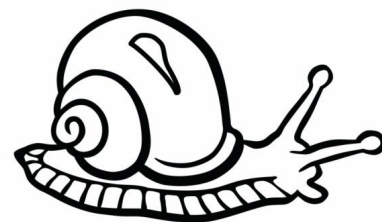


Автоматическое распознавание речи. Акустические признаки

П. А. Холявин

p.kholyavin@spbu.ru





Задача распознавания речи

Задача АРР – сопоставить акустическому сигналу последовательность слов.
Более формально: каково наиболее вероятное предложение из всех возможных в языке L при условии акустического сигнала O ?

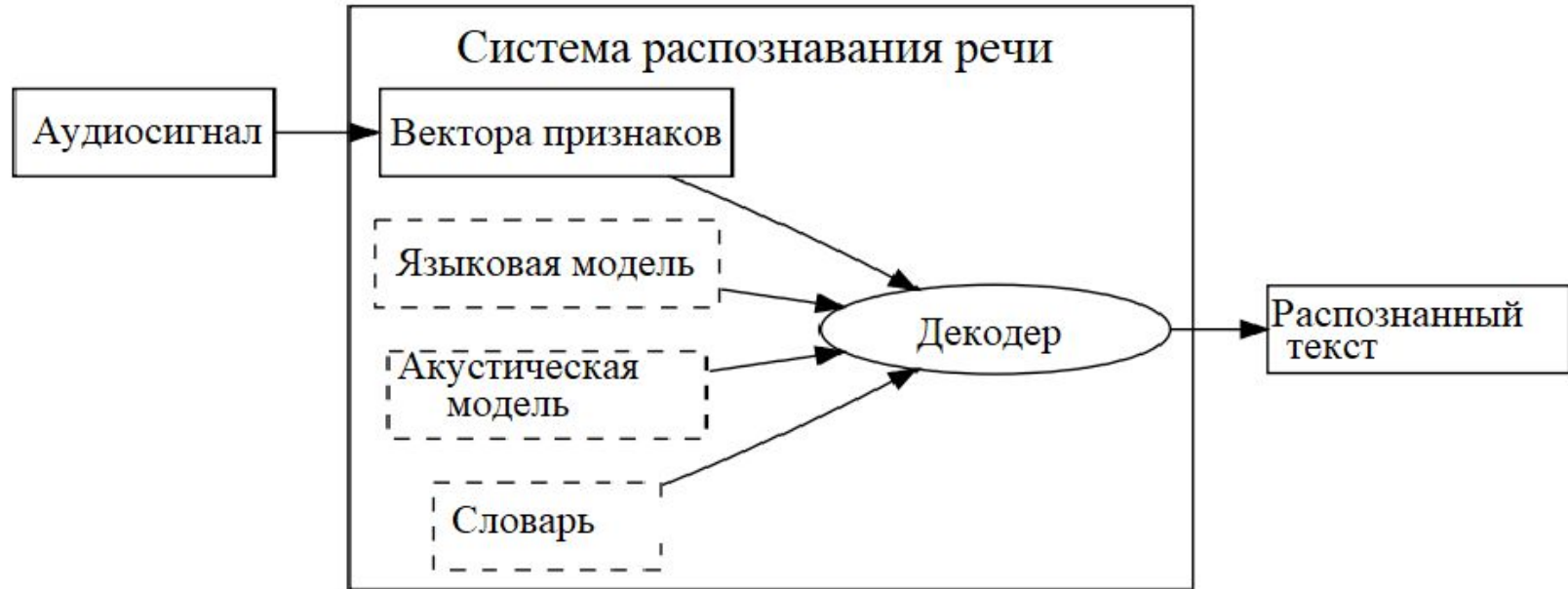
Если $O = o_1, o_2, \dots, o_n$ – звуковая последовательность,
 $W = w_1, w_2, \dots, w_n$ – последовательность слов, то

$$\hat{W} = \underset{W \in L}{\operatorname{argmax}} P(W|O)$$

$$\hat{W} = \underset{W \in L}{\operatorname{argmax}} \frac{P(O|W) P(W)}{P(O)} = \underset{W \in L}{\operatorname{argmax}} P(O|W) P(W)$$



Части системы АРР

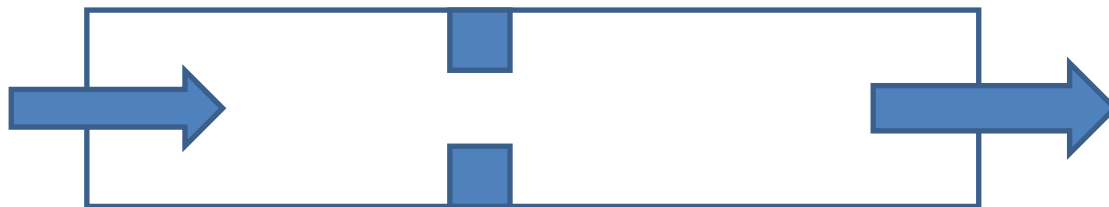




Акустика речеобразования

Движения речевых органов создают условия, необходимые для образования колебаний:

1. Образование потока воздуха.
2. Образование колебаний путём наложения возмущений на этот поток воздуха:
 - а) голосовыми связками
 - б) преградами, образованными органами речи.

