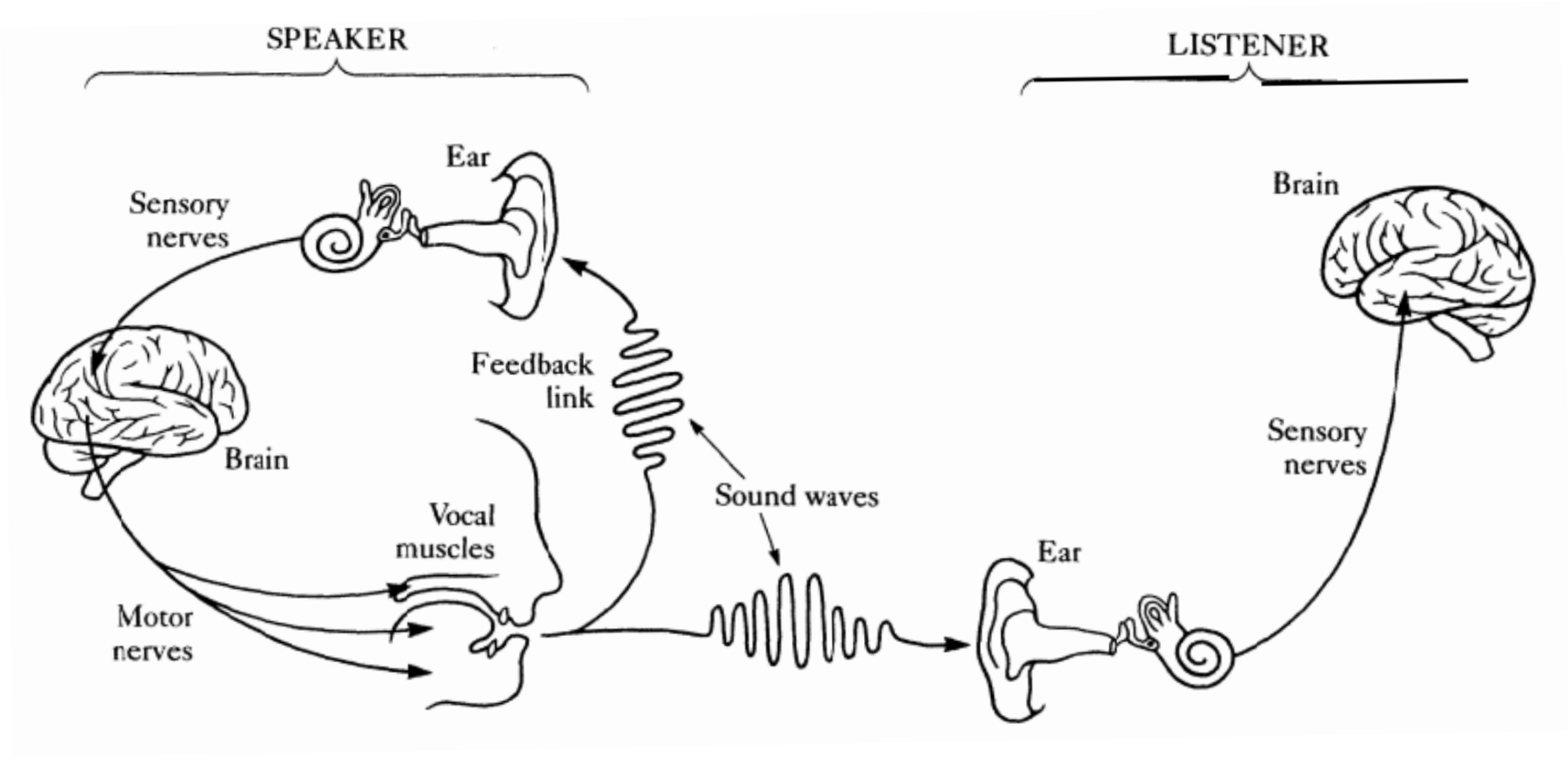
Речь

Биологические аспекты, форманты

Речевая цепь

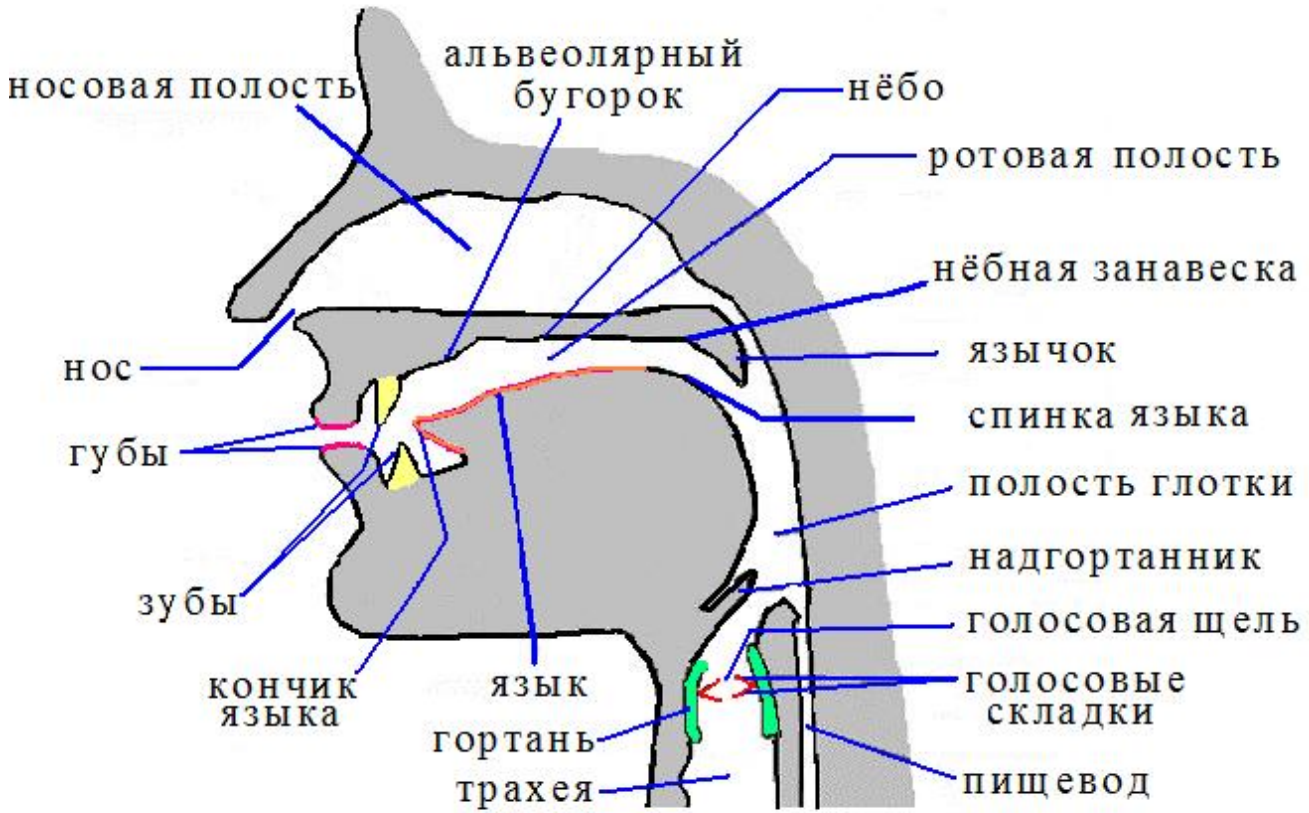


Denes and Pinson

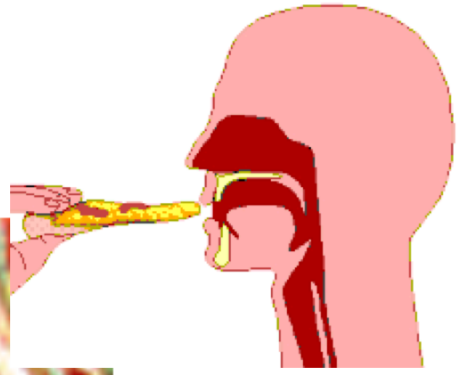
Речевой тракт человека

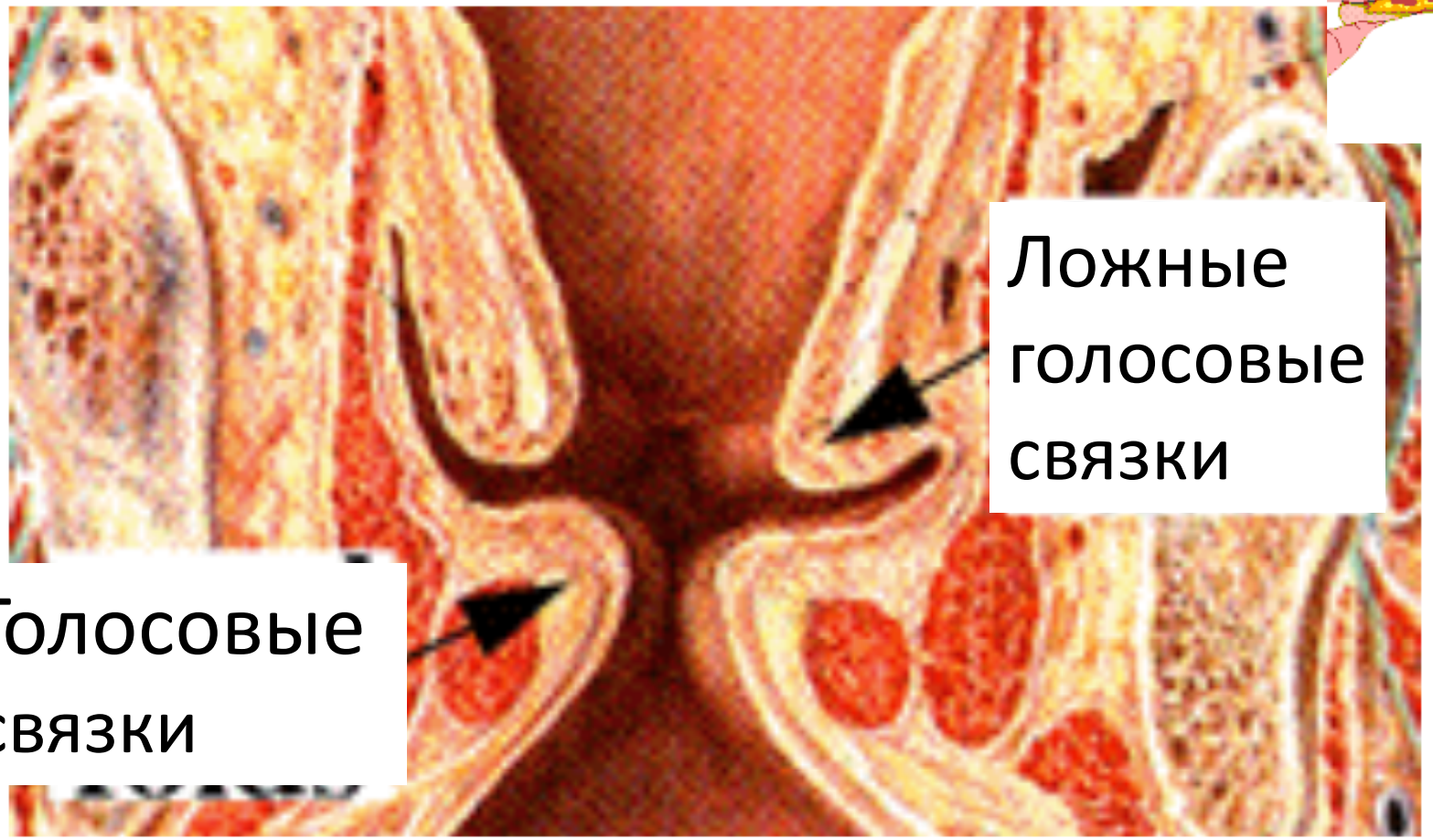
Речевого тракт как музыкальный инструмент:

- генератор — дыхательная система - резервуар легкие
- вибраторы — голосовые связки;
- резонаторы —резонансные полости -глотка, рот и нос, называемые артикуляционной системой.

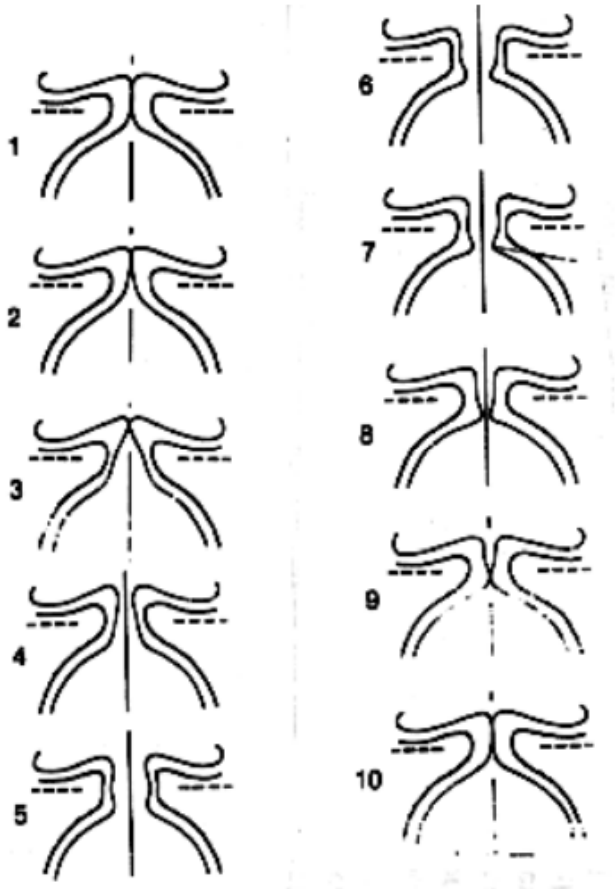


Вертикальный разрез гортани





Произношение



1. Перед началом должны быть сведены, что приводит к избыточному давлению

2,3,4 – Воздух прокладывает себе путь через голосовые связки, давя на них

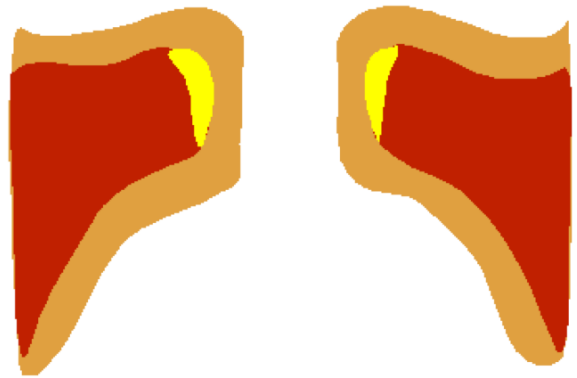
5. Проходя через щель давление начинает падать:
   - Газ проходит через суженый проход – скорость увеличивается
   - Увеличение скорости приводит к падению давления

6-10 Давление падает голосовые связки смыкаются

- Процесс повторяется снова. Цикл ~ 1/100 сек

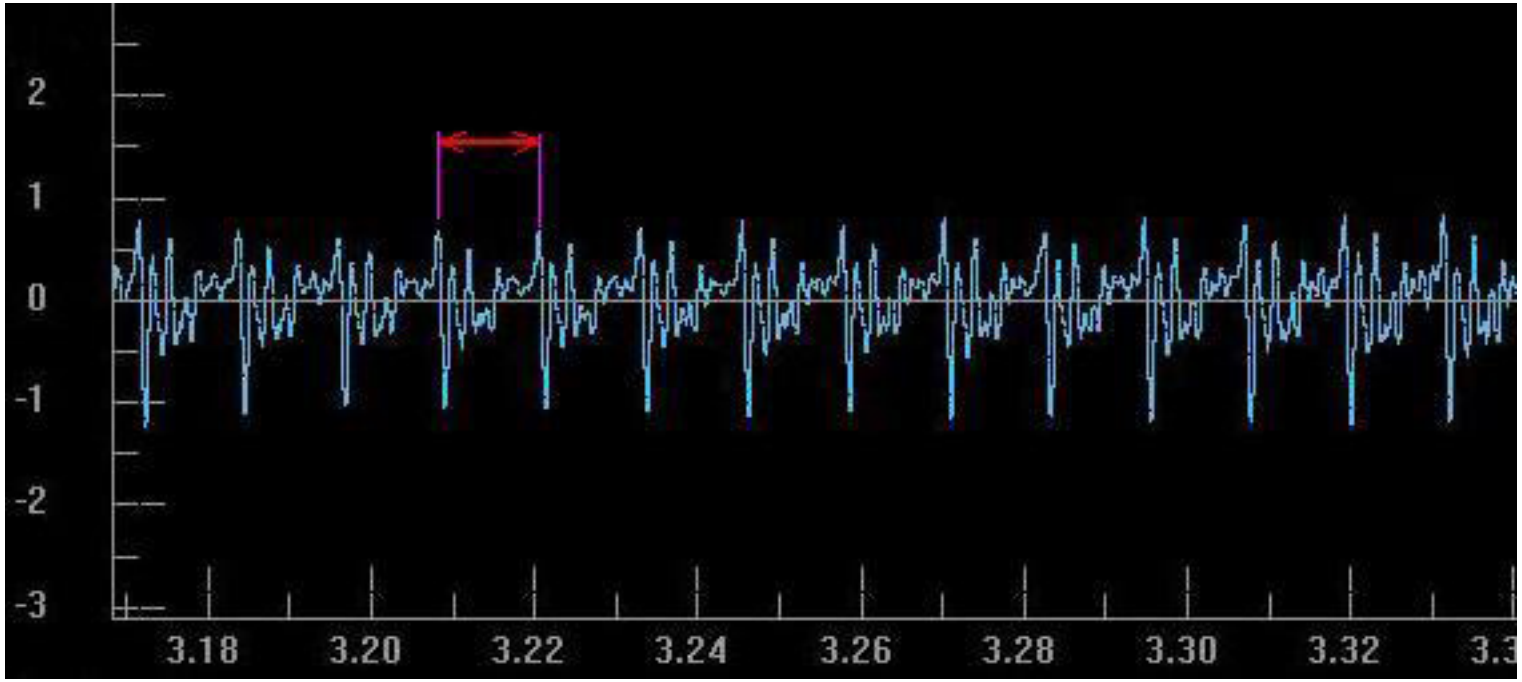
Частота основного тона

- Голосовой тракт открыт полностью, воздух проходит беззвучно
- Если остается небольшая щель, скорость воздуха не падает до нуля, и слышен шум (шёпот)
  - Беззвучные звуки п / т / к / с / ф / ш/ щ / ч
- Частота колебаний голосовых связок называется частотой основного тона



https://youtu.be/P2pLJfWUjc8

Основной тон



Голосовыми связками формируются вокализированные звуки, которые трансформируются в речевом тракте

Осциллограмма голоса - звук [а]. По вертикальной оси - амплитуда (в отсчетах), по горизонтальной — время (в секундах). Высокие по амплитуде пики обозначают время начала раскрывания голосовых складок. 8 циклов (периодов) на интервале [3.22, 3.32] с, т.е. 80 периодов в 1 с, следовательно, частота ОТ для данного диктора ~ 80 Гц.

Согласные и гласные

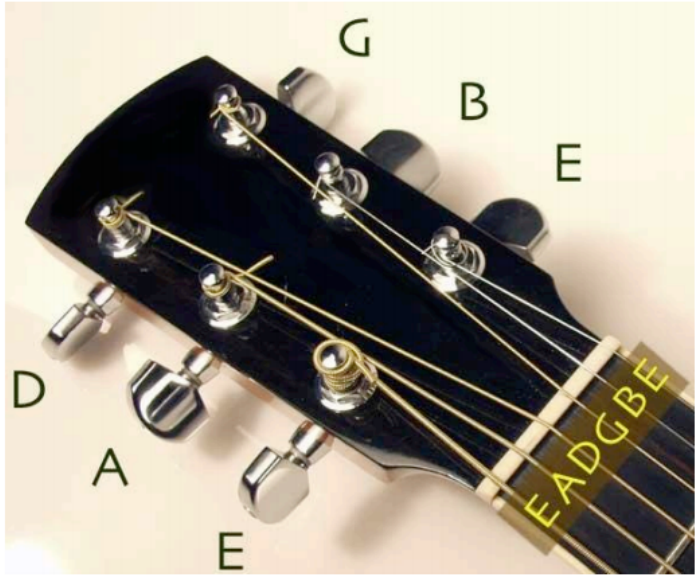
- Гласные – производятся из вокализированных звуков изменением формы речевого тракта без препятствий воздушному потоку
- Согласные – воздушный поток встречает препятствия на пути в речевом тракте
  - *Фрикативные* ([в],[ф],[з],[щ],[ж]) образуются при форсированном прохождении звука через сужения речевого тракта.
  - *Взрывные* звуки ([р],[с],[т],[к]) образуются вследствие полного перекрытия речевого тракта, создания большого давления перед этим барьером и последующего резкого снятия препятствия.

Основная частота (F0)

- Гитара
  У гитар есть струны. Каждая струна имеет разную толщину. Это заставляет каждую струну вибрировать с разной частотой, что приводит к разным тонам (который в музыке мы называем нотами):

(F0): каждый объект имеет свою основную частоту, это частота, с которой этот объект вибрирует.

E	329.6 Hz
B	246.9 Hz
G	196 Hz
D	146.8 Hz
A	110 Hz
E	82.4 Hz



Основная частота (F0)

Голос человека

- Ваш голос также имеет основную частоту (F0). Он создается вашими голосовыми связями (которые на самом деле являются двумя закрылками или мембранами, которые вибрируют):

Для мужчин это около 130 Гц (До малой октавы)
для женщин это около 220 Гц (До малой октавы – разница почти в октаву!)
BTW: До первой октавы - 261,6 Гц

Stroboscopical imaging of the vocal fold movement using the LED laryngoscope

Гармоники

- Когда объект вибрирует на своей собственной частоте, гармоники также активируются:

F0	100 Hz	F0	200 Hz	F0	400 Hz
1st	200 Hz	1st	400 Hz	1st	800 Hz
2nd	300 Hz	2nd	600 Hz	2nd	1200 Hz
3rd	400 Hz	3rd	800 Hz	3rd	1600 Hz
4th	500 Hz	4th	1000 Hz	4th	2000 Hz
5th	600 Hz	5th	1200 Hz	5th	1400 Hz

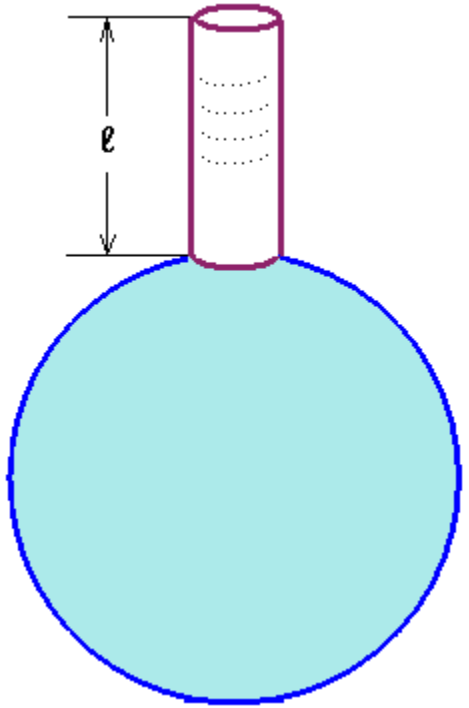
Резонанс

- Простой пример из жизни – почему акустическая гитара звучит громче электрогитары.
- Любая масса воздуха заключенная в ограниченный объем является акустическим резонатором, т.е. колебательной системой, имеющей свои собственные частоты колебаний

Резонанс

- Акустический резонатор или резонатор Гельмгольца

Упрощенная модель резонатора учитывает колебания воздуха только в горловине, поскольку эти резонаторы имеют применения только в слышимом диапазоне частот

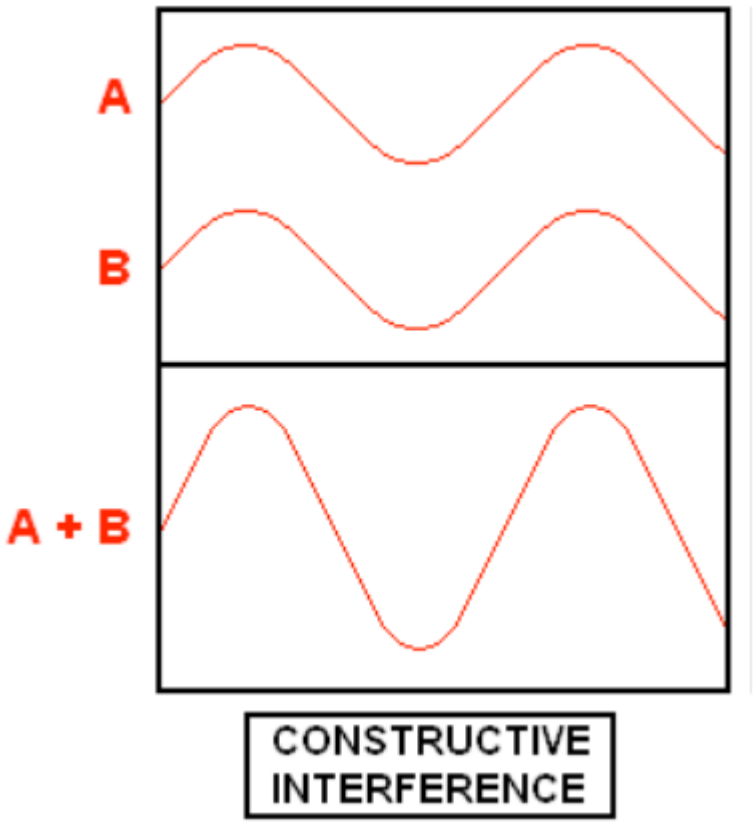


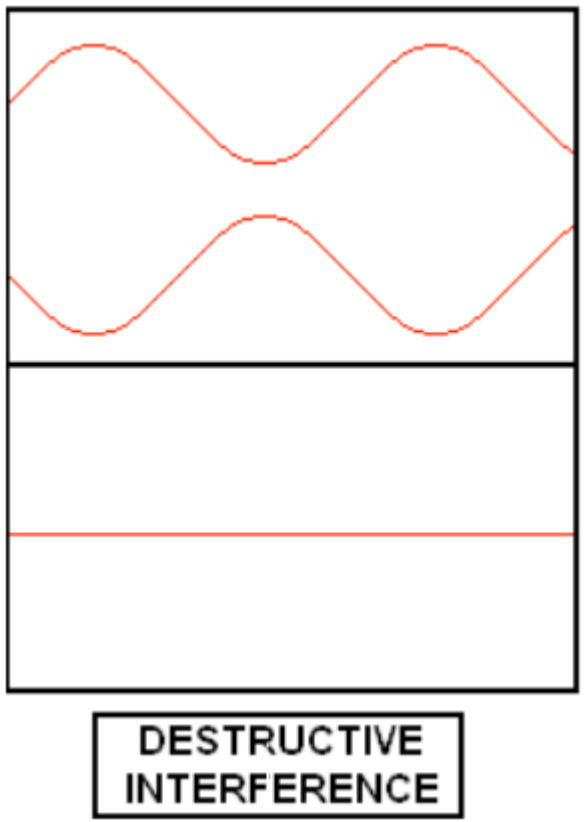
В этом предположении собственная частота колебаний воздуха в горловине (или частота резонатора Гельмгольца) равна

$$\omega_0 = c\sqrt{\frac{S}{Vl}}$$

где c — скорость звука, S— площадь поперечного сечения горловины, l — ее длина, V — объем колбы резонатора.

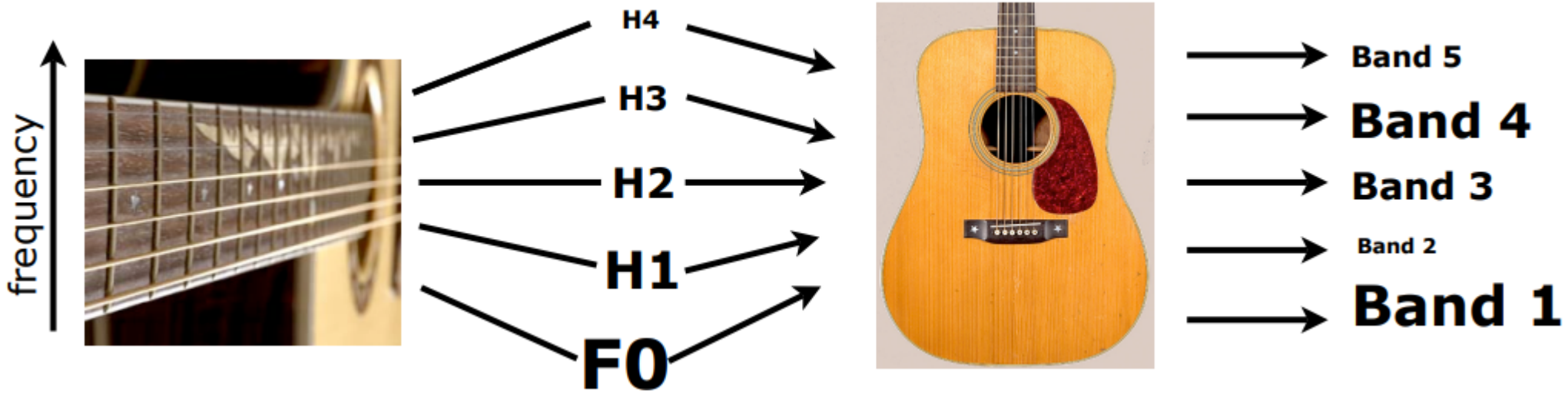
Резонанс


A
B
A + B
CONSTRUCTIVE
INTERFERENCE

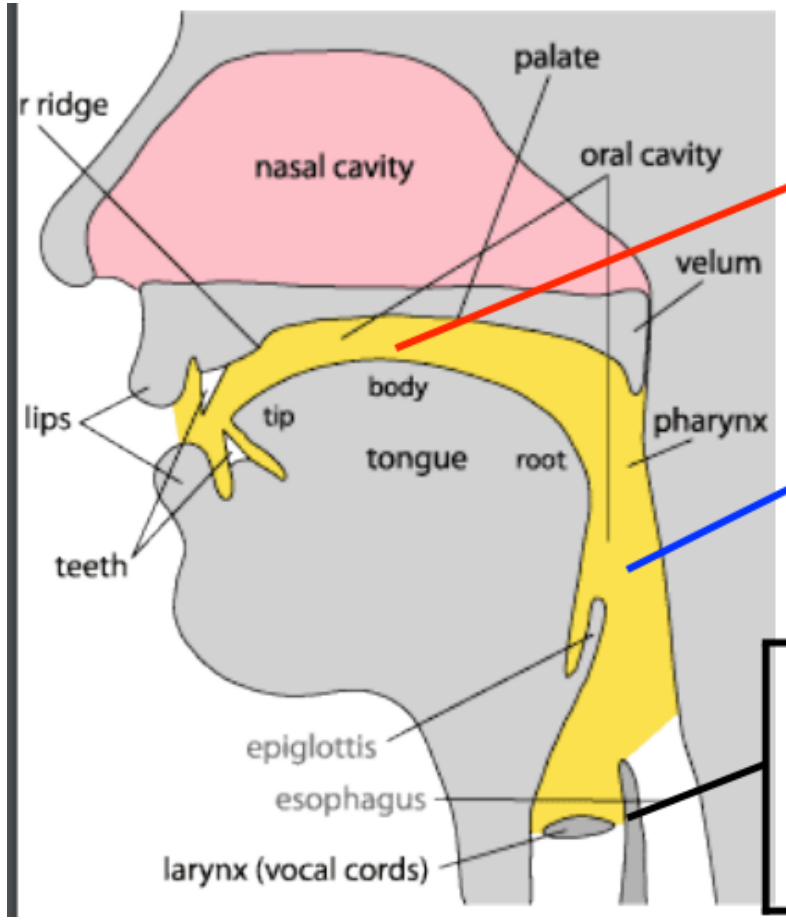

DESTRUCTIVE
INTERFERENCE

Резонансные полосы

- В жизни очень редко, происходит точное совпадение частот. Говорят о частотах, которые близки к собственным частотам воздушного столба в тракте.
- Например, резонирование вокруг гармоники на частоте 440 Гц, возможно в полосе 400-480 Гц.



Форманты



Band 5 = F2 (the second formant)
Band 4
Band 3
Band 2
~~Band 1~~

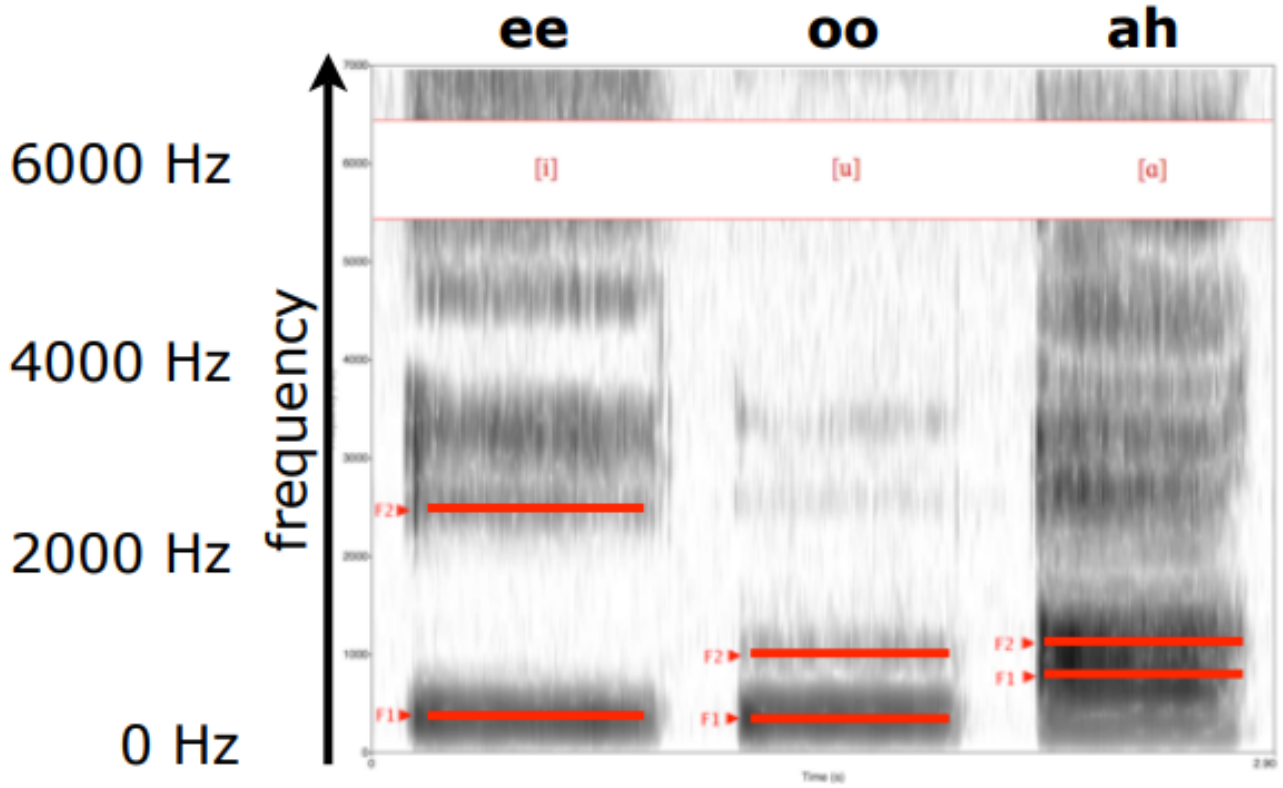
Band 5
Band 4 = F1 (the first formant)
Band 3
Band 2
~~Band 1~~

H4
H3
H2
H1
F0

We call these FORMANTS because they are the frequency bands that are used to "form" the speech sound

Оценка формант

- Самый простой способ оценить форманты – визуализация спектрограммы

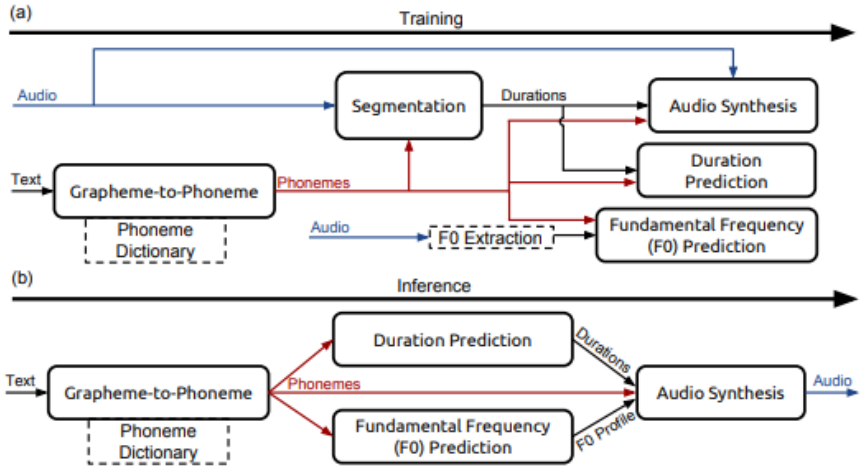


Pitch (тональность звука) это не частота

- Тональность звука - это психологическое ощущение или перцепция, скоррелированная с F0
  - Связь между тональностью звука и F0 не является линейной;
- восприятие тона человеком является наиболее точным между 100Гц и 1000Гц.
  - Линейные в этом диапазоне
  - Логарифмические выше 1000Hz
- *Mel* шкала - это одна из моделей отображения F0-в тон
- *Mel* является единицей тона определенной таким образом, что пара звуков, которые являются перцептивно равноотстоящим в тональностях разделяются равным числом mel.
- Частота в *mels* = 1127 * ln * (1 + f / 700)

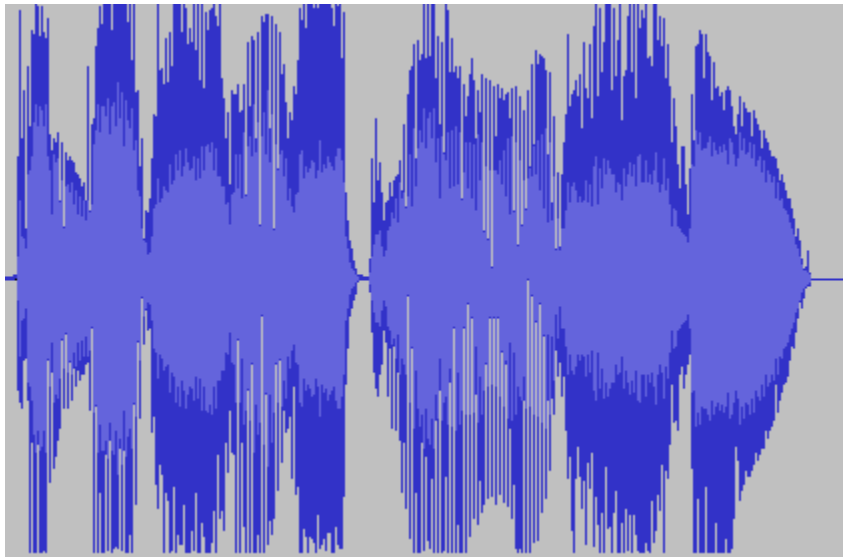
F0 оценка

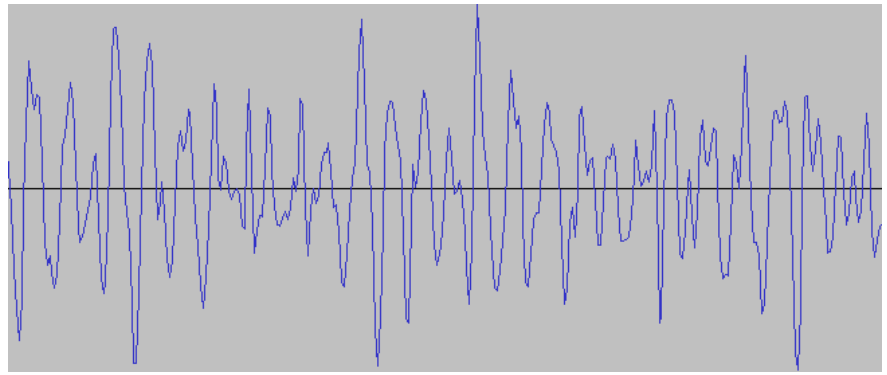
- Где может пригодится
- В синтезе речи, предсказываем параметры длительность и F0, по которым синтезируем речь
- В оценке музыкальных произведений - автоматическое транскрибирование
- Метаданные для индексирования мультимедиа



F0 оценка.

- Речь это нестационарный сигнал
- Основной тон может достаточно сильно меняться во времени
- Дополнительный шум, очень сильно влияет на оценку
- Резонанс речевого тракта может усиливать гармоники, что может повлиять на оценку
- Амплитуда голоса может сильно меняться в большом диапазоне
- Речевой сигнал включает фрагменты тишины (unvoiced)
- Все эти факторы делают оценку основного тона сложной задачей.





Как измерить ошибку?

- Что есть ground-truth?
  - Данные размеченные в ручную.
  - Запись при помощи ларингографа.
  - Доступны публичные базы данных
    - PTDB-TUG
    - TIMIT
    - SPEECON
- Типы ошибок:
  - VDE – Voice Detection Error. Доля правильно распознанных фрагментов с голосом или без
  - FFE – F0 Frame Error. Доля фреймов с правильно распознанной частотой. Не должна отличаться более чем на 20% в вокализированных фрагментах

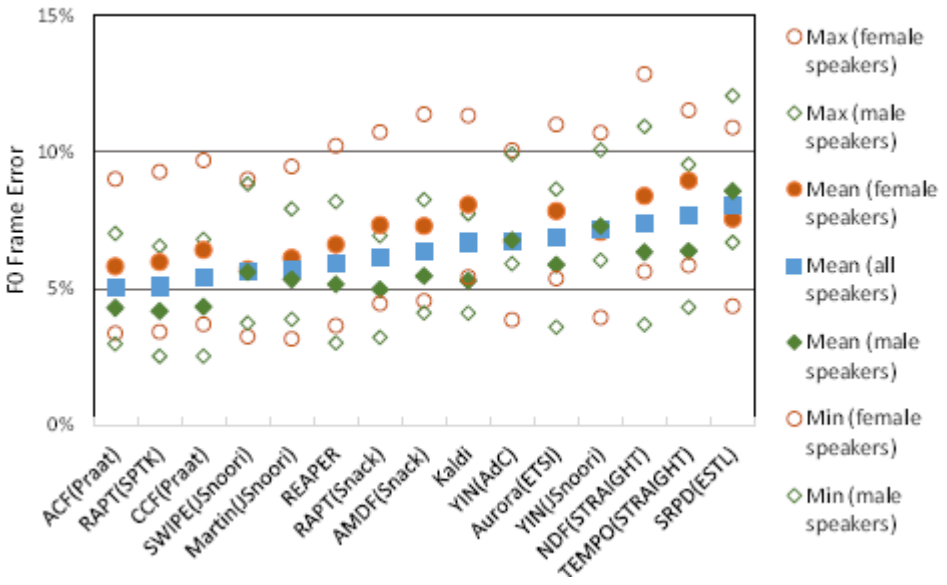
Задача оценки F0

- Из-за сложности оценки нет единого алгоритма.

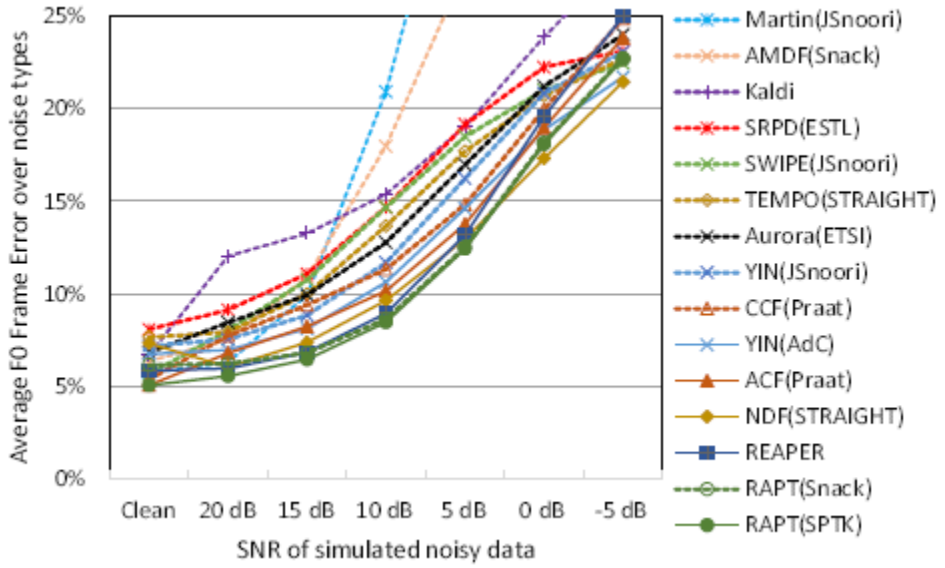
Name in this paper	Algorithm	Toolkit
ACF (Praat)	ACF [1]	Praat[a][25]
AMDF (Snack)	AMDF [19]	Snack library[b]
Aurora (ETSI)	Aurora [7]	ETSI[c]
CCF (Praat)	CCF	Praat[a]
Kaldi	enhanced RAPT [9]	Kaldi[d]
Martin (JSnoori)	Spectral-based [22]	JSnoori[e]
NDF (STRAIGHT)	NDF [8]	STRAIGHT[f]
REAPER	REAPER	REAPER[g]
RAPT (SPTK)	RAPT [2]	SPTK[h]
RAPT (Snack)	RAPT [2]	Snack library[b]
SHS (Praat)	SHS [24]	Praat[a]
SRPD (ESTL)	SRPD [20], [21]	ESTL[i]
SWIPE (JSnoori)	SWIPE [6], [23]	JSnoori[e]
SWIPE (SPTK)	SWIPE [6], [23]	SPTK[h]
TEMPO (STRAIGHT)	TEMPO [4], [5]	STRAIGHT[f]
YIN (AdC)	YIN [3]	YIN[j]
YIN (JSnoori)	YIN [3]	JSnoori[e]

Performance Analysis
of Several Pitch Detection Algorithms
on Simulated and Real Noisy Speech Data. Denis Jouvet and Yves Laprie

Задача оценки F0



F0 Frame Error на чистых данных

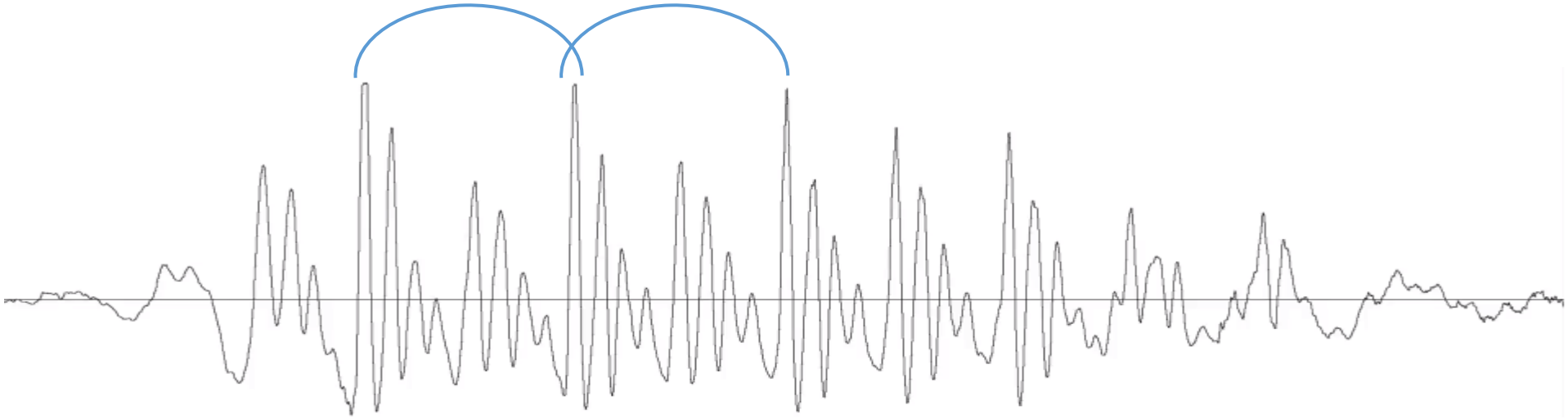


F0 Frame Error с добавление шума

Performance Analysis
of Several Pitch Detection Algorithms
on Simulated and Real Noisy Speech Data. Denis Jouvet and Yves Laprie

F0 оценка

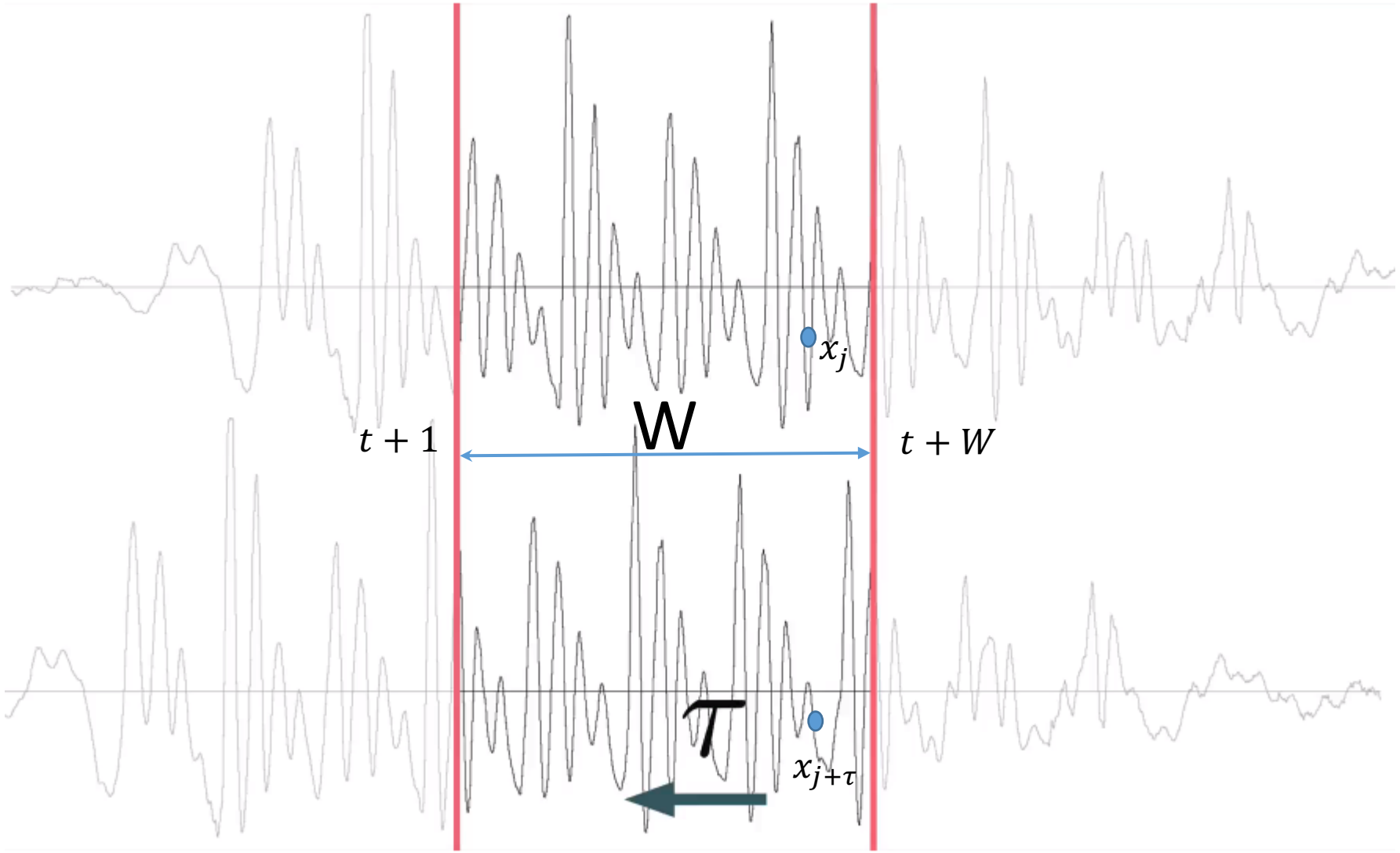
- Для оценки F0 нужно найти частоту колебания голосовых связок
- При увеличении осциллограммы волны можно увидеть периоды колебаний
- Мы можем оценить период через автокорреляцию. Подобие фрагментов должно быть максимально



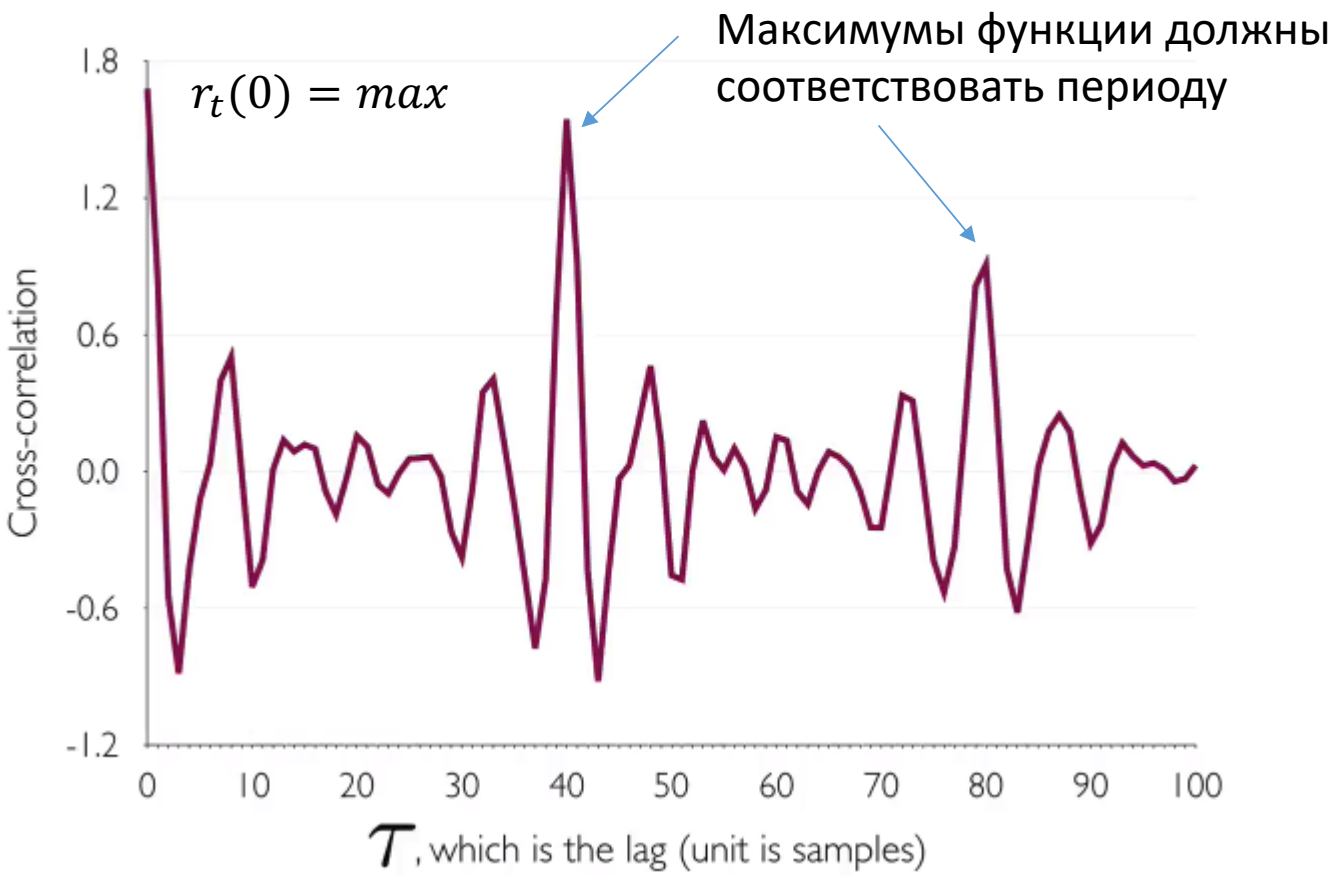
Оценка через автокорреляцию

- $r_t(\tau) = \sum_{j=t+1}^{t+W} x_j x_{j+\tau}$
- $r_t(\tau) -$ функция автокорреляции
- t – индекс времени
- W – размер окна сэмплирования

Оценка через автокорреляцию



Оценка через автокорреляцию

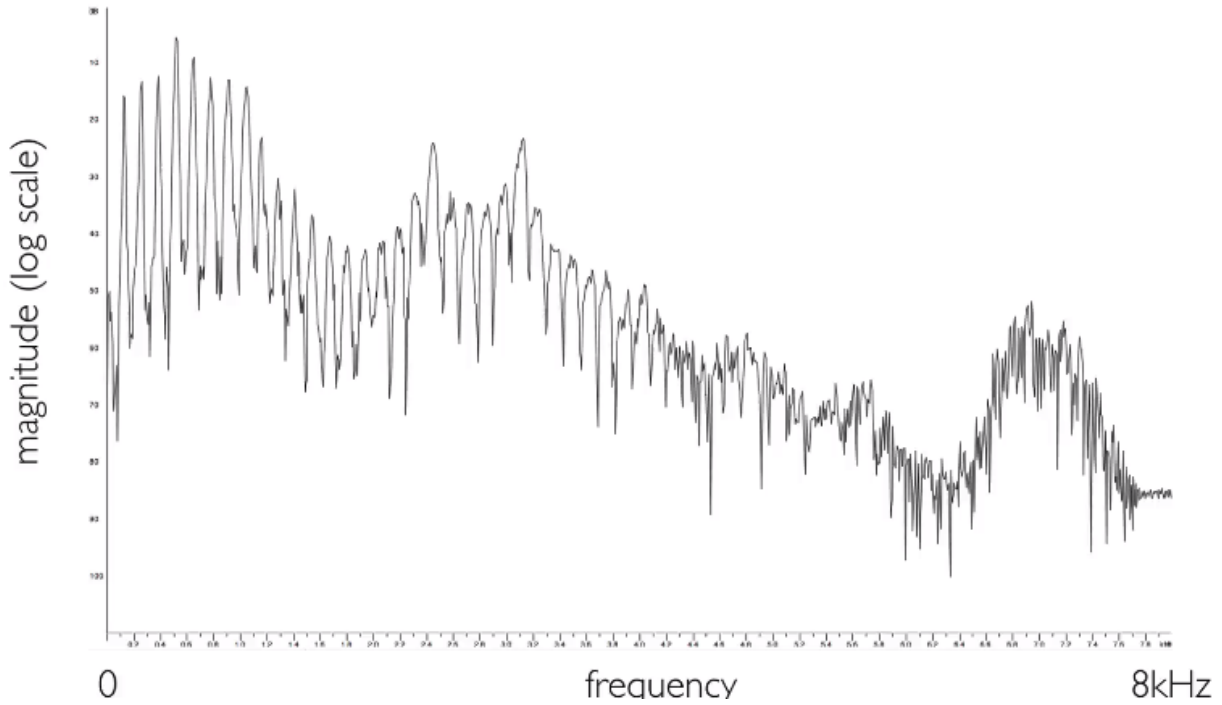


Оценка через автокорреляцию

- Ищем пики в автокорреляционной функции
- Самый большой пик в нуле,
- Ищем следующий максимальный пик с лагом не равным 0 это период, максимумы должны повторятся с заданным периодом
-
- Как всегда это не так легко как звучит
  - Реальный сигнал не точно периодический
  - Форманты могут приводить к искажению в оценке периода

Автокорреляции не достаточно

- Можно добавить фильтр высоких частот
- Можно комбинировать со снижением частоты дискретизации для уменьшения сложности вычислений



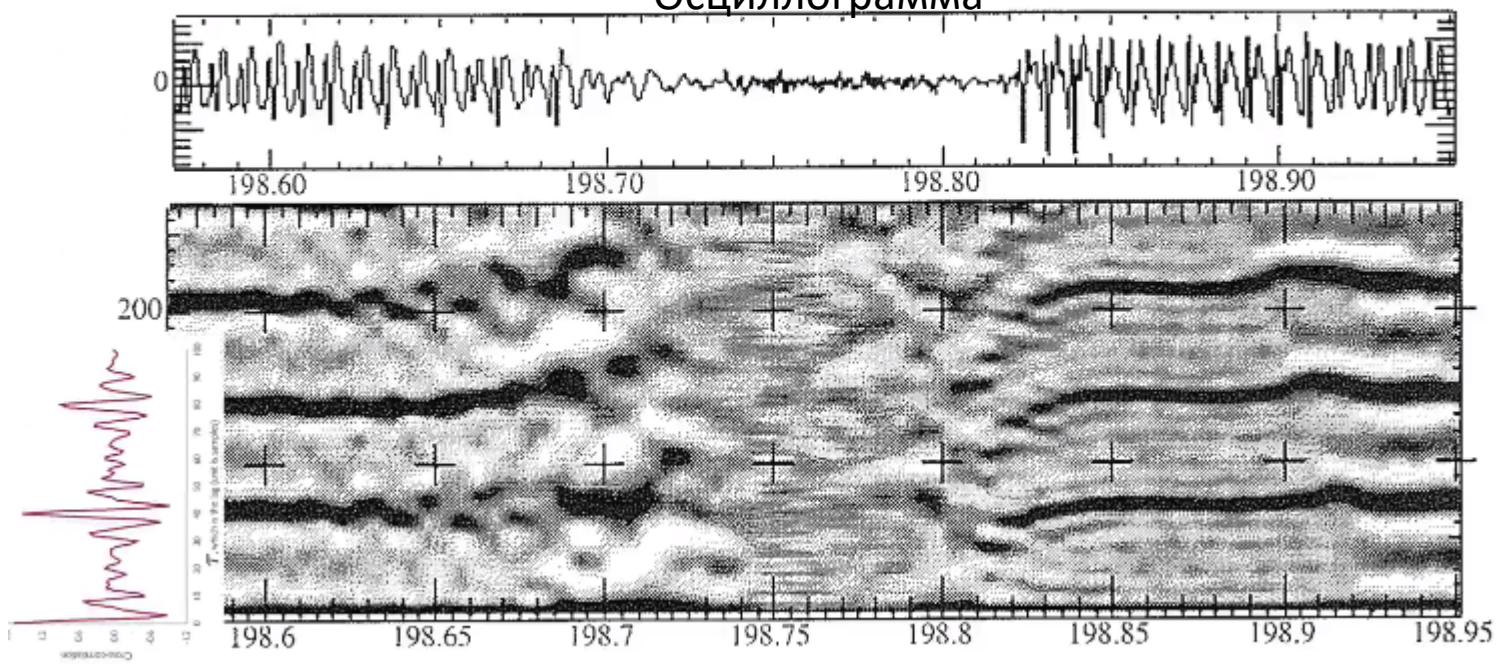
Pitch Tracking (RAPT)

A Robust Algorithm for Pitch Tracking (RAPT)

David Talkin

sampling rate = 8kHz

Осциллограмма



$r_t(\tau)$

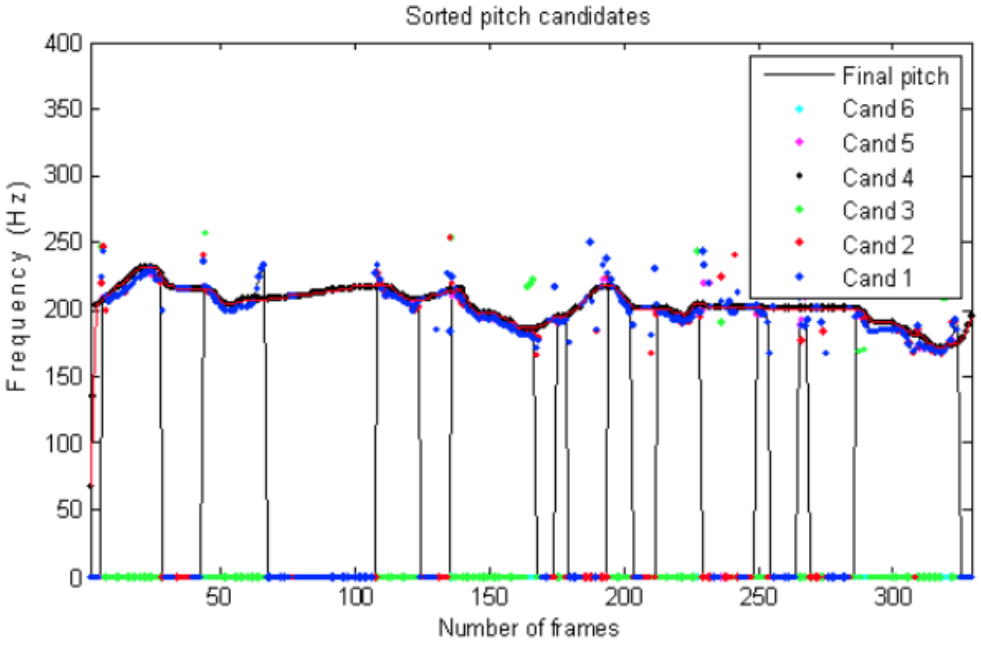
t

Pitch Tracking (RAPT)

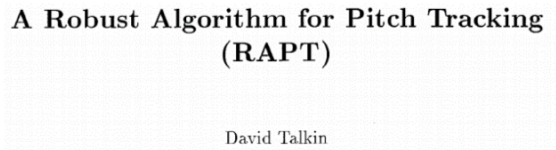
A Robust Algorithm for Pitch Tracking (RAPT)

David Talkin

- Даунсамплим основной трэк
- Считаем периоды по треку с низким сампл рейтом. Оцениваем через нормализованную функцию кросс корреляции $r_t(\tau) = \frac{\sum_{j=t+1}^{t+W} x_j x_{j+\tau}}{\sqrt{e_t e_{t+\tau}}}$; $e_j = \sum_{l=j}^{j+W} s_l^2$
- Если мы нашли пики на треке с низким рейтом, то пересчитываем их в окрестности по треку с высоким рейтом
- Каждый пик генерирует кандидата. Каждый фрейм генерирует гипотезу unvoiced
- Используют динамическое программирование, что бы выбрать набор пиков или unvoiced фреймов, которые лучше удовлетворяют условиям алгоритма.



Pitch Tracking (RAPT)

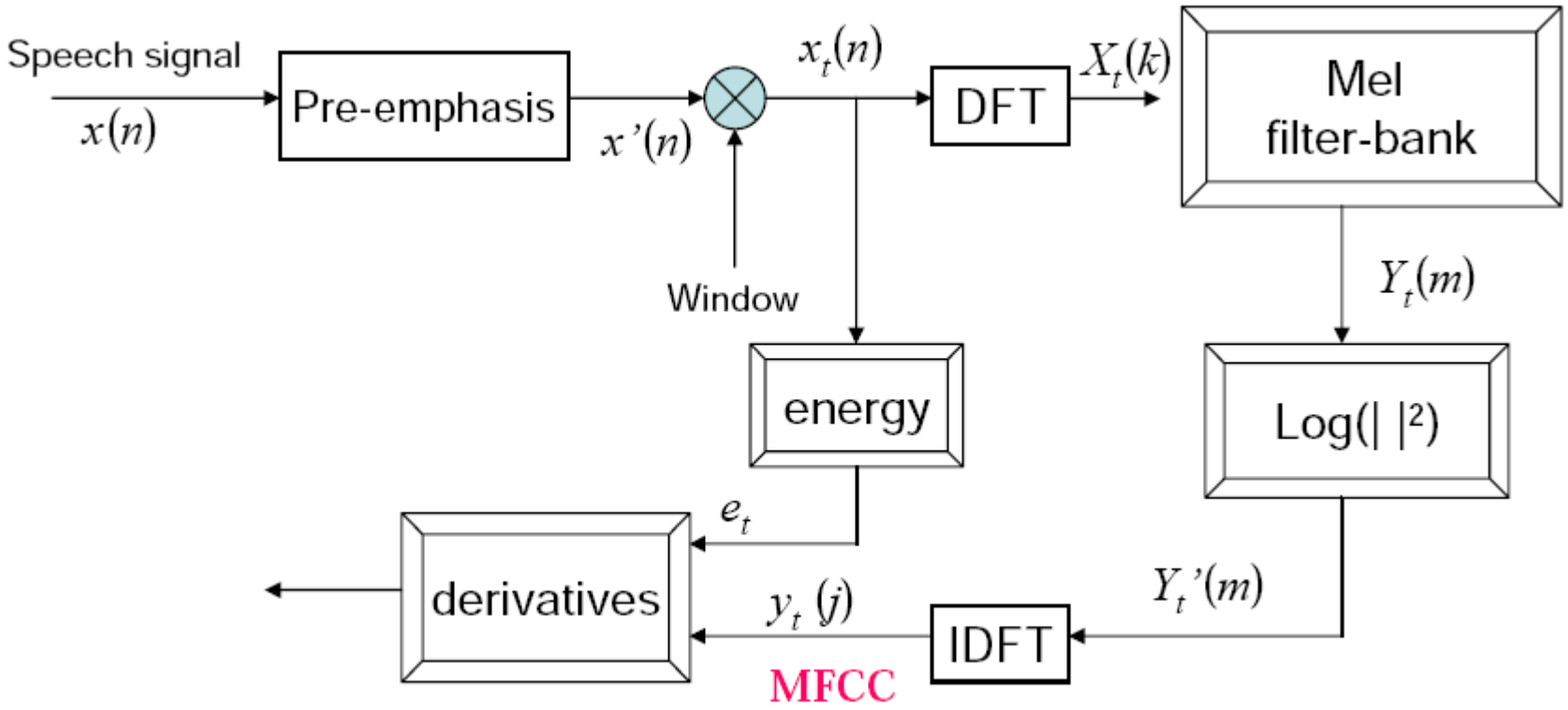

A Robust Algorithm for Pitch Tracking (RAPT)

David Talkin

- Пре-процессинг + автокорреляция + процессинг
- Типичная архитектура для алгоритмов F0
- Есть огромное количество параметров для настройки

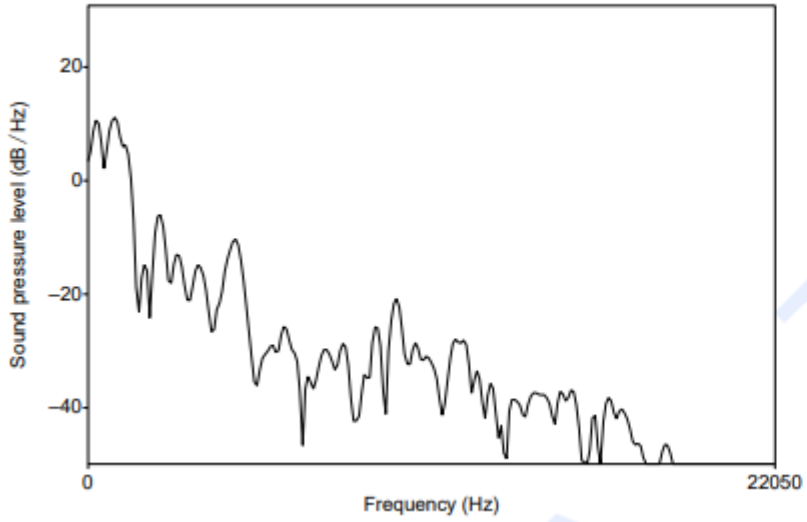
Constant	Meaning	Value
$F0_{min}$	minimum F0 to search for (Hz)	50
$F0_{max}$	maximum F0 to search for (Hz)	500
t	analysis frame step size (sec)	.01
w	correlation window size (sec)	.0075
CAND_TR	minimum acceptable peak value in NCCF	.3
LAG_WT	linear lag taper factor for NCCF	.3
FREQ_WT	cost factor for F0 change	.02
VTRAN_C	fixed voicing-state transition cost	.005
VTR_A_C	delta amplitude modulated transition cost	.5
VTR_S_C	delta spectrum modulated transition cost	.5
VO_BIAS	bias to encourage voiced hypotheses	0.0
DOUBL_C	cost of exact F0 doubling or halving	.35
A_FACT	term to decrease ϕ of weak signals	10000
N_CANDS	max. number of hypotheses at each frame	20

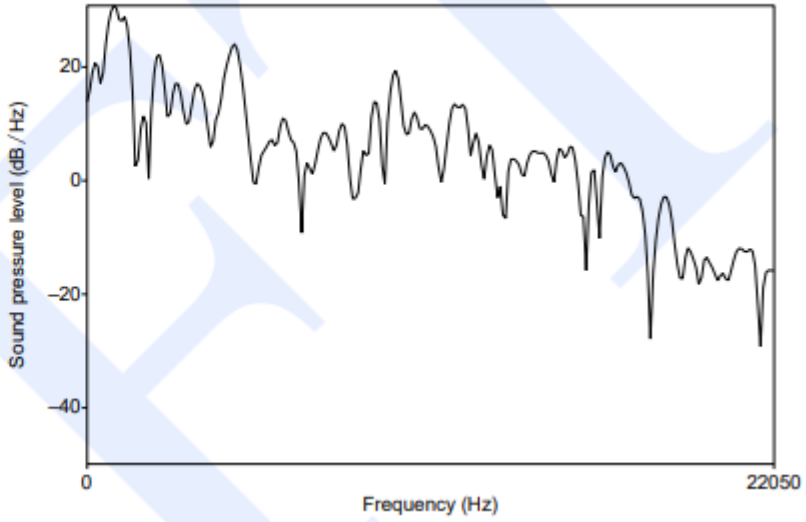
MFCC (Mel-Frequency Cepstral Coefficient)



Preemphasis

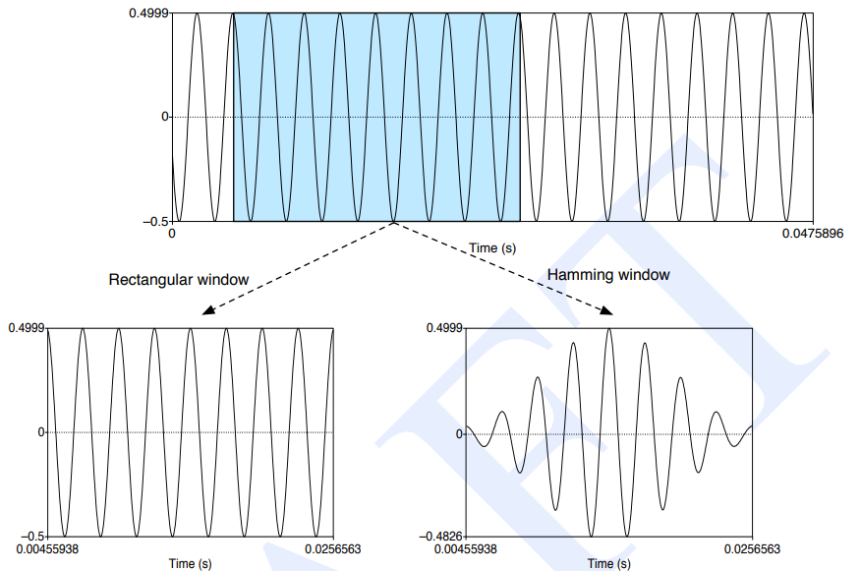
- High-pass фильтр $y[n] = x[n] - \alpha x[n-1]$.
- Усиливает высокие частоты

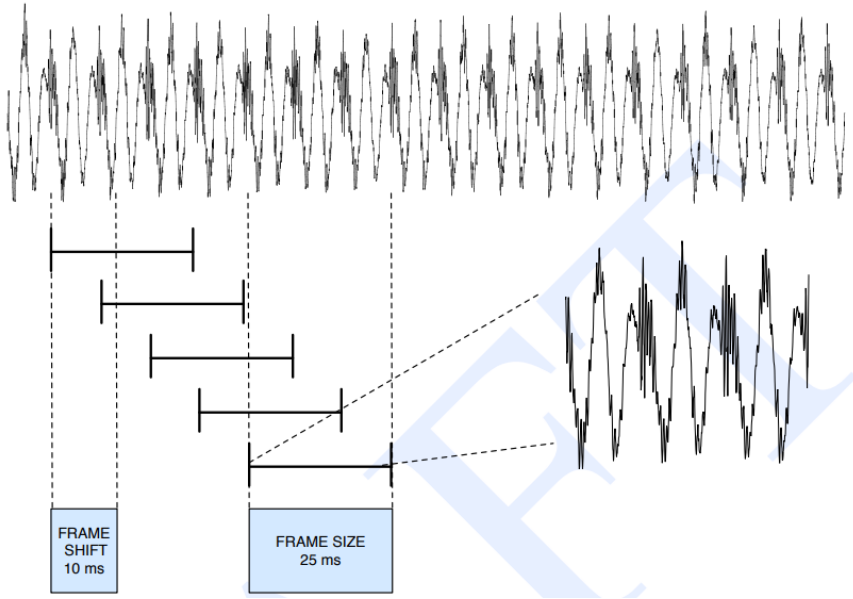




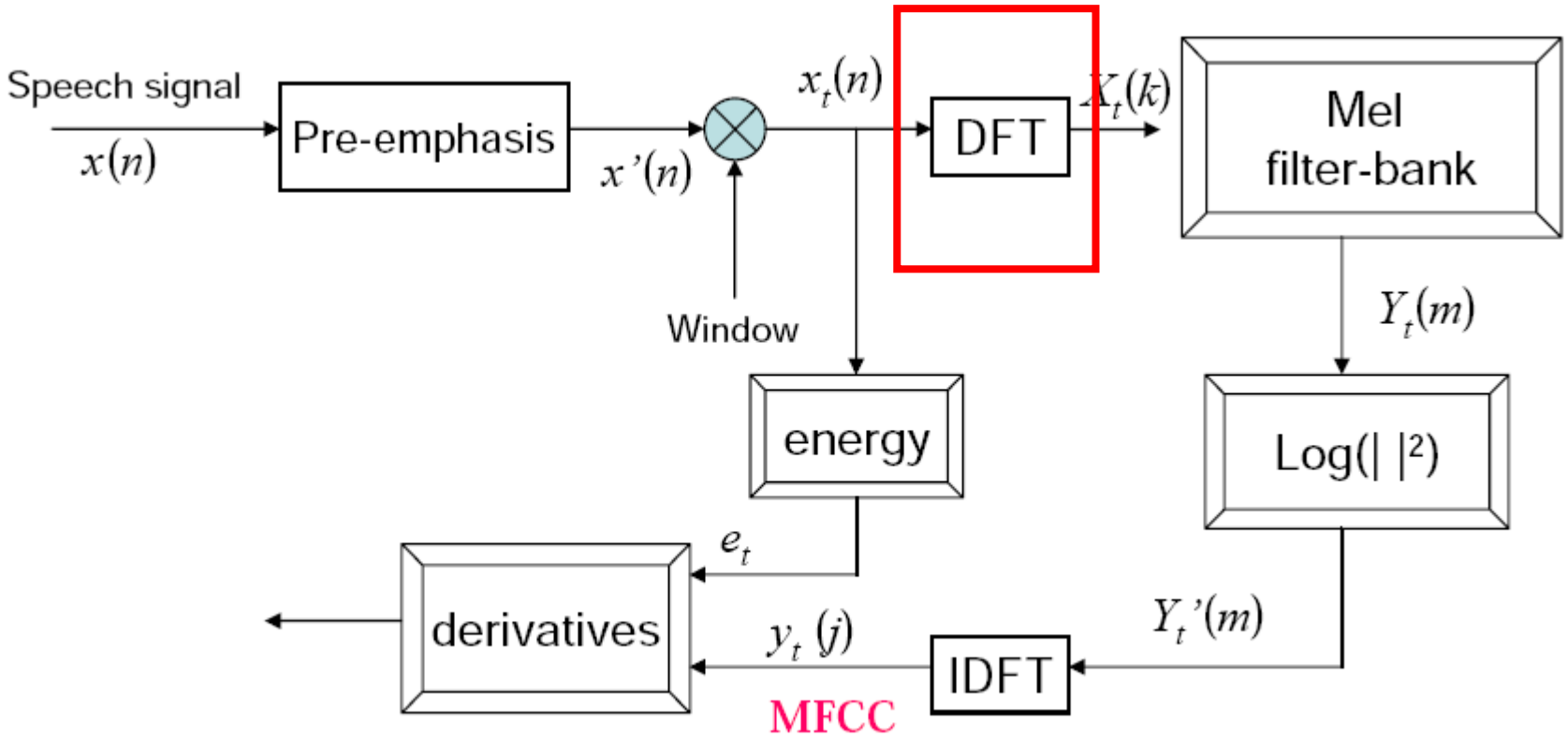
Окно

- $y[n] = w[n]x[n]$ – где: $w[n]$ – функция окна
- $w[n] = \begin{cases} 1 & 0 \le n \le L-1 \\ 0 & \text{иначе} \end{cases}$ - прямоугольное окно
- $w[n] = \begin{cases} 0.54 - 0.46\cos(\frac{2\pi n}{L}) & 0 \le n \le L-1 \\ 0 & \text{иначе} \end{cases}$ - hamming



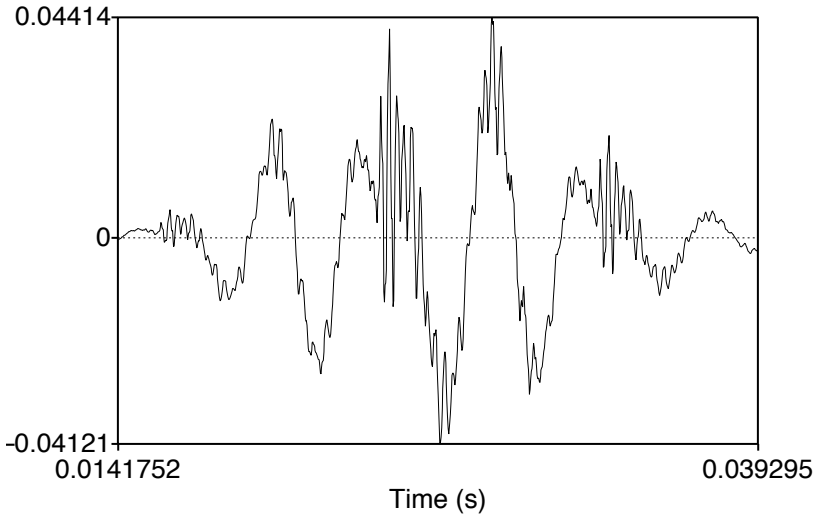


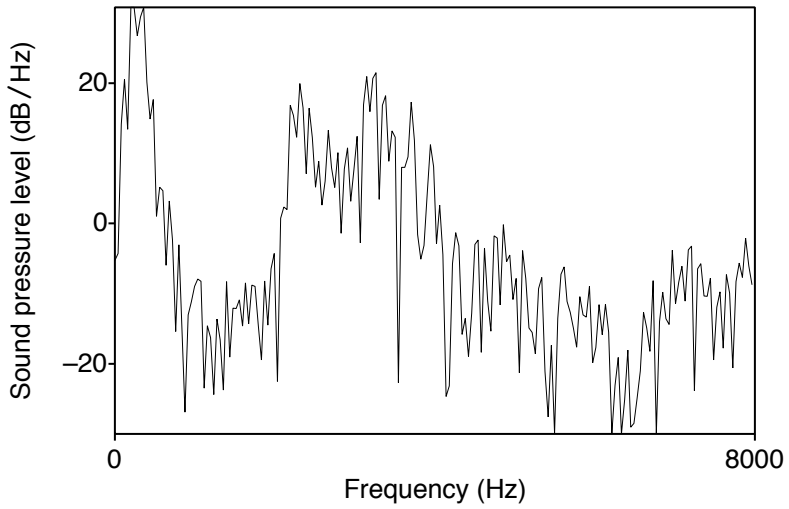
MFCC


Speech signal
$x(n)$
Pre-emphasis
$x'(n)$
Window
$x_t(n)$
DFT
$X_t(k)$
Mel
filter-bank
$Y_t(m)$
Log(| |²)
$Y_t'(m)$
IDFT
$y_t(j)$
energy
e_t
derivatives
MFCC

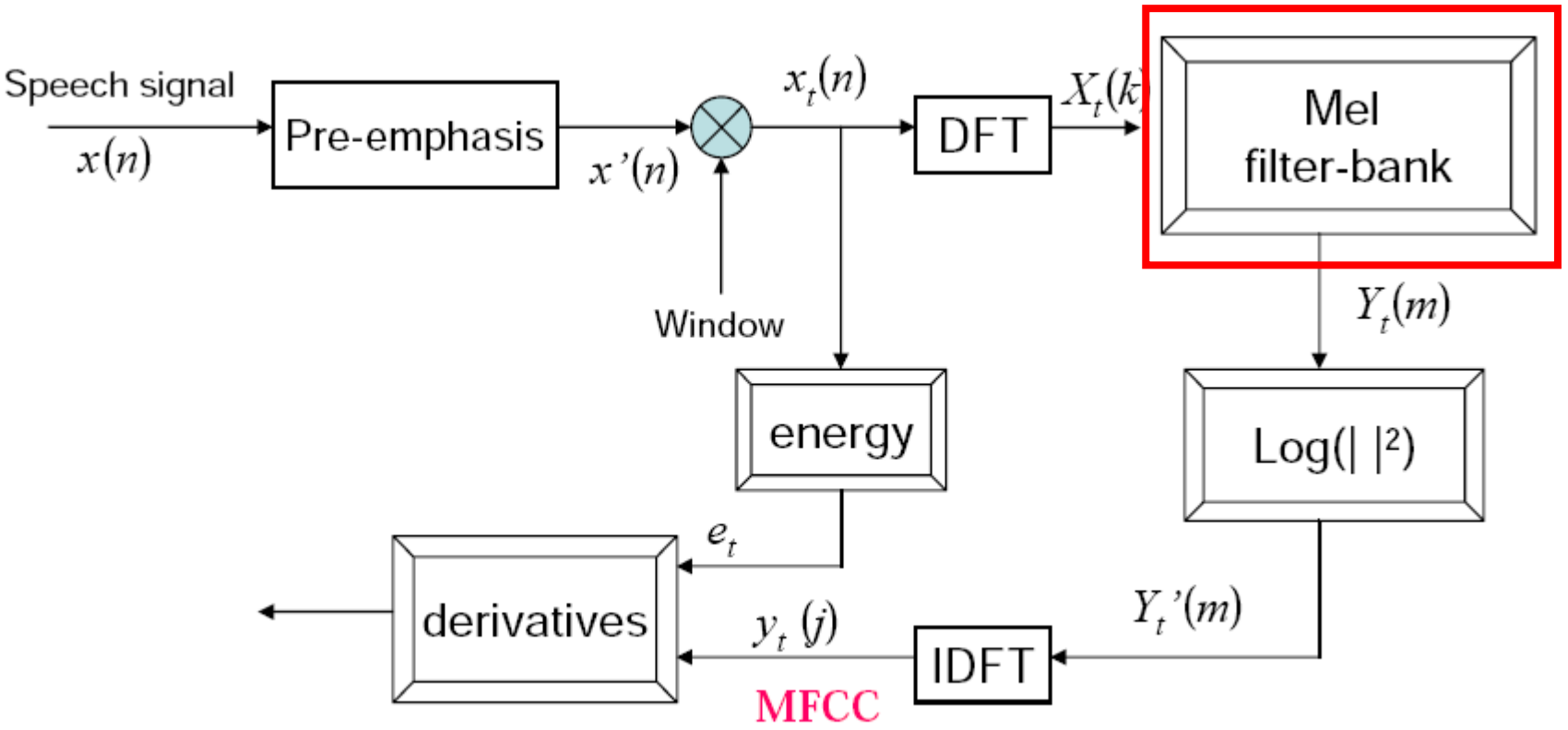
DFT

- Считаем спектр



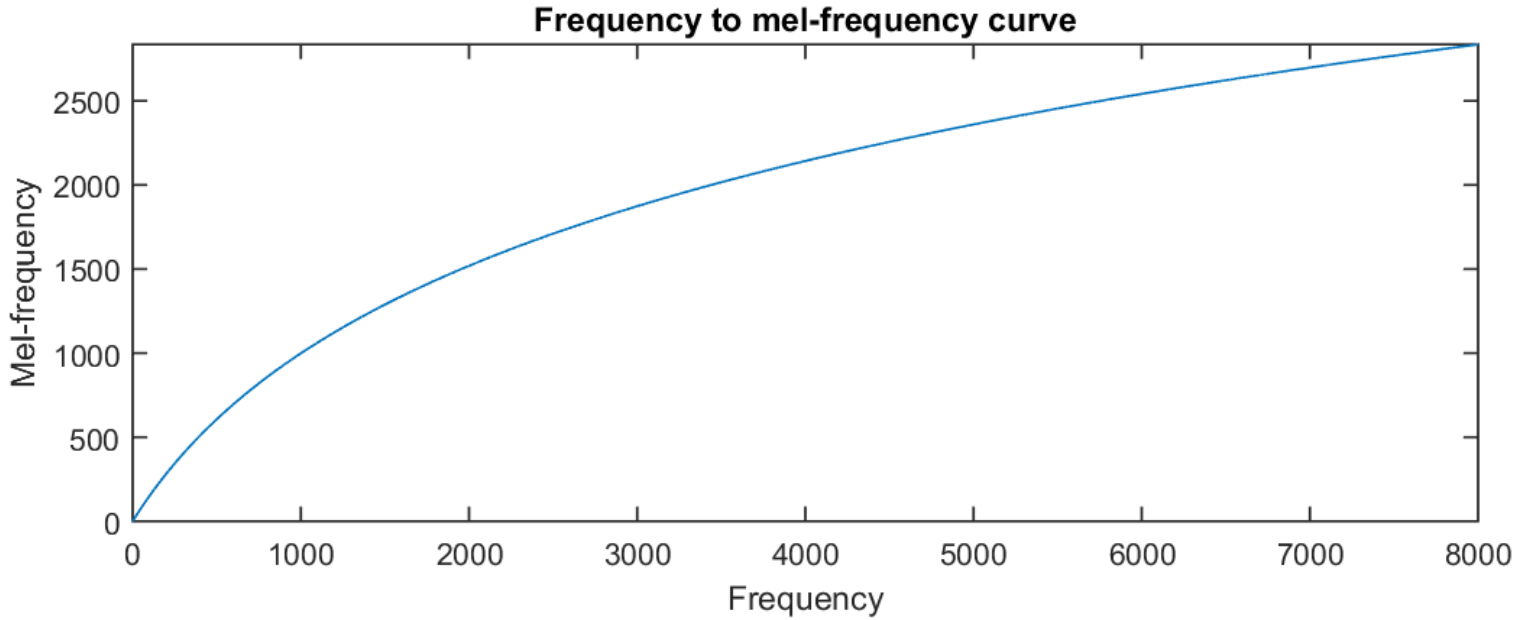


MFCC


Speech signal
$x(n)$
Pre-emphasis
$x'(n)$
Window
$x_t(n)$
DFT
$X_t(k)$
Mel
filter-bank
$Y_t(m)$
Log(| |²)
$Y_t'(m)$
IDFT
$y_t(j)$
energy
e_t
derivatives
MFCC

Mel filterbank

- Человек восприятие звука не одинаково на разных частотах
- Оно менее чувствительно на частотах выше 1000Гц

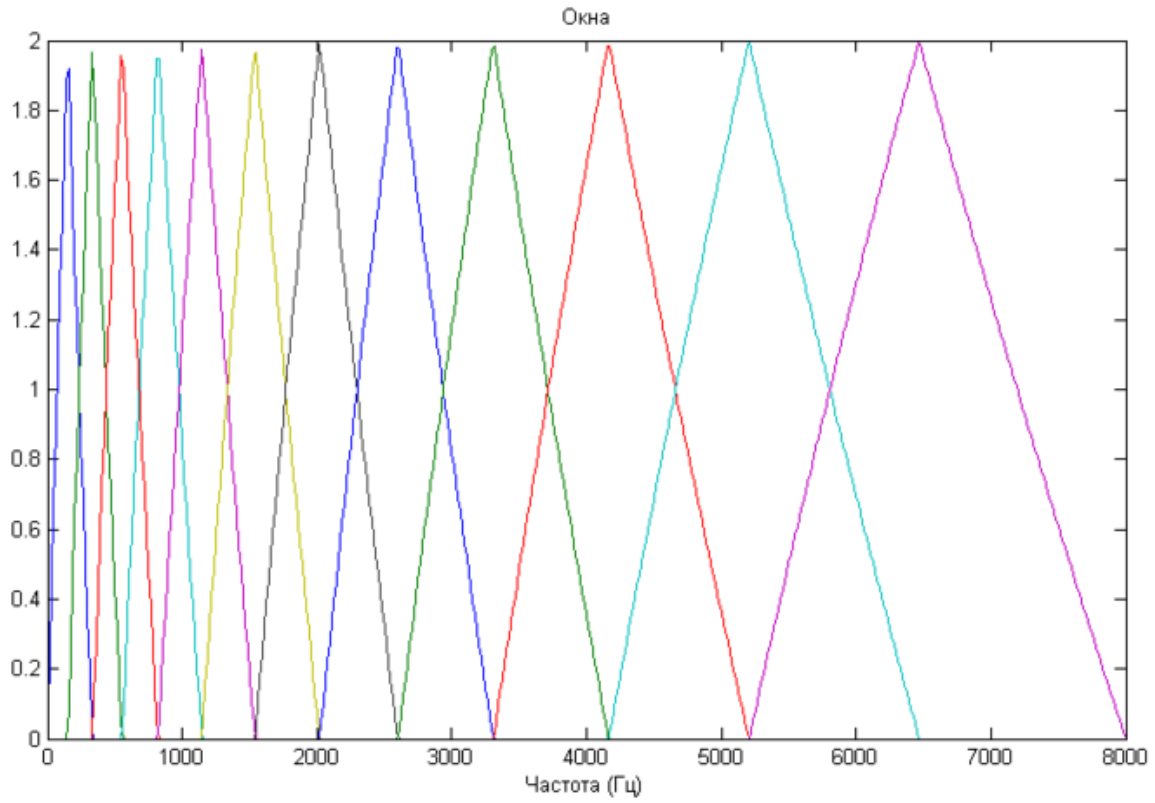


$$mel(f) = 1127 \ln(1 + \frac{f}{700})$$

Mel filterbank

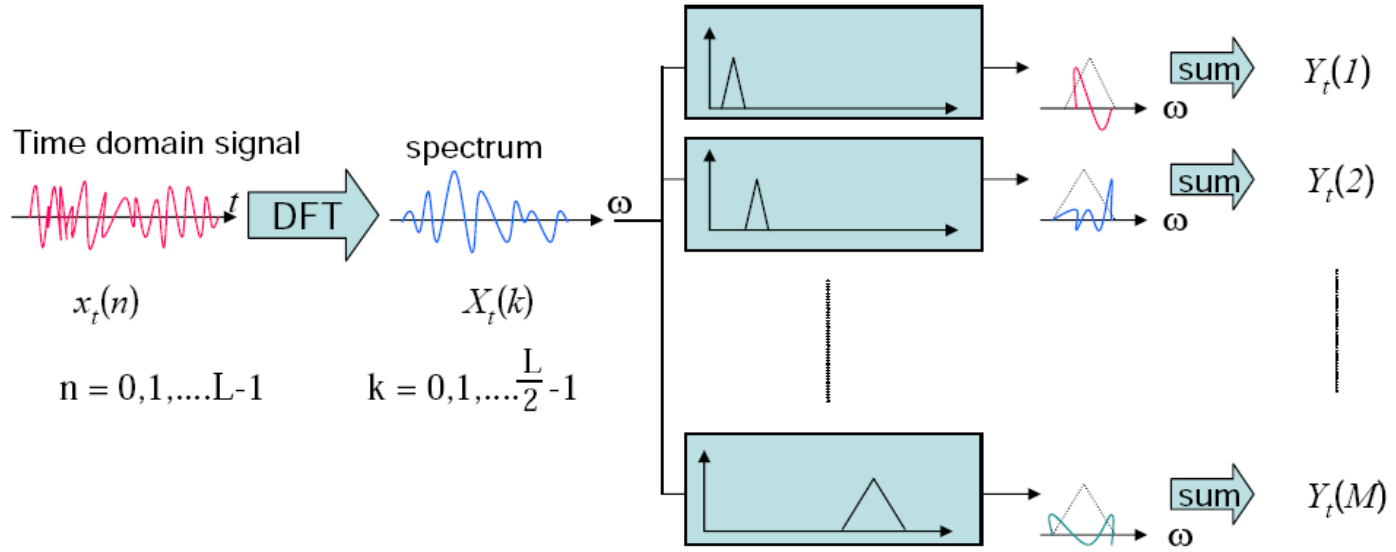
- Мы адаптируем алгоритм к человеческому восприятию

$$H_m'[k] = \begin{cases} 0 & k < f[m-1] \\ \dfrac{(k - f[m-1])}{(f[m] - f[m-1])} & f[m-1] \le k \le f[m] \\ \dfrac{(f[m+1] - k)}{(f[m+1] - f[m])} & f[m] \le k \le f[m+1] \\ 0 & k > f[m+1] \end{cases}$$

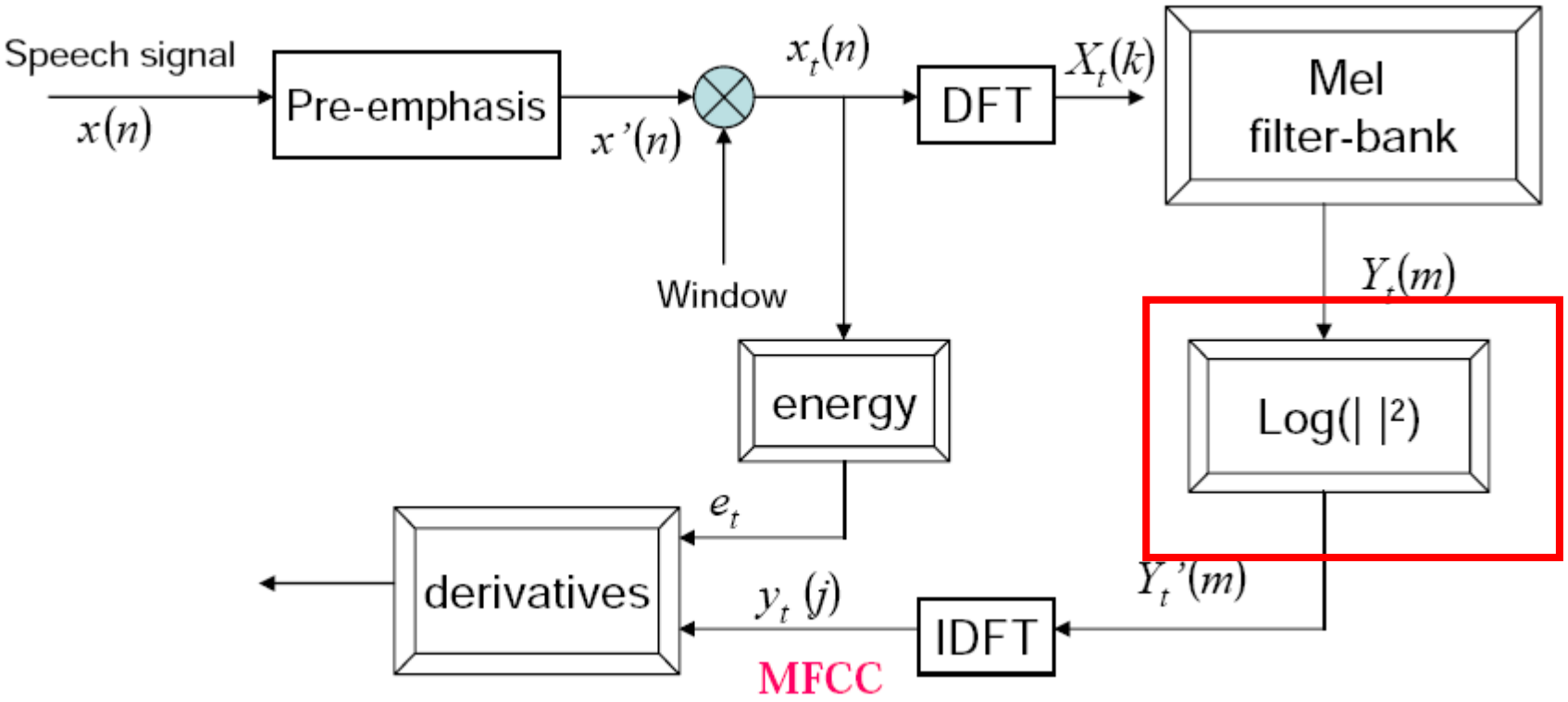


Mel filterbank обработка

- Применяем банк мел фильтров к спектру
- Каждый выход фильтра есть сумма отфильтрованных спектральных компонент

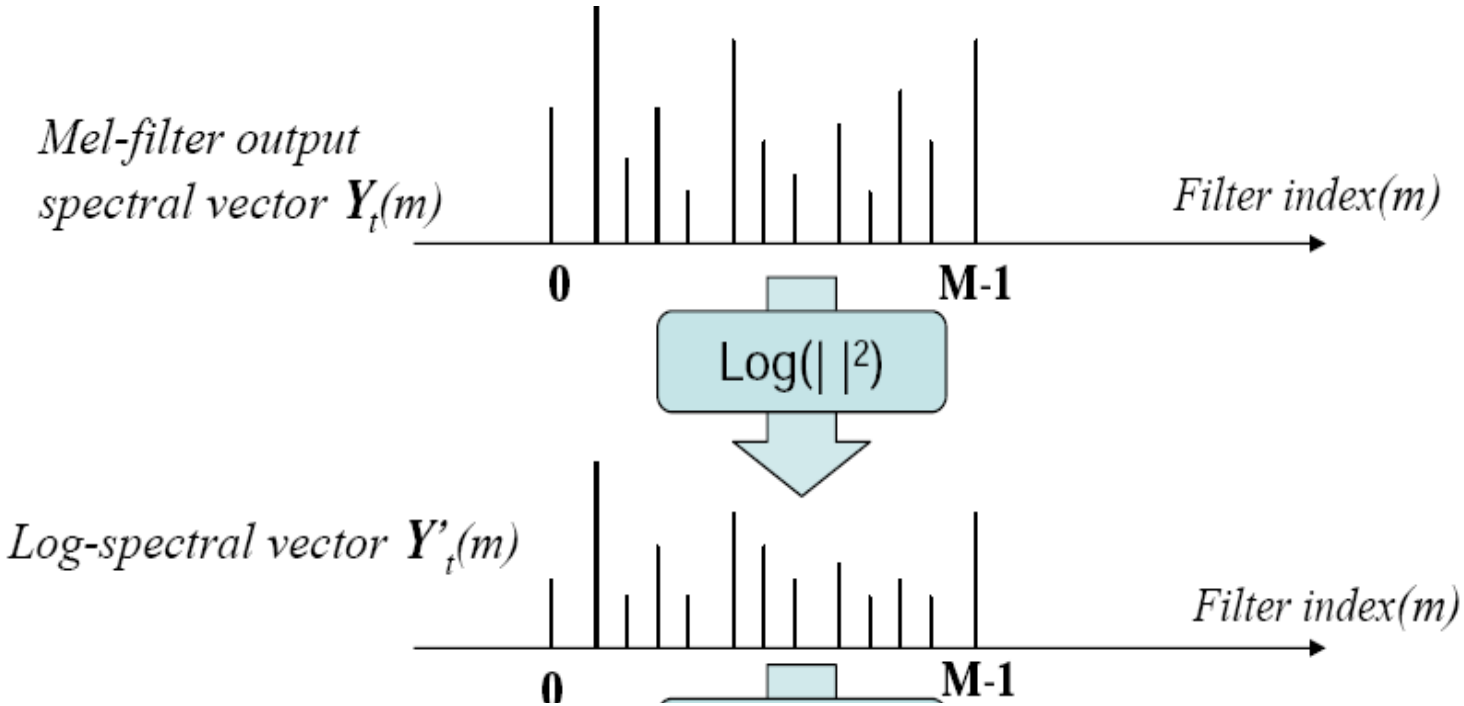


MFCC

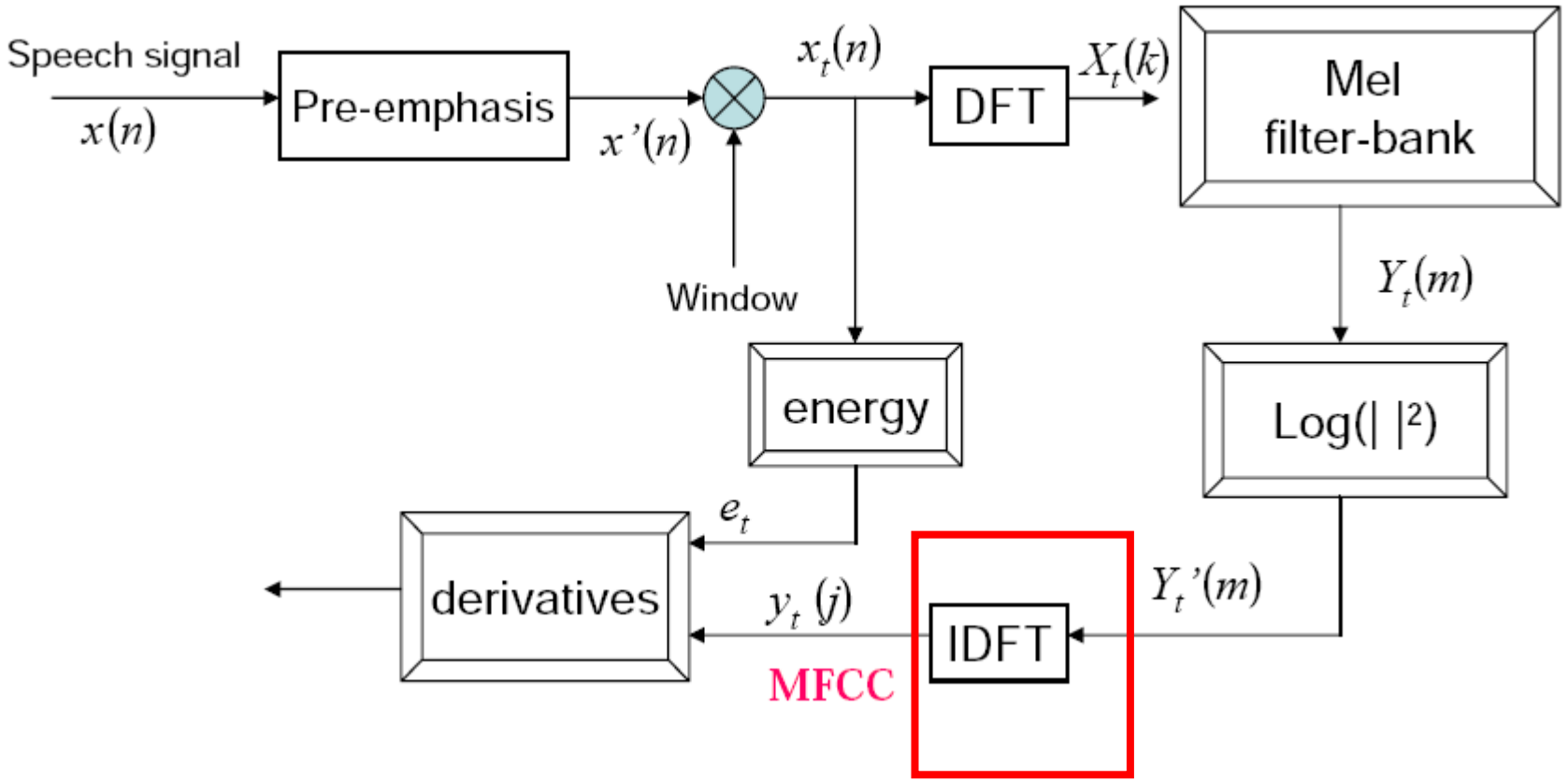

Speech signal
$x(n)$
Pre-emphasis
$x'(n)$
$x_t(n)$
DFT
$X_t(k)$
Mel
filter-bank
$Y_t(m)$
Log(| |²)
$Y_t'(m)$
IDFT
$y_t(j)$
MFCC
Window
energy
e_t
derivatives

Log energy

- Считаем логарифм амплитуды мел фильтра



MFCC


Speech signal
$x(n)$
Pre-emphasis
$x'(n)$
Window
$x_t(n)$
DFT
$X_t(k)$
Mel
filter-bank
$Y_t(m)$
Log(| |²)
$Y_t'(m)$
IDFT
$y_t(j)$
MFCC
energy
e_t
derivatives

Кепструм

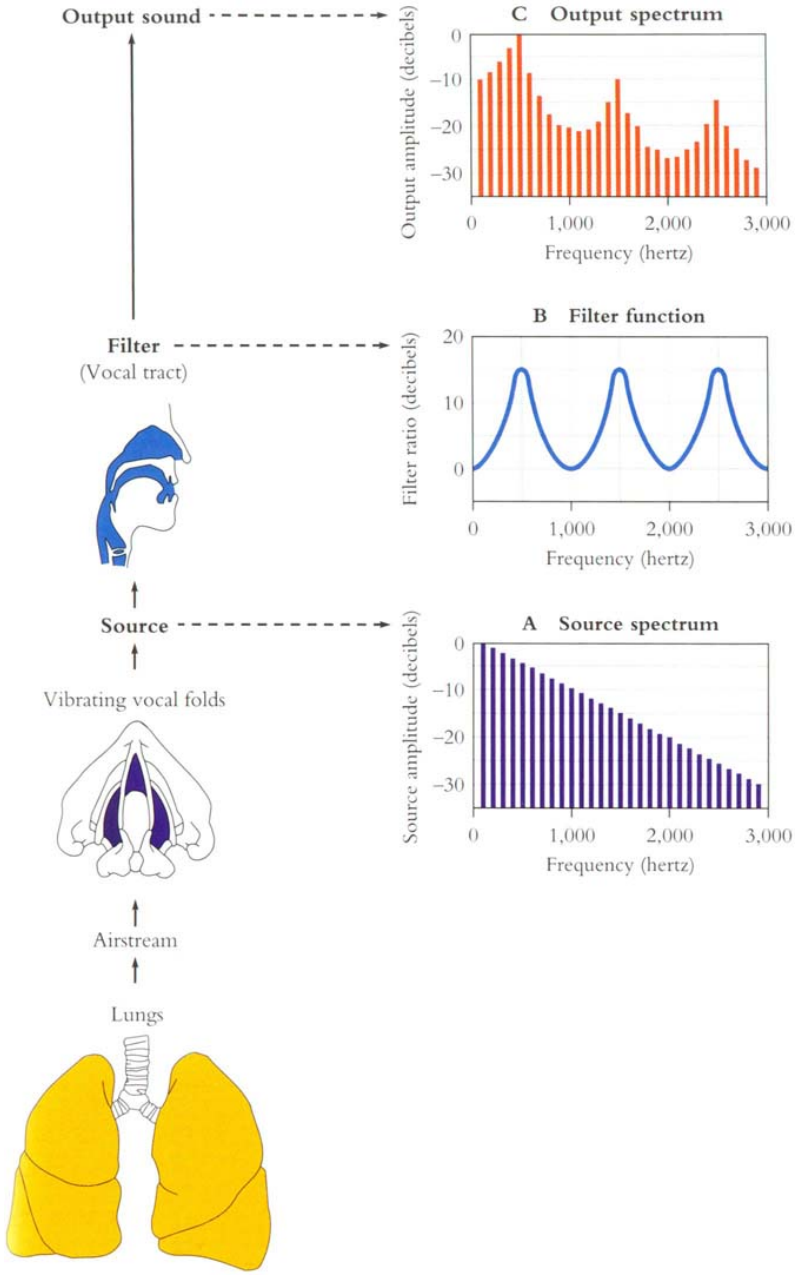
- Как разделить источник и фильтр
- Речевая волна создается
  - Голосовыми связками
  - Речевым трактом, который является, в свою очередь, фильтром
- Речевой тракт создает гармоники
- Ротовая полость является усилителем
- Усиление гармоник зависит от формы ротовой полости

Кепструм

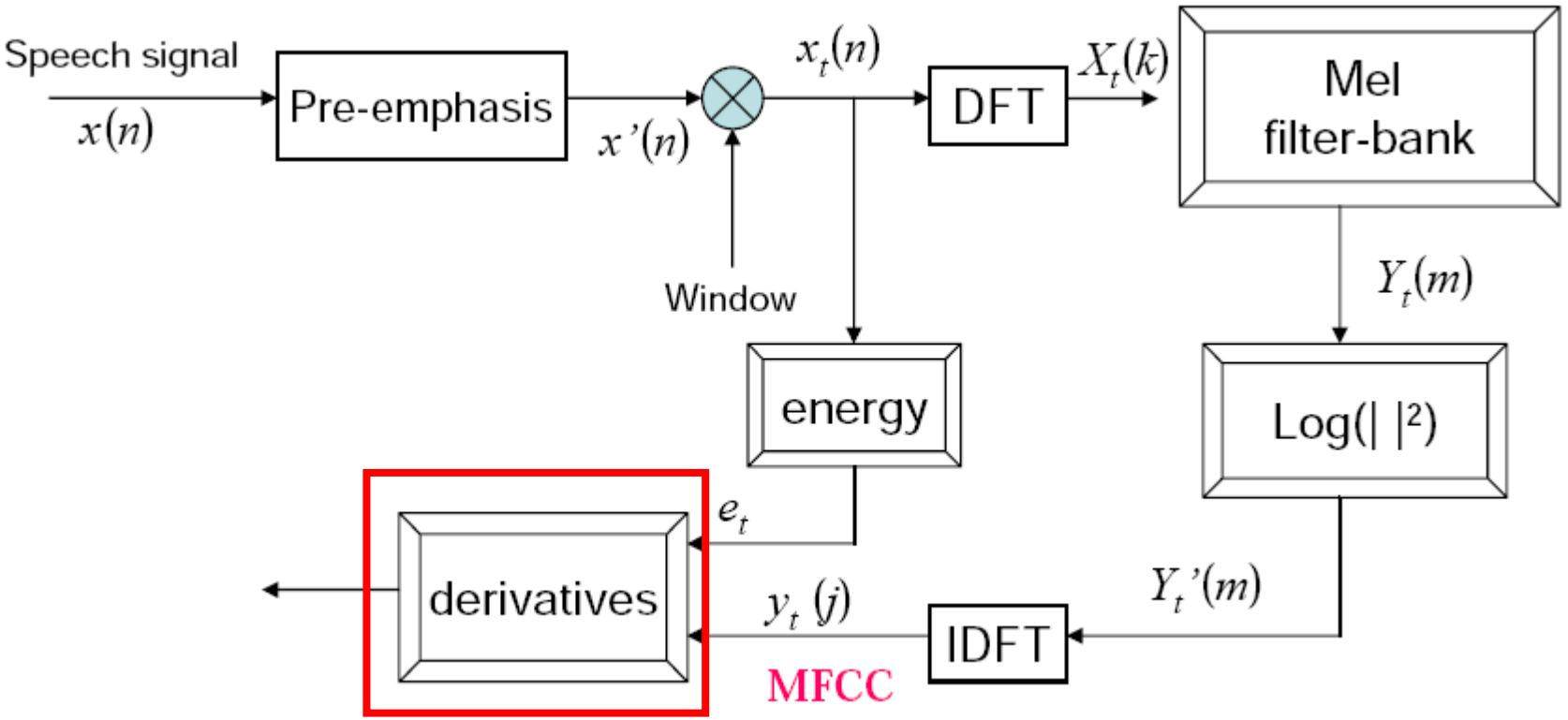
- Нужны характеристики фильтра
  - Делаем Фурье преобразование
  - Дискретное косинусное преобразование

$$c[n] = \sum_{m=0}^{M-1} S[m] \cos\left(\pi n (m + 1/2) / M\right) \qquad 0 \le n < M$$

получаются не коррелированные признаки



MFCC


Speech signal
$x(n)$
Pre-emphasis
$x'(n)$
Window
$x_t(n)$
DFT
$X_t(k)$
Mel
filter-bank
$Y_t(m)$
Log(| |²)
$Y_t'(m)$
IDFT
$y_t(j)$
energy
e_t
derivatives
MFCC

Дельты

- Можно считать дополнительно энергию:

$$Energy = \sum_{t=t_1}^{t_2} x^2[t]$$

- И изменение по фреймам кепстров c(t)

$$d(t) = \frac{c(t+1) - c(t-1)}{2}$$

Типичные параметры MFCC

- Window size: 25ms
- Window shift: 10ms
- Pre-emphasis coefficient: 0.97
- MFCC:
- 12 MFCC (mel frequency cepstral coefficients)
- 1 фича энергии
- 12 дельта MFCC фич
- 12 изменение дельты MFCC фичей
- 1 дельта энергии
- 1 изменение дельты энергии

Аугментация данных

- Добавление шума
  - Реальный фоновый шум
  - Фоновая болтовня
- Реверберации
  - Генерируется умный импульсный ответ для параметров помещения
- Шум добавляется как отношение к максимальной амплитуде сигнала signal-noise-ratio (SNR) – (-5, 0, 5, 10, and 15 dB).
- Для реверберации случайно выбираются параметры и создается импульсный ответ

Speech Enhancement

- Очистка от шума.

Table 10.1 Performance of speech enhancement techniques on CHiME-3

Test data enhancement	XE (%WER)
None	48.86
WPE	45.36
Autoencoder	**30.58**

WER word error rate

Аугментация

Table 10.2 Results on CHiME-3 with different data augmentation variants

DNN training data		Test data enhancement	XE (%WER)
Noise type	Reverb		
None	None	None	48.86
Babble	Artificial	None	26.41
Stationary	Artificial	None	25.8
Stationary	None	None	24.26
Stationary	None	WPE	**22.72**

Результат аугментации шумом и реверберацией и обучение DNN