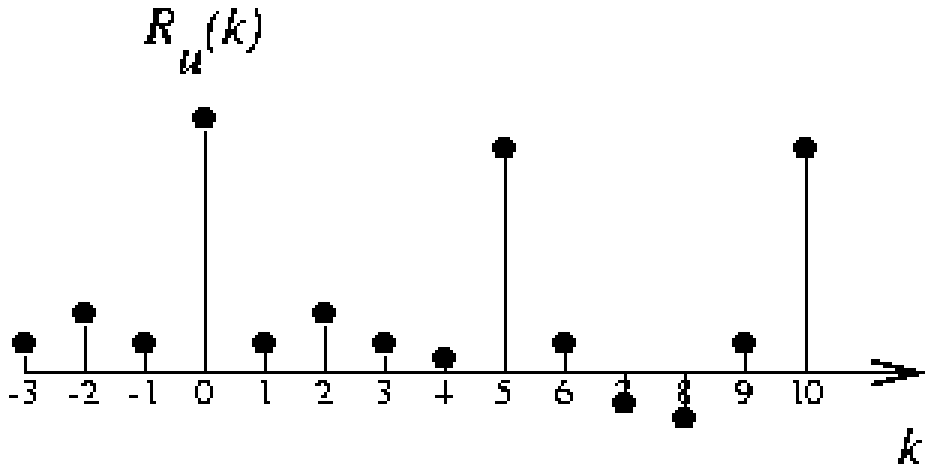
$$R_u(k) = \sum_{n=0}^{159-k} u(n)u(n+k)$$

Далее на основании автокорреляции принимается решение о виде звука (огласованный или неогласованный)

LPC анализ



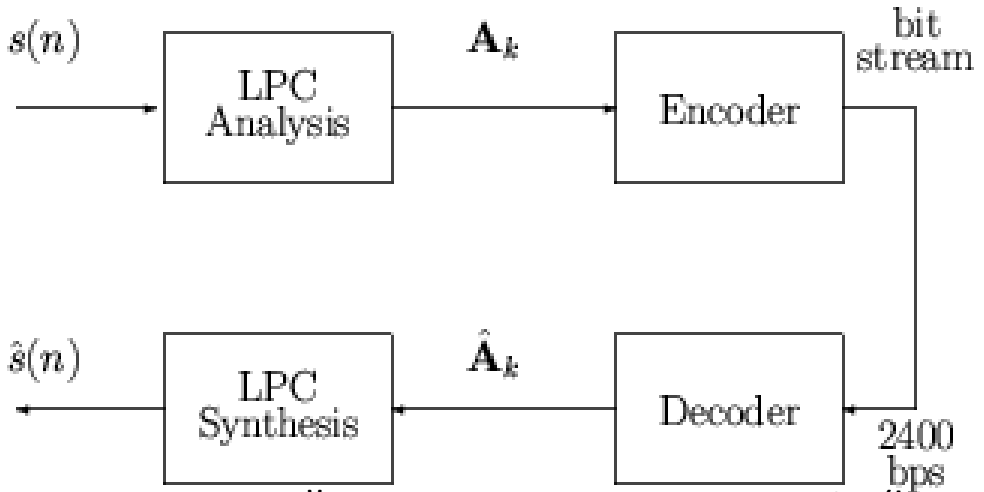
UV



V, T=5

Вокодер LPC 2,4 кбит/сек

- Блок-схема вокодера:



- Коэффициенты LPC представляются через линейные спектральные пары LSP (line spectrum pair)
- LSP математически эквивалентны коэффициентам LPC, но лучше подходят для процедуры квантования
- LSP вычисляются следующим образом:

$$\begin{aligned} P(z) &= 1 + (a_1 - a_{10})z^{-1} + (a_2 - a_9)z^{-2} + \ldots + (a_{10} - a_1)z^{-10} - z^{-11} \\ Q(z) &= 1 + (a_1 + a_{10})z^{-1} + (a_2 + a_9)z^{-2} + \ldots + (a_{10} + a_1)z^{-10} + z^{-11} \end{aligned}$$

- Факторизация этих уравнений дает:

$$\begin{aligned} P(z) &= (1 - z^{-1}) \prod_{k=2,4,\ldots,10} (1 - 2\cos\omega_k z^{-1} + z^{-2}) \\ Q(z) &= (1 + z^{-1}) \prod_{k=1,3,\ldots,9} (1 - 2\cos\omega_k z^{-1} + z^{-2}) \end{aligned}$$

- $\{\omega_k\}_{k=1}^{10}$ параметры LSP

Вокодер LPC 2,4 кбит/сек

- Параметры LSP упорядочены и ограничены:

$$0 < \omega_1 < \omega_2 < \ldots < \omega_{10} < \pi$$

- Они более коррелированы от кадра к кадру, чем коэффициенты LPC
- Размер кадра – 20 мсек, то есть 50 кадров в сек.
  2400 бит/сек соответствует 48 битам на кадр
- Распределение бит представлено в таблице:

Parameter Name	Parameter Notation	Rate (bits/frame)
LPC (LSP)	$\{a_k\}_{k=1}^{10}$ ($\{\omega_k\}_{k=1}^{10}$)	34
Gain	G	7
Voiced/Unvoiced & Period	$V/UV, T$	7
Total		48

- 34 бита LSP распределены в соответствии с таблицей:

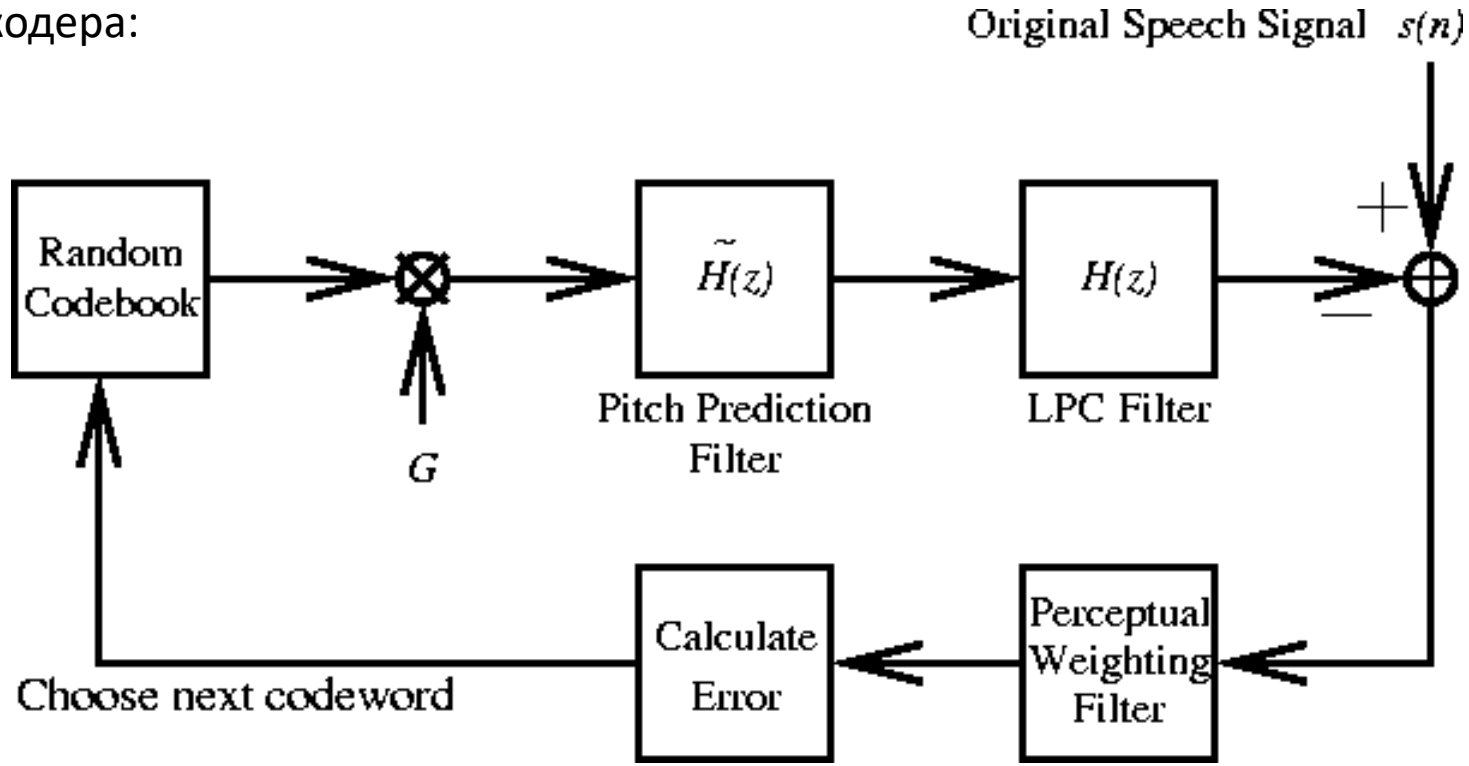
LSP	No. of Bits
ω_1	3
ω_2	4
ω_3	4
ω_4	4
ω_5	4
ω_6	3
ω_7	3
ω_8	3
ω_9	3
ω_{10}	3
Total	34

- Для усиления $\boldsymbol{G}$ используется 7-битный неоднородный скалярный квантователь
- Для огласованной речи величины $\boldsymbol{T}$ задаются в диапазоне от 20 до 146
- $\boldsymbol{V/UV}$, $\boldsymbol{T}$ совместно кодируются как показано в таблице:

V/UV	T	Encoded Value
UV	—	0
V	20	1
V	21	2
V	22	3
V	23	4
⋮	⋮	⋮
⋮	⋮	⋮
V	146	127

Кодер CELP 4,8 кбит/сек

- CELP – Code-Exited Linear Prediction – линейное предсказание с кодовым возбуждением
- Принципы кодирования аналогичны LPC, за исключением:
  - размер кадра – 30 мсек (240 отсчетов)
  - $\boldsymbol{u(n)}$ кодируются непосредственно
  - используется большее количество бит, более сложные вычисления
  - используется фильтр предсказания основного тона (pitch)
  - используется векторное квантование
- Блок-схема кодера:



Кодер CELP 4,8 кбит/сек

- Фильтр предсказания основного тона: $\tilde{H}(z) = \frac{1}{1 + bz^{-T}}$
- Фильтр перцептуального взвешивания: $W(z) = \frac{H(z/\gamma_2)}{H(z/\gamma_1)} \qquad \gamma_1 = 0.9, \gamma_2 = 0.5$
- Каждый кадр разделен на 4 подкадра. В каждом подкадре кодовая книга содержит 512 кодовых векторов
- Усиление передается 5-ю битами в каждом подкадре
- Параметры LSP передаются 34-мя битами аналогично вокодеру LPC
- При 30 мсек на кадр 4,8 кбит/сек соответствует 144 битам на кадр, распределенным следующим образом:

Parameters	No. of Bits
LSP	34
Pitch Prediction Filter	48
Codebook Indices	36
Gains	20
Synchronization	1
FEC	4
Future Expansion	1
Total	144

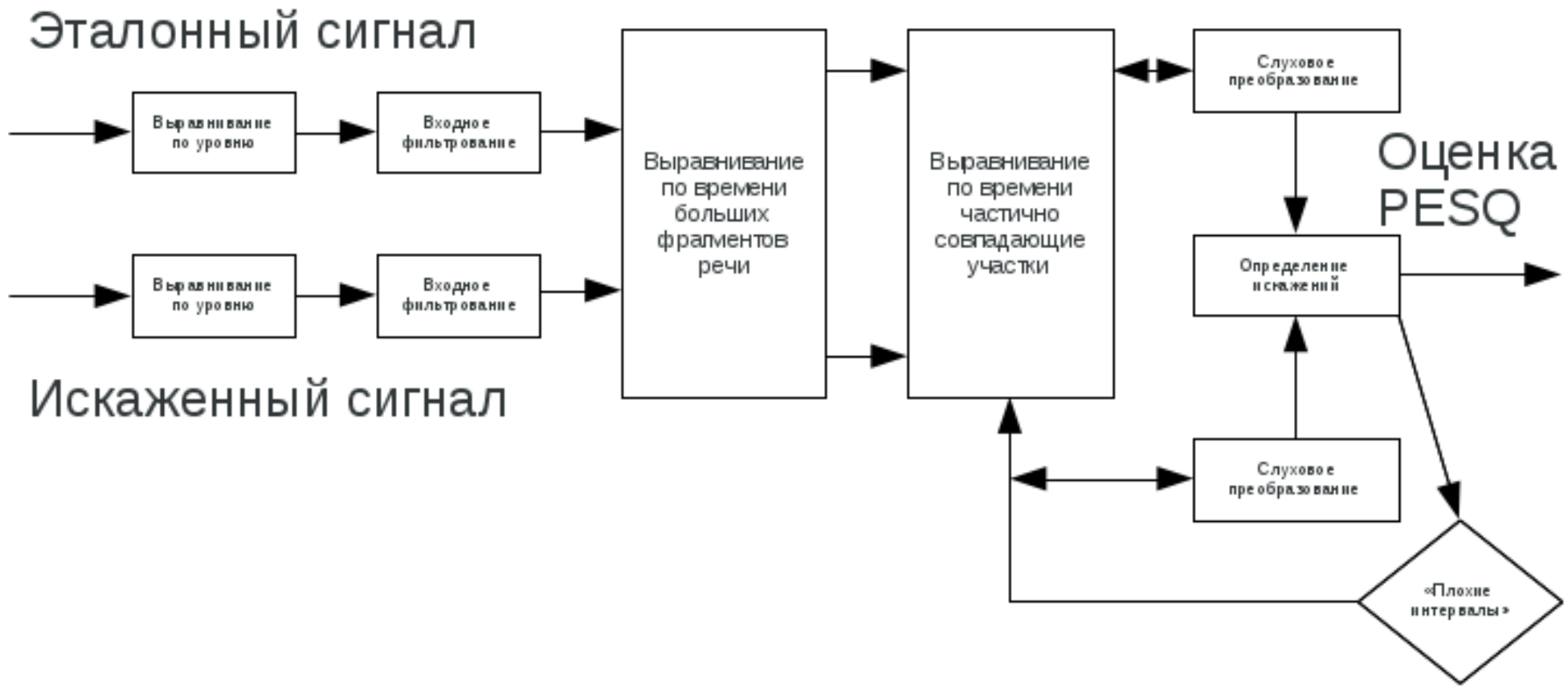
Кодер CS-ACELP 8,0 кбит/сек

- CS-ACELP – Conjugate-Structured Algebraic CELP – сопряженно-структурированное алгебраическое линейное предсказание с кодовым возбуждением
- Принципы кодирования аналогичны 4,8 кбит/сек CELP, за исключением:
  - размер кадра – 10 мсек (80 отсчетов)
  - кадр делится на два подкадра по 5 мсек (40 отсчетов)
  - параметры LSP кодируются с использованием двухстадийного векторного квантования
  - усиление также кодируется с использованием векторного квантования
- При 10 мсек на кадр 8 кбит/сек соответствует 80 битам на кадр, распределенным следующим образом:

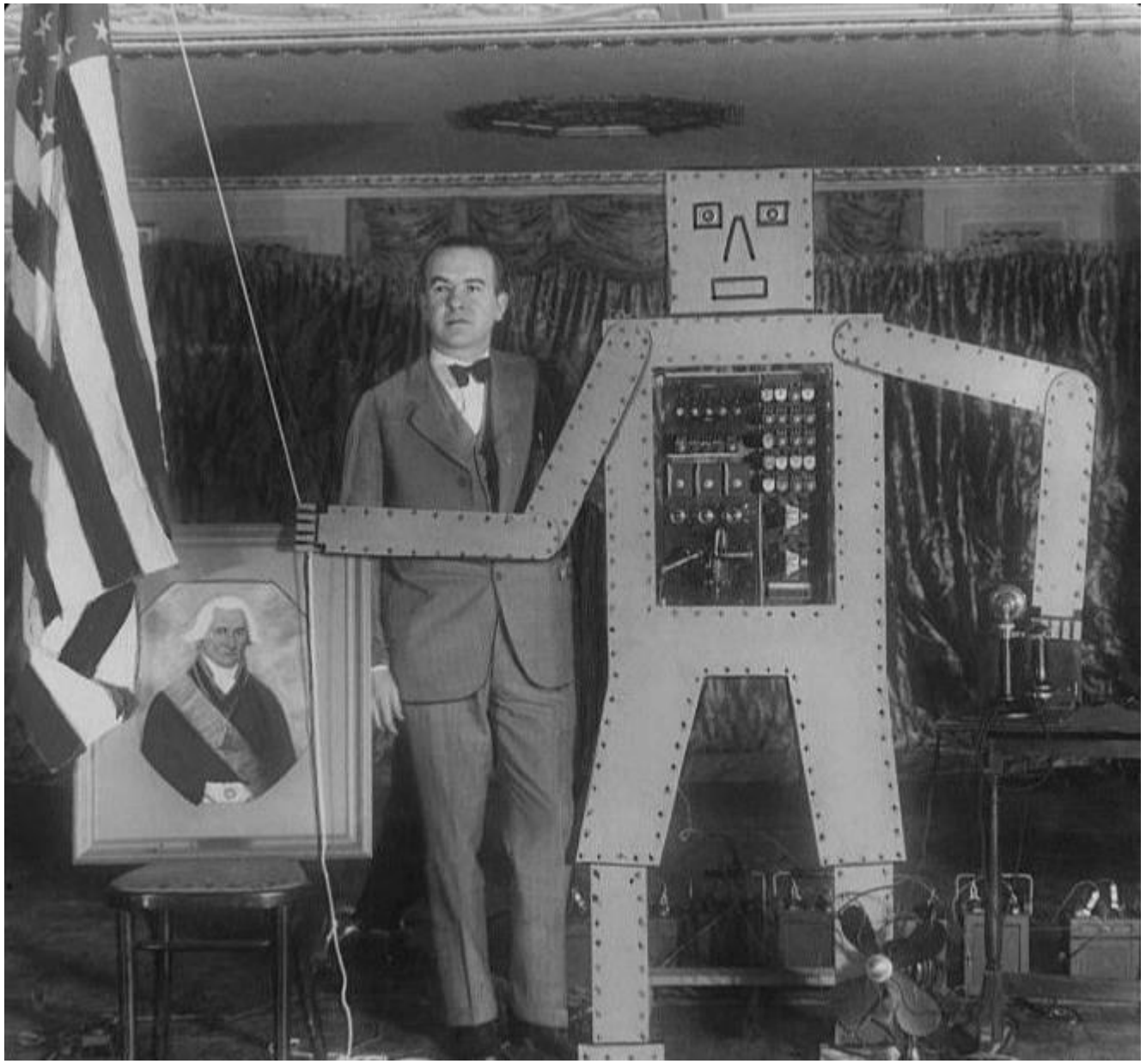
Parameters	No. of Bits
LSP	18
Pitch Prediction Filter	14
Codebook Indices	34
Gains	14
Total	80

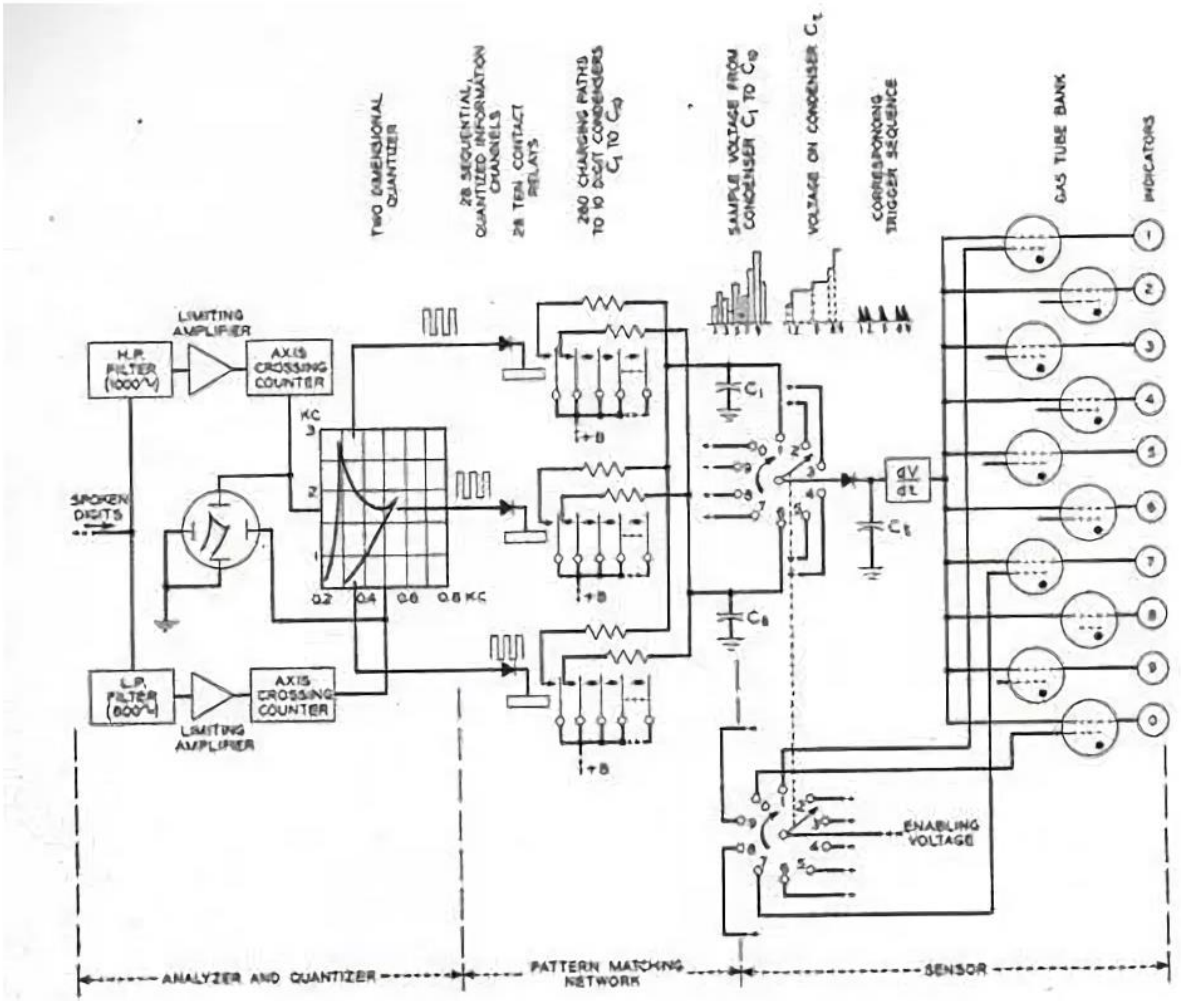
PESQ

Наиболее адекватная мера качества – MOS (mean opinion score).
Для определения качества передачи речи в PESQ (Perceptual Evaluation of Speech Quality) предусмотрено сравнение входного, или эталонного, сигнала с его искаженной версией на выходе системы связи.



Распознавание речи





https://www.youtube.com/watch?v=rQco1sa9AwU

Распознавание речи

30-е: аналоговые технологии на камертонах. Робот Телевокс. Общение свистом или короткими командами.

Машина Audrey, разработанная в 1952 году. Распознавала 10 цифр с точностью 70% от обученного диктора, 50% от случайных людей.

1961 год: IBM Shoebox, в видео, способна выполнять арифметические операции. Вводимые через микрофон слова сравнивались с базовым словарем.

К 1976 году Университет Карнеги-Меллона представил Harpy, способную оперировать словарём из 1011 слов. Harpy не сличала целиком услышанные слова с образцами, а разделяла их на аллофоны (образец звучания фонемы в зависимости от окружающих её букв). https://www.youtube.com/watch?v=32KKg3aP3Vw

1985 - IBM Tangora могла научиться понимать речь любого диктора с любым акцентом, диалектом и особенностями произношения, для этого лишь требовалась 20-минутная тренировка, в ходе которой накапливалась база образцов фонем и аллофонов. Работала на марковской модели.

В 1996 году появилась первая коммерческая программа, способная различать не отдельные слова, а беспрерывный поток естественной речи — IBM MedSpeak/Radiology. Продукт IBM был специализированным, он использовался в медицине для стенографирования описания результатов рентгенограммой, произносимых врачом в ходе исследования.

Первым универсальным движком распознавания естественной речи стала программа Dragon Naturally Speaking 1997-го года. При работе с нею диктору не требовалось проходить тренировку или оперировать определённым лексиконом, как в случае с MedSpeak, — с NaturallySpeaking мог работать любой человек.

Нейросетевые методики распознавание речи

В наше время лучшим средством для создания движка распознавания речи стала рекуррентная нейросеть (RNN), на которой построены все современные сервисы распознавания голоса, музыки, изображений, лиц, объектов, текста. RNN позволяет с высочайшей точностью понимать слова, а также предсказывать наиболее вероятное слово в рамках контекста, если оно не было распознано.

Рекуррентная нейросеть для распознавания речи хороша тем, что после длительной тренировки базой различных произношений она научится с высокой точностью различать фонемы и составлять из них слова вне зависимости от качества и характера произношения. И даже «додумывать» с высокой точностью в рамках контекста слова, которые не удалось распознать однозначно из-за фоновых шумов или нечёткого произношения.

Но с предсказаниями RNN есть нюанс — рекуррентная нейросеть может «додумать» пропущенное слово только опираясь на самый ближайший контекст примерно в пять слов.

Архитектура долгой краткосрочной памяти (Long short-term memory, LSTM) для рекуррентных нейросетей, созданная в 1997 году. Она специально разрабатывалась для того, чтобы добавить RNN умение учитывать контекст, удалённый от обрабатываемого события, — результаты решения предыдущих задач (то есть, распознаваний слов) проносятся сквозь весь процесс распознавания, сколь бы длинным не был монолог, и учитываются в каждом случае сомнений. Причём расстояние удаления почти не влияет на эффективность работы архитектуры.

Лабораторная работа по теме «Основы обработки речи в MATLAB»

Задача: измерить основной тон своего голосового тракта.

Рекомендации:

1) Используйте какую-нибудь стороннюю утилиту и запишите wav файл длиной примерно секунд 10-15 с звуком «а», максимально расслабив голосовые связки и открыв рот.

2) Вам понадобится изучить логарифм модуля спектра полученного сигнала и найти нужные форманты.

3) Не забудьте почистить речевой сигнал от ненужной информации, как высокочастотной, так и от резких колебаний спектра.

4) Голосовые связки имеют несколько гармоник, основной тон это первая. Используйте весовую функцию, чтобы убрать вторичные гармоники.

5) Мужской голос ниже женского.