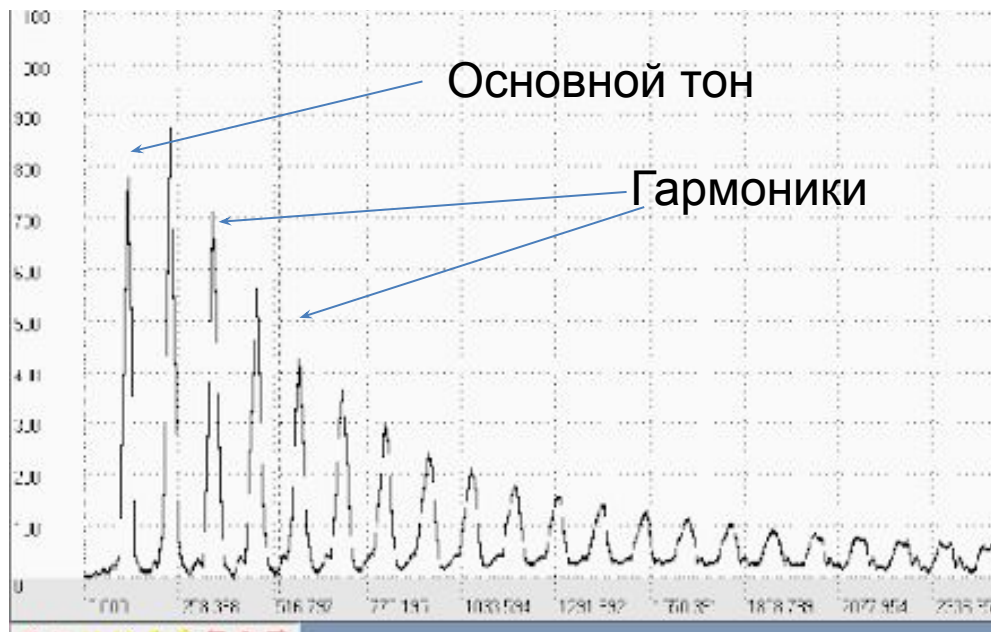


Частотная фильтрация: $P(f) = S(f) * T(f)$



Голосовой источник звука

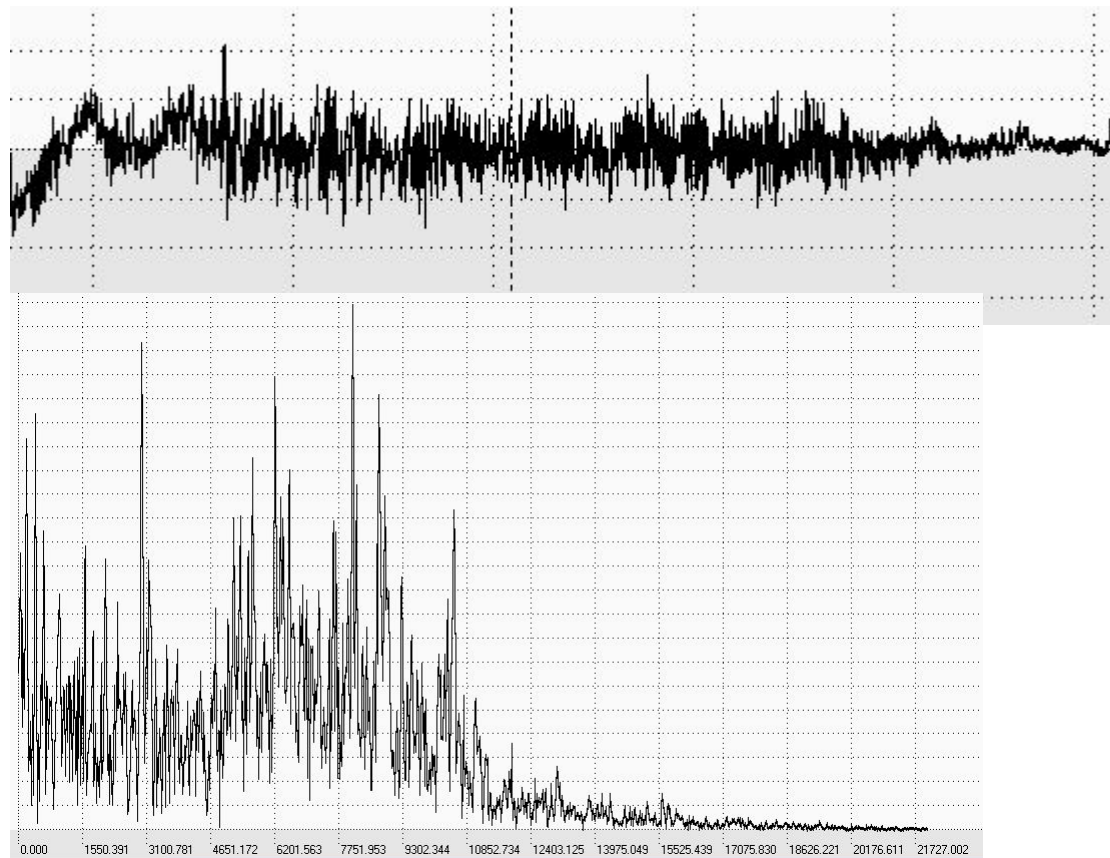
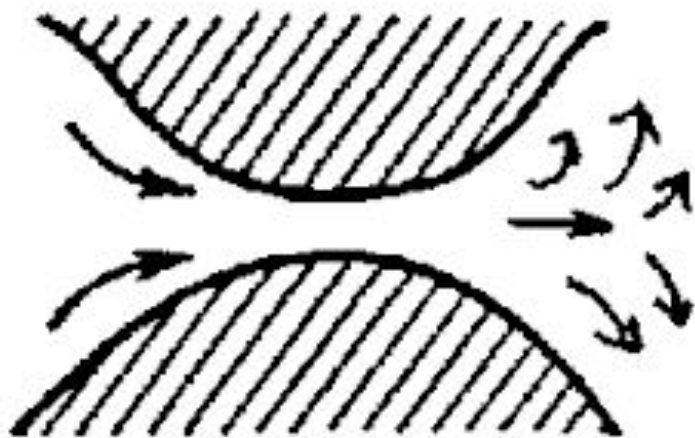
Спектр глоттальной волны:





Турбулентный источник звука

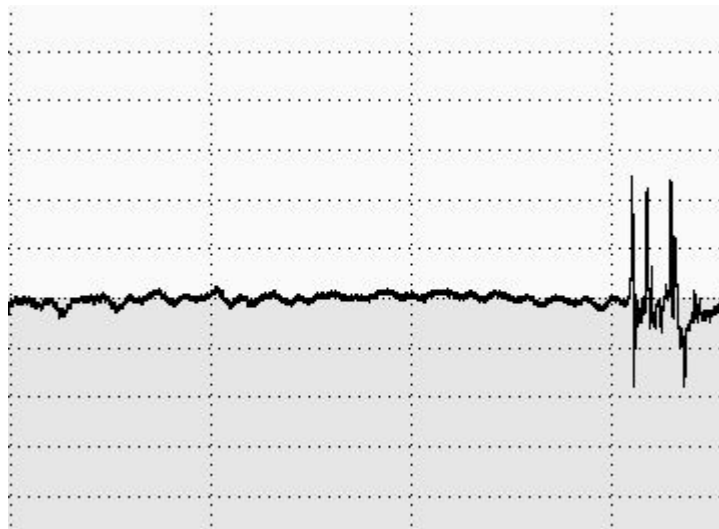
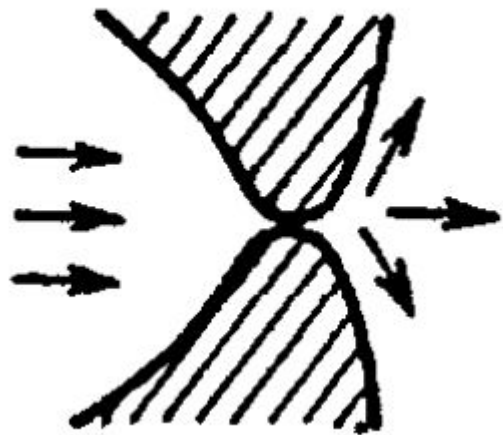
Шум образуется в результате прохождения воздуха через сужение





Импульсный источник звука

Шум образуется в результате резкого движения речевых органов





Источники звука в речи

Г: гласные [a], сонанты [n]

Т: глухие щелевые согласные [s]

И (*потом Т): глухие смычные согласные [t]

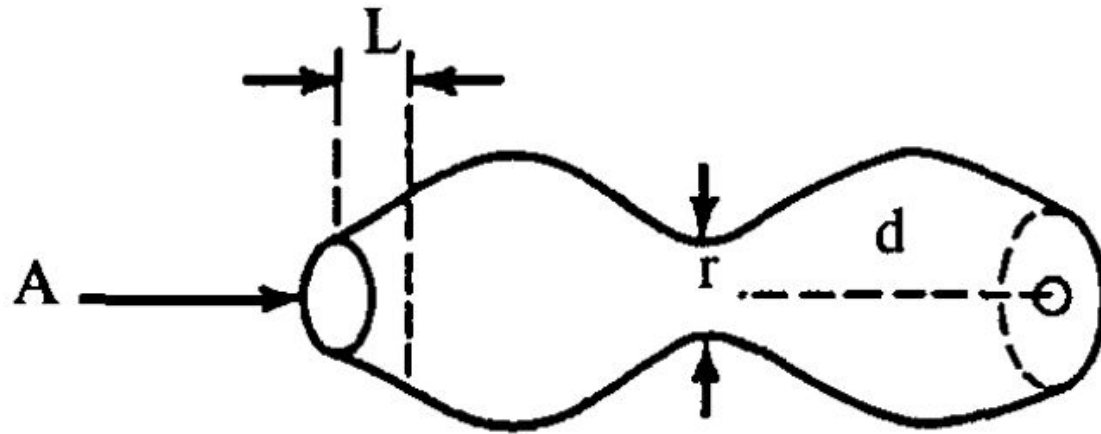
Г + Т: звонкие щелевые согласные [z]

Г + И (*потом Т): звонкие смычные согласные [d]



Передаточная функция речевого тракта

Схематичное изображение
речевого тракта:



A – площадь сужения на
выходе (губы)
L – его длина
r – площадь язычного
сужения
d – расстояние от
язычного сужения до
голосовой щели



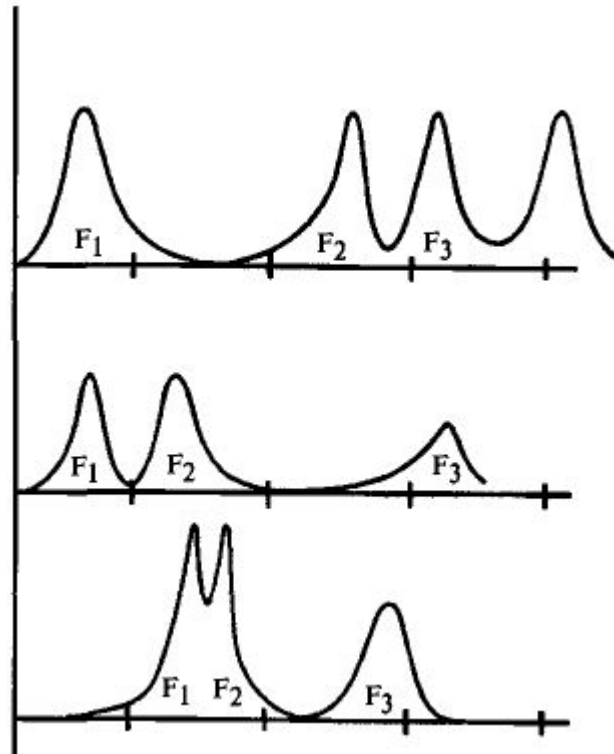
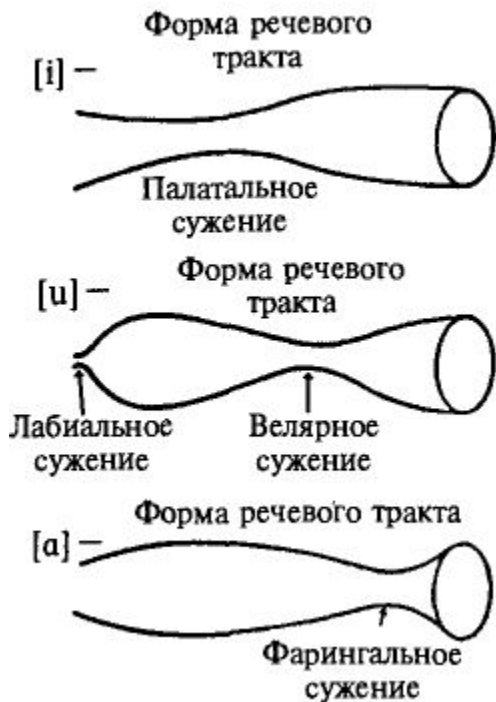
Передаточная функция речевого тракта

Воздушный столб, заключённый в речевом тракте – сложная колебательная система.

Её передаточная функция – сложная кривая. Её максимумы – **форманты**: F_1 , F_2 , $F_3 \dots$

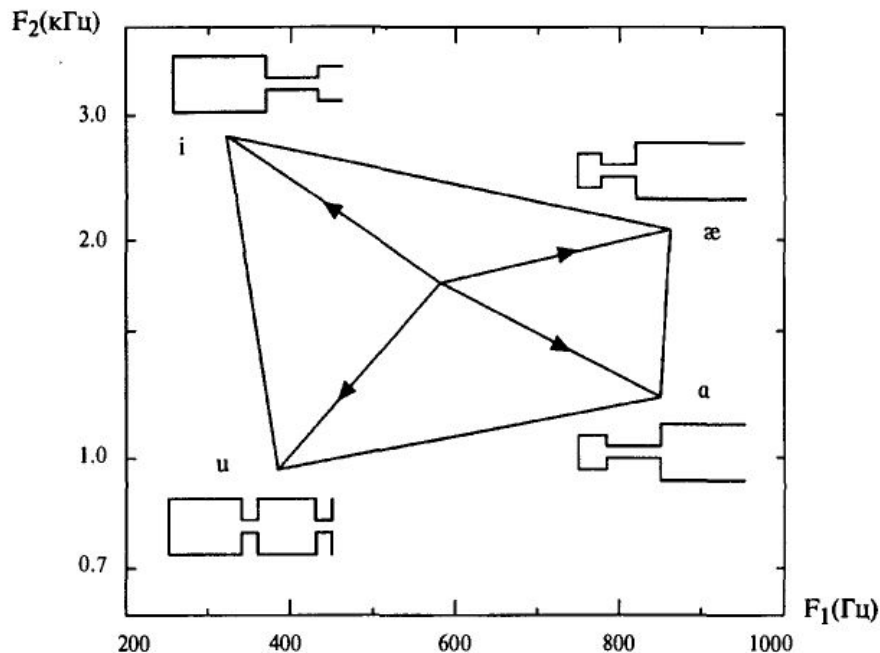


Передаточная функция речевого тракта

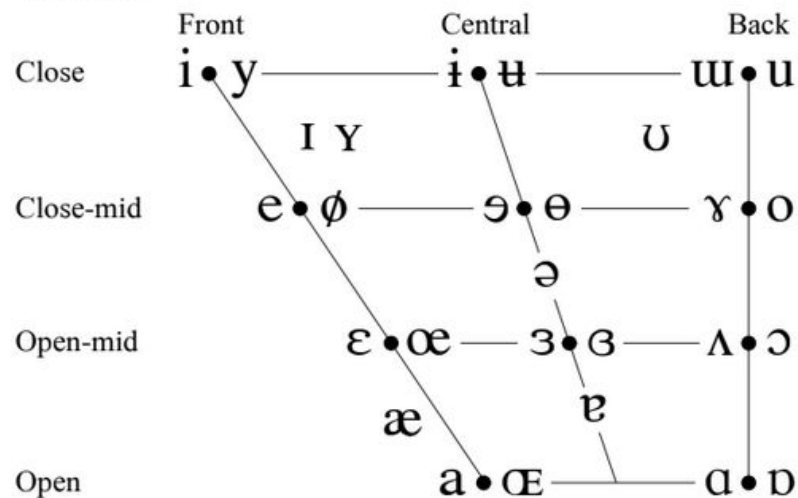




Передаточная функция речевого тракта



VOWELS

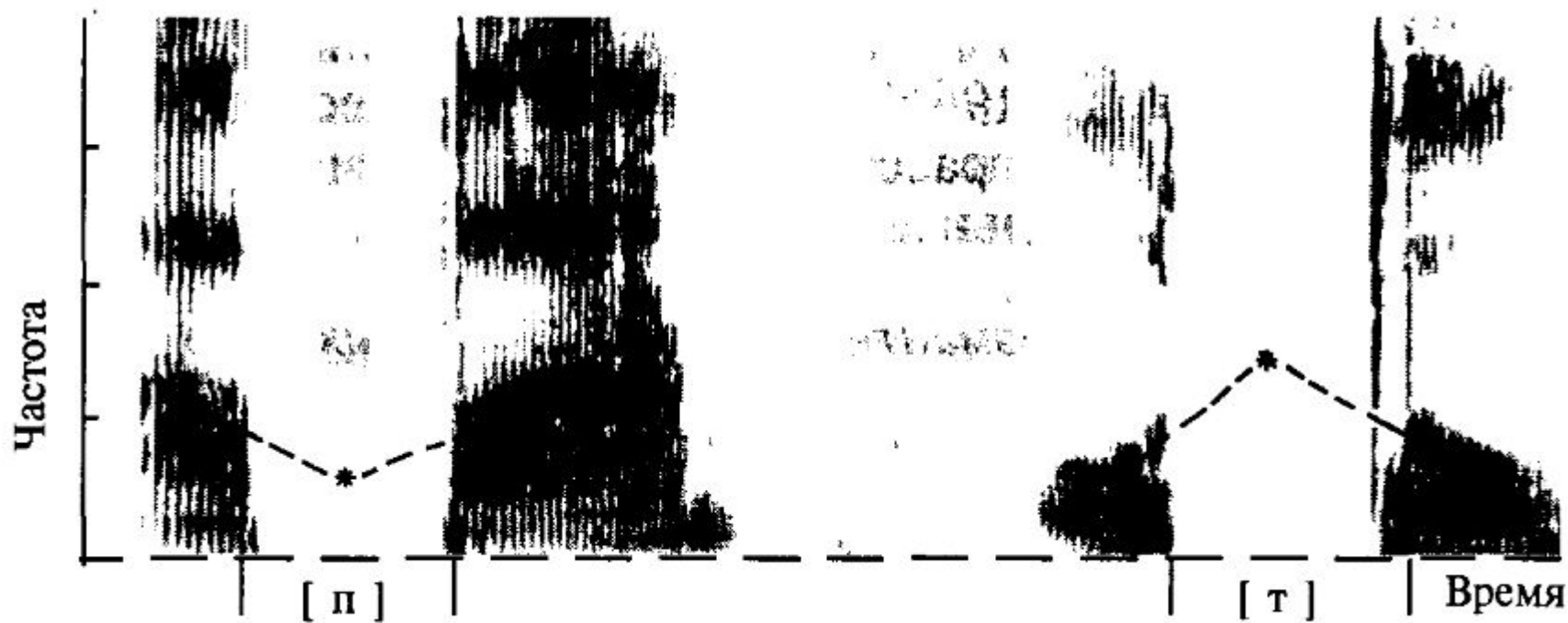


Where symbols appear in pairs, the one to the right represents a rounded vowel.



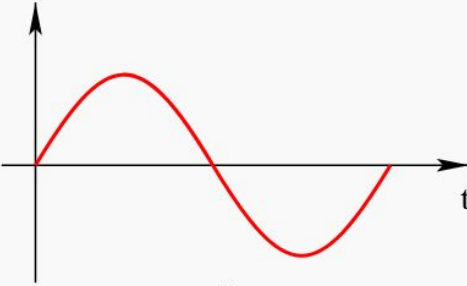
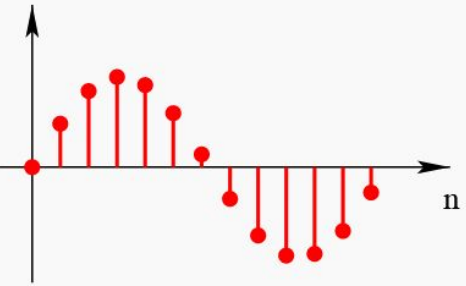
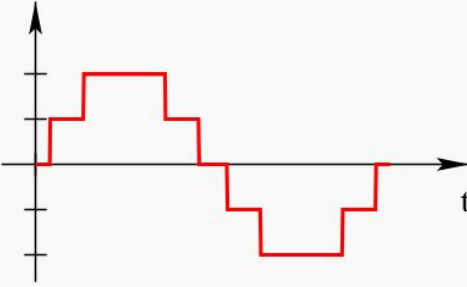
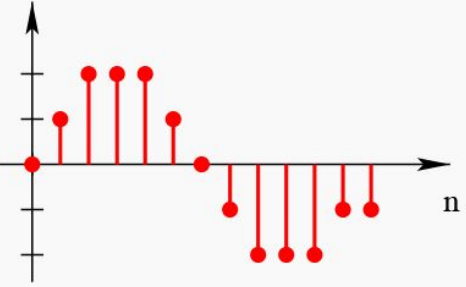
Локусы согласных

Согласные на слух различаются также с опорой на форманты близлежащих участков. Экстраполированные значения формант называются **локусами**.





Аналого-цифровое преобразование

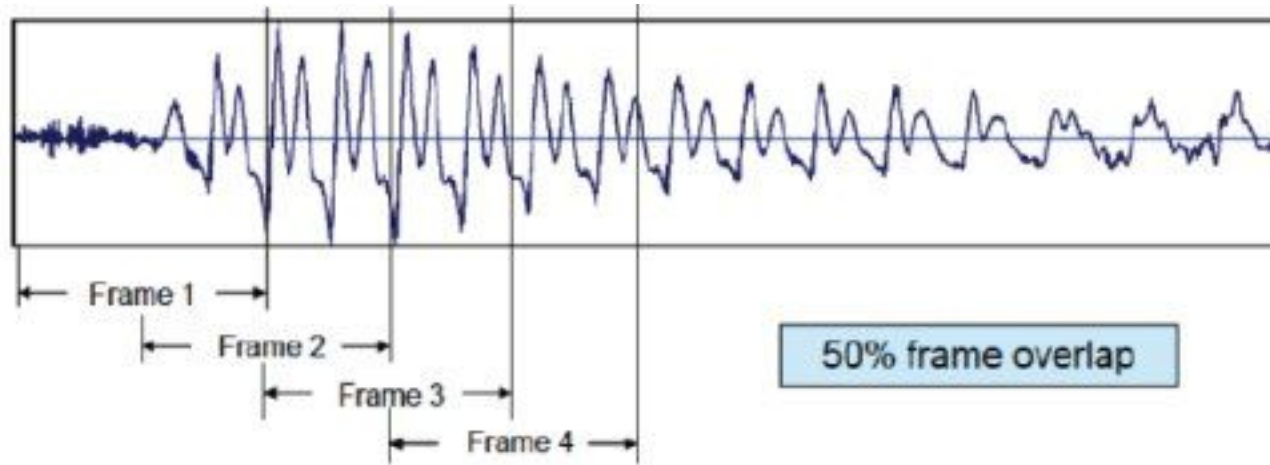
	Непрерывные	Дискретные
Неквантованные	 <p>Аналоговый сигнал</p>	
Квантованные	 <p>Цифровой сигнал</p>	 <p>Дискретный (квантованный)</p>

$$F_{\text{дискр}} = 1 / T$$



Обработка сигнала

1. Разбиение сигнала на окна (frames):



Сколько окон в сигнале длиной N отсчётов, если длина окна = N_F и шаг = N_H ?



Обработка сигнала

<https://brianmcfee.net/dstbook-site/content/ch09-stft/Framing.html>

$$K = 1 + \left\lfloor \frac{N - N_F}{N_H} \right\rfloor.$$

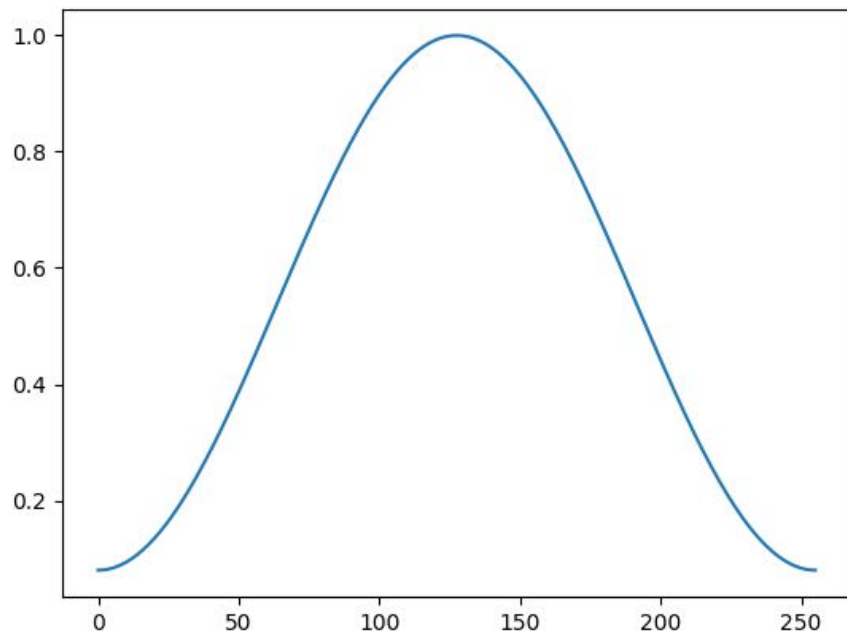


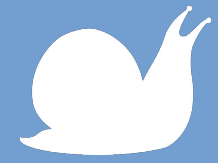
Обработка сигнала

2. Оконные функции:

а) прямоугольное окно

б) окно Хэмминга (Hamming)

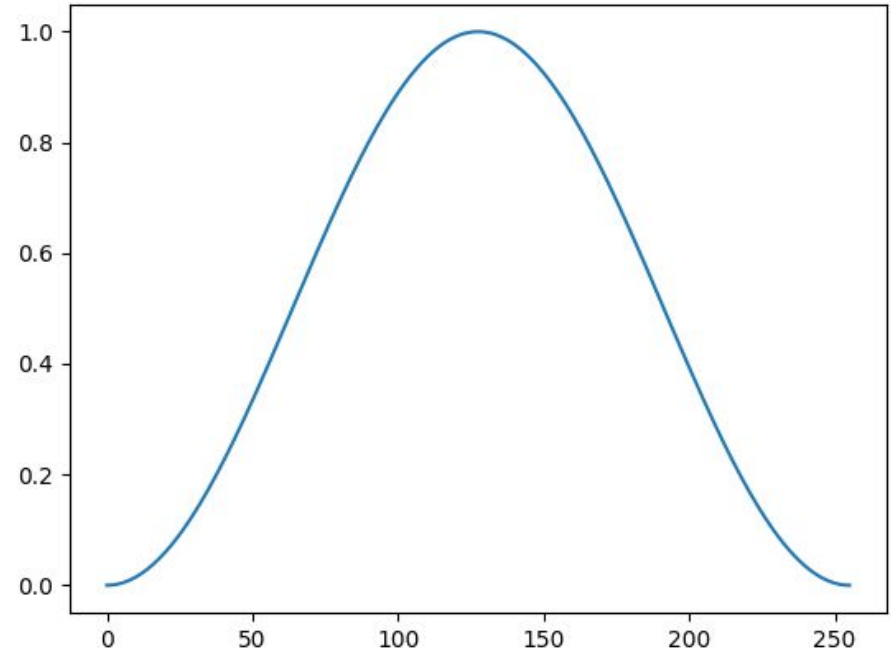




Обработка сигнала

2. Оконные функции:

в) окно Ханна/Ханнинга (Hann/Hanning)



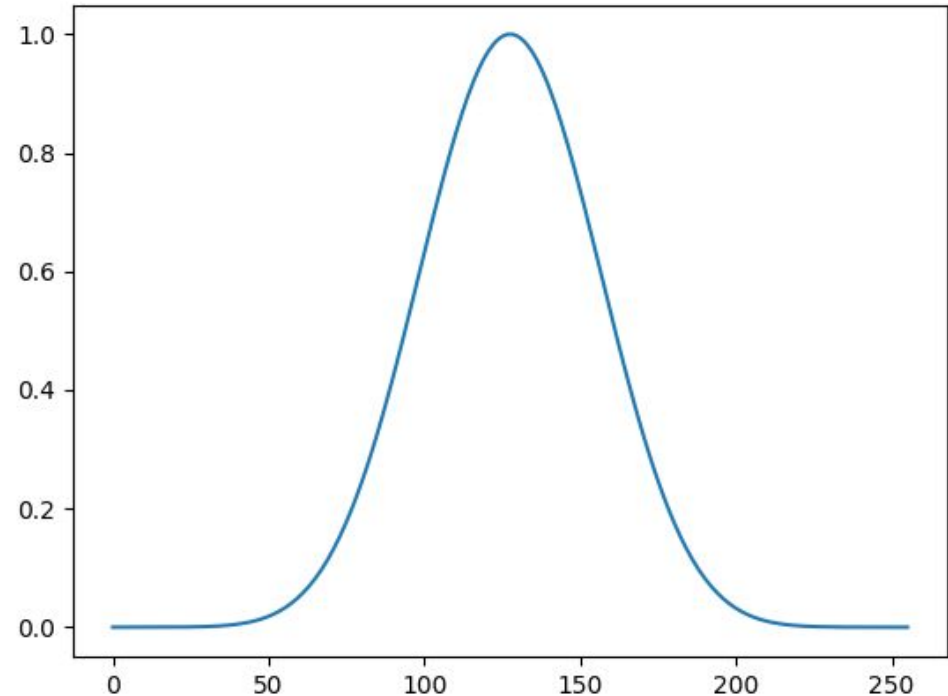


Обработка сигнала

2. Оконные функции:

г) окно Кайзера (Kaiser)

$$w(n) = I_0 \left(\beta \sqrt{1 - \frac{4n^2}{(M-1)^2}} \right) / I_0(\beta)$$





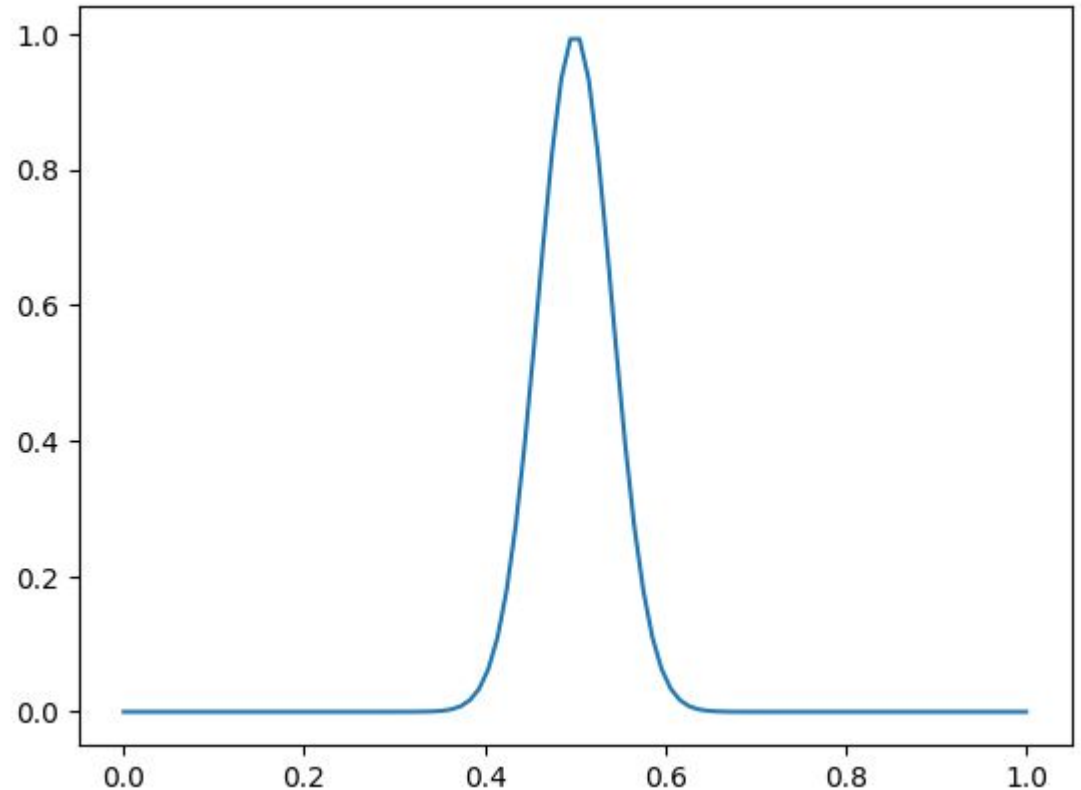
Обработка сигнала

2. Оконные функции:

д) Гауссово окно

$$w[n] = \exp\left(-\frac{1}{2}\left(\frac{n - N/2}{\sigma N/2}\right)^2\right), \quad 0 \leq n \leq N.$$

$$\sigma \leq 0.5$$





Обработка сигнала

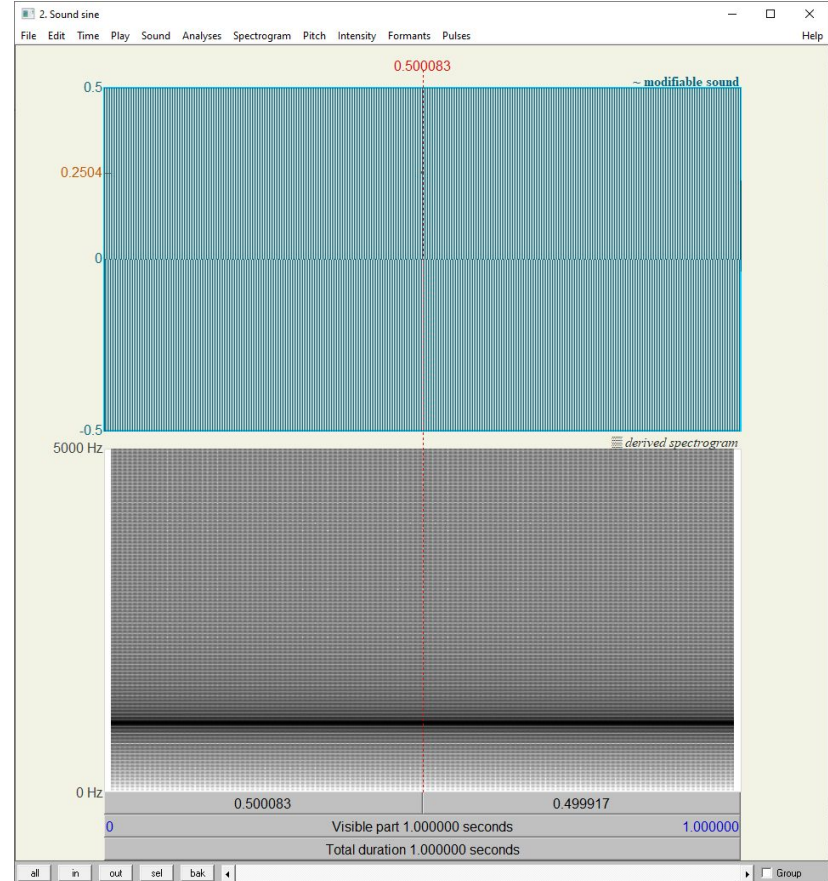
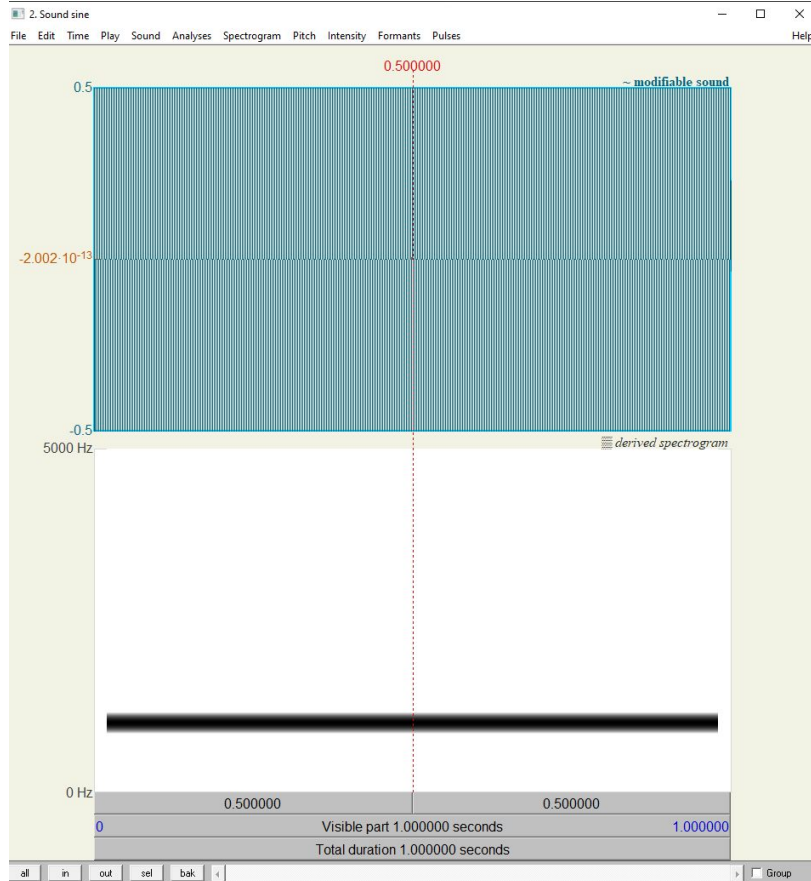
Sidelobes; anybody wants to win a cake?

The Gaussian window is the only shape that we can consider seriously as a candidate for the analysis window. To see this, create a 1000-Hz sine wave with [Create Sound from formula...](#) by typing $1/2 * \sin(2*\pi*1000*x)$ as the formula, then click **View & Edit**. If the window shape is Gaussian, the spectrogram will show a horizontal black line. If the window shape is anything else, the spectrogram will show many horizontal grey lines (*sidelobes*), which do not represent anything that is available in the signal. They are artifacts of the window shapes.

We include these other window shapes only for pedagogical purposes and because the Hanning and Hamming windows have traditionally been used in other programs before computers were as fast as they are now (a spectrogram is computed twice as fast with these other windows). Several other programs still use these inferior window shapes, and you are likely to run into people who claim that the Gaussian window has disadvantages. We promise such people a large cake if they can come up with sounds that look better with Hanning or Hamming windows than with a Gaussian window. An example of the reverse is easy to find; we have just seen one.



Обработка сигнала

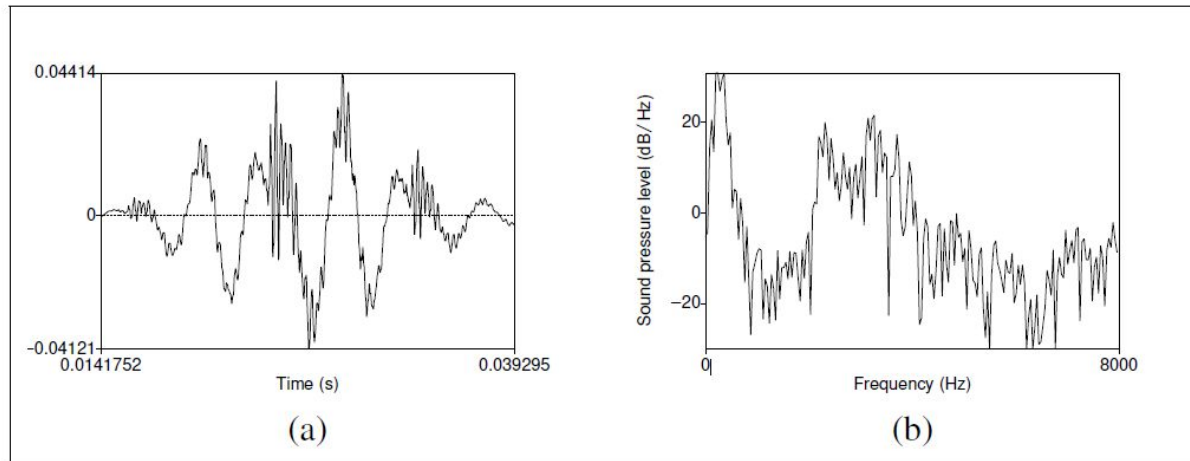




Обработка сигнала

3. Дискретное преобразование Фурье (DFT)

$$X_k = \sum_{n=0}^{N-1} x_n \cdot e^{-\frac{i2\pi}{N} kn}$$

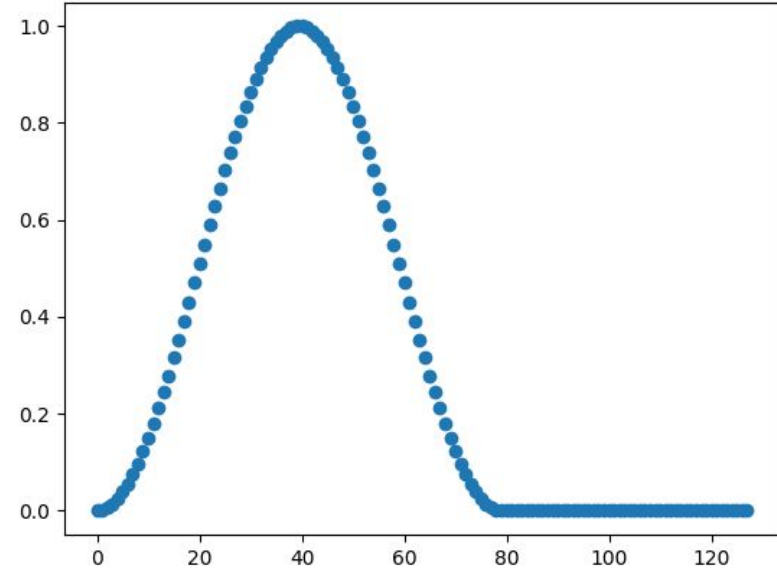
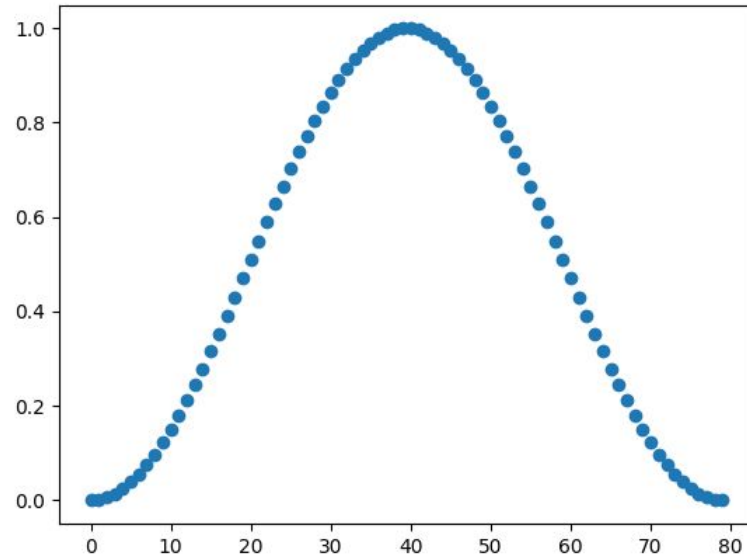




Обработка сигнала

3a. zero-padding

$F_s = 8000$ Hz, frame = 0.01 s



3b. Быстрое преобразование Фурье (FFT)



Акустические признаки

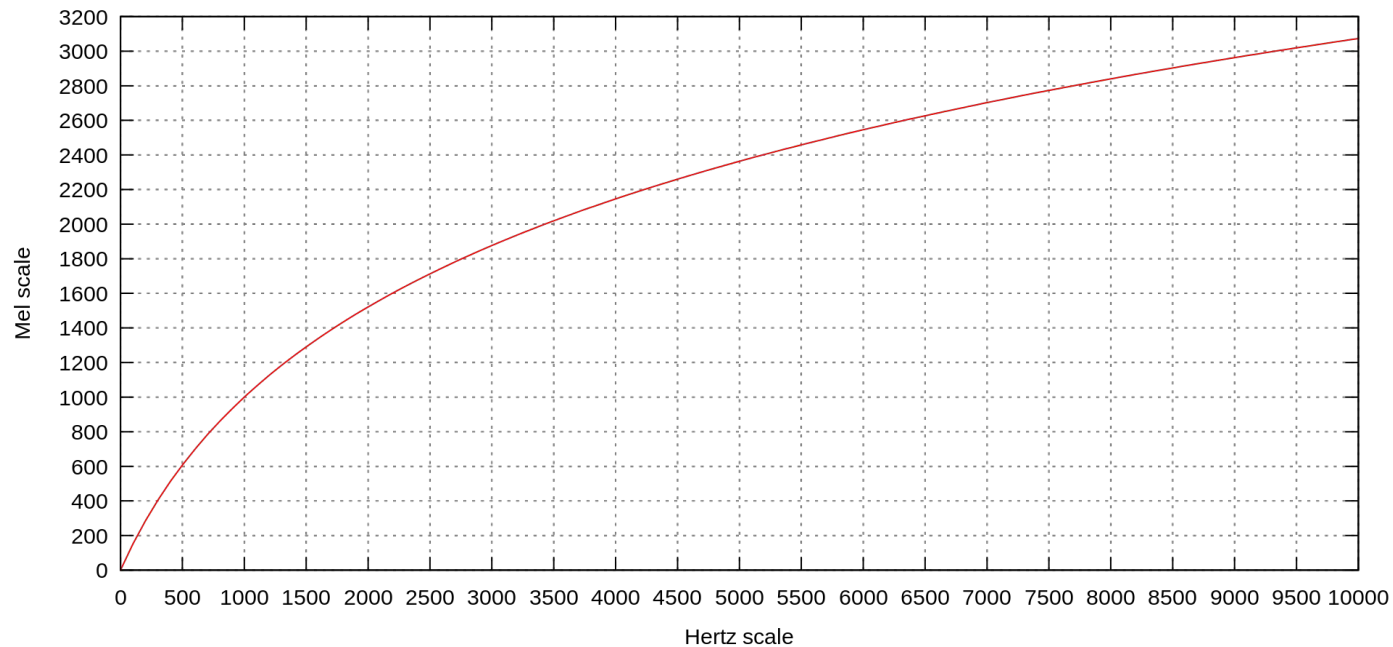
1. Мел-кепстральные коэффициенты (MFCC)
2. Коэффициенты линейного предсказания (LPC)
3. Перцептивное линейное предсказание (PLP)



MFCC

Шкала мелов

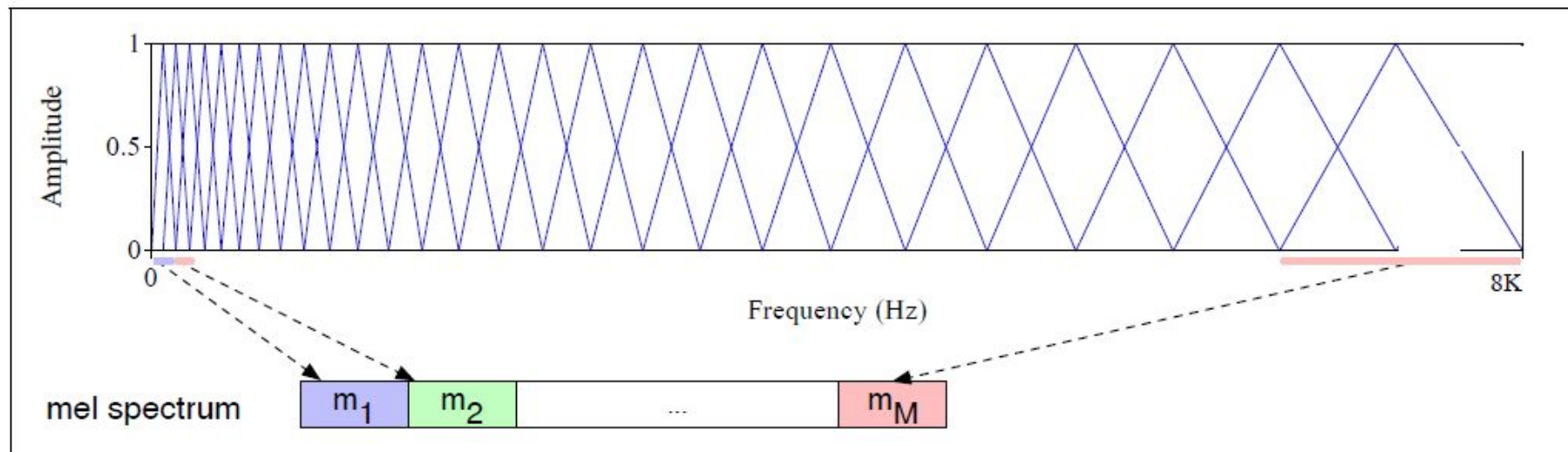
$$m = 2595 \log_{10} \left(1 + \frac{f}{700} \right)$$





MFCC

1. Создание банка мел-фильтров





MFCC

2. Логарифм

3. Дискретное косинусное преобразование (вычисление кепстра)

$$X_k = \frac{1}{2}x_0 + \sum_{n=1}^{N-1} x_n \cos\left[\frac{\pi}{N} \left(k + \frac{1}{2}\right) n\right] \quad \text{for } k = 0, \dots, N-1.$$

4. Берутся первые T коэффициентов результата.



LPC

LPC = Linear Predictive Coding

1. Попробуем предсказывать каждый следующий отсчёт как линейную комбинацию p предшествующих:

$$\tilde{x}[n] = \sum_{k=1}^p a_k x[n-k]$$

Тогда ошибка предсказания:

$$e[n] = x[n] - \tilde{x}[n] = x[n] - \sum_{k=1}^p a_k x[n-k]$$



LPC

2. Вычисляем p коэффициентов линейного предсказания с тем расчётом, чтобы ошибка предсказания была минимальной.
3. Эти коэффициенты – полюса передаточной функции:

$$H(z) = \frac{X(z)}{E(z)} = \frac{1}{1 - \sum_{k=1}^p a_k z^{-k}} = \frac{1}{A(z)}$$

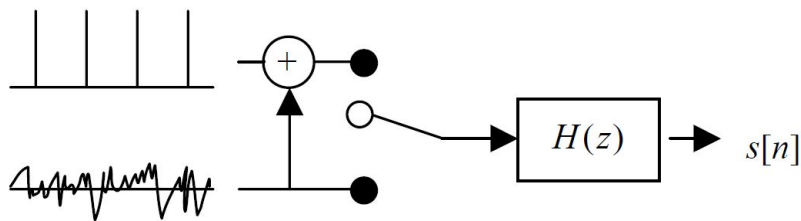
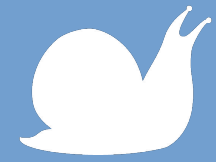


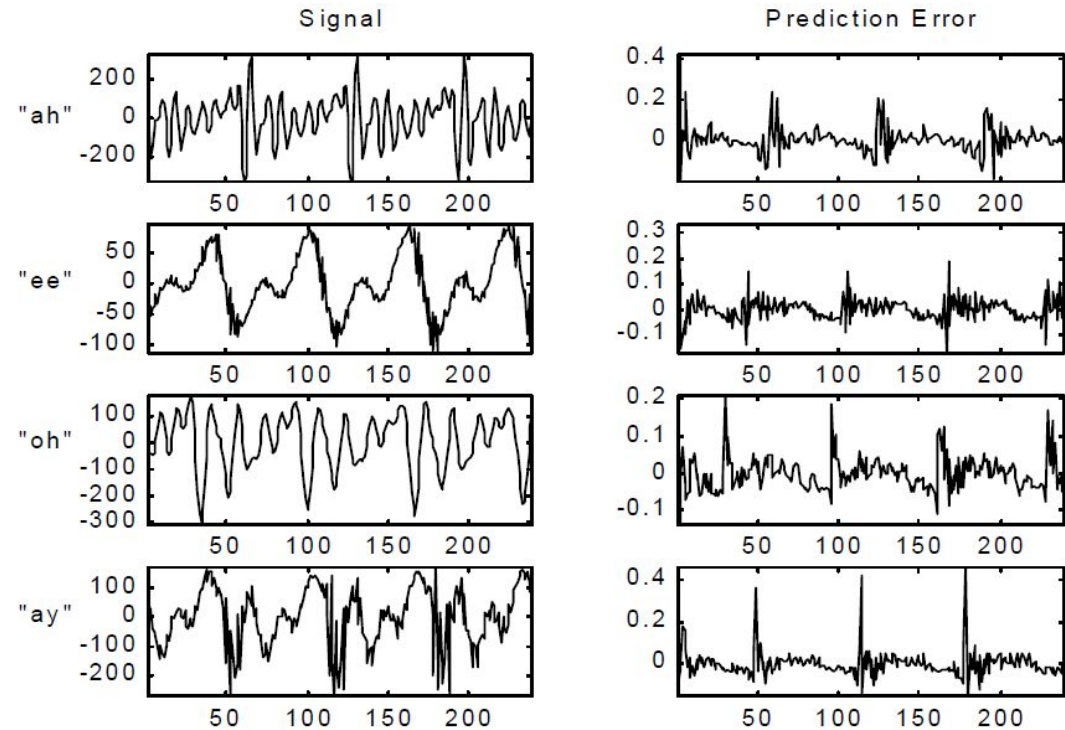
Figure 6.16 A mixed excitation source-filter model of speech.

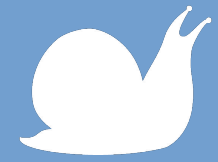


LPC

Ошибка предсказания:

1. Глухие участки – белый шум
2. Звонкие участки:





PLP

PLP = Perceptual Linear Prediction

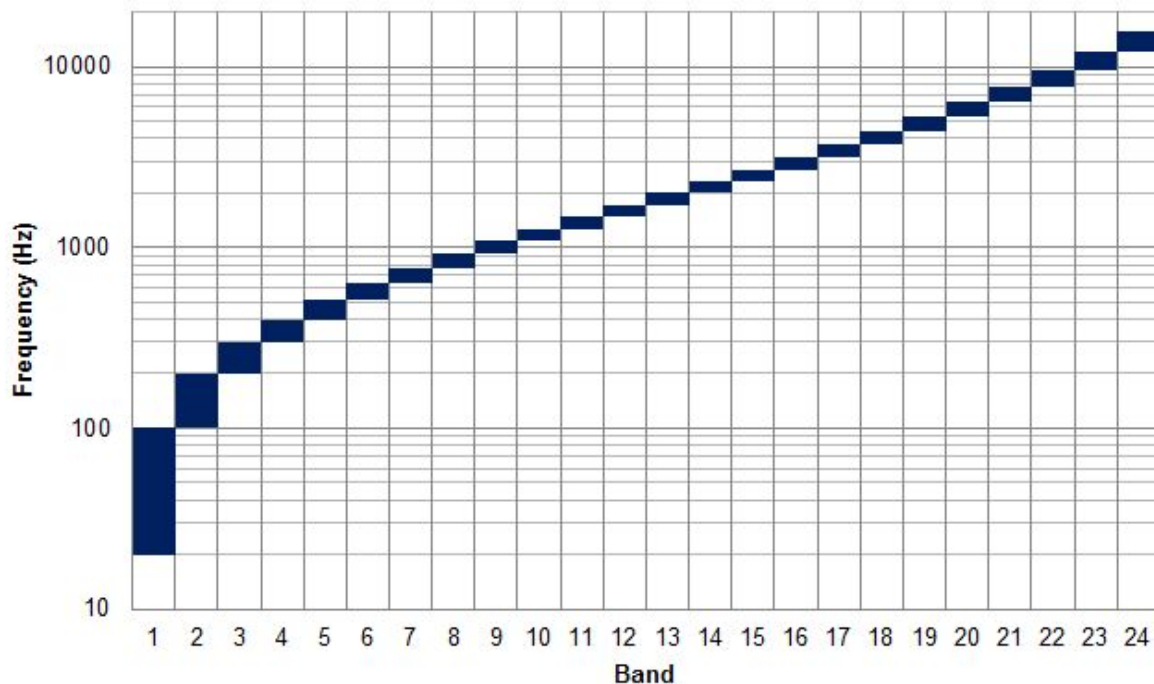
1. При вычислении LPC можно перевести спектр в шкалу барков:



Шкала барков

Шкала барков

$$\text{Bark} = 13 \arctan(0.00076f) + 3.5 \arctan((f/7500)^2)$$





PLP

2. Используем кривые равной громкости для выравнивания.
3. Вычисляем полюса фильтра, как раньше.



Дополнительные моменты

1. Дельта и дельта-дельта признаки:

$$\Delta \mathbf{c}_k = \mathbf{c}_{k+2} - \mathbf{c}_{k-2}$$

$$\Delta \Delta \mathbf{c}_k = \Delta \mathbf{c}_{k+1} - \Delta \mathbf{c}_{k-1}$$

2. Снижение размерности с помощью метода главных компонент (principal component analysis, PCA)

3. Гибридные признаки



Сравнение некоторых признаков

Feature set	Relative error reduction
13th-order LPC cepstrum coefficients	Baseline
13th-order MFCC	+10%
16th-order MFCC	+0%
+1st- and 2nd-order dynamic features	+20%
+3rd-order dynamic features	+0%



Полезные библиотеки

Окна:

<https://docs.scipy.org/doc/scipy/reference/signal.windows.html>

```
scipy.signal.windows.hann()  
scipy.signal.windows.hamming()  
scipy.signal.windows.kaiser()
```

Преобразование Фурье:

<https://docs.scipy.org/doc/scipy/tutorial/fft.html>

```
scipy.fft.fft()  
scipy.fft.ifft()
```



Полезные библиотеки

Вычисление MFCC:

1. librosa (pip install librosa)

<https://librosa.org/doc/main/generated/librosa.feature.mfcc.html>

`librosa.feature.mfcc()`

<https://librosa.org/doc/main/generated/librosa.feature.melspectrogram.htm>

`librosa.feature.melspectrogram()`

2. https://github.com/jameslyons/python_speech_features

`python_speech_features.base.mfcc()`



Наборы акустических признаков

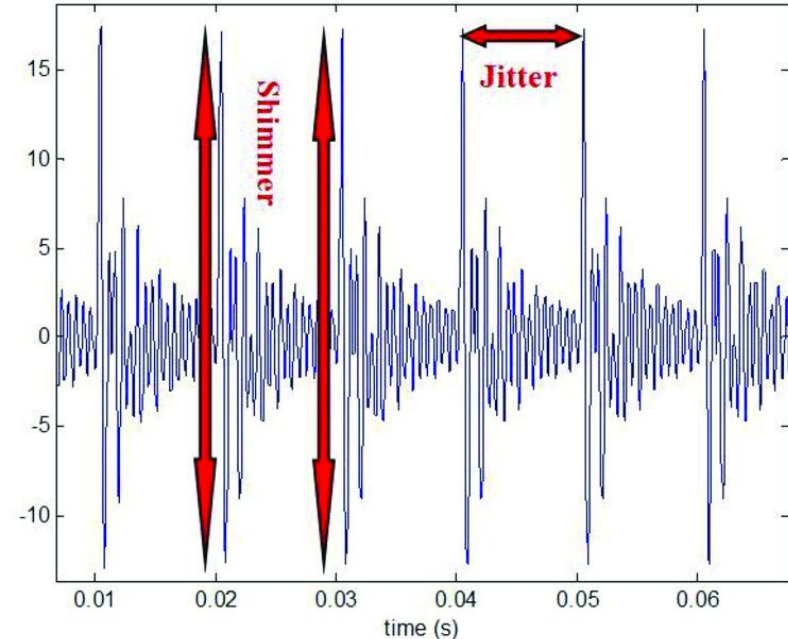
The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing

Frequency related parameters:

- **Pitch**, logarithmic F_0 on a semitone frequency scale, starting at 27.5 Hz (semitone 0).
- **Jitter**, deviations in individual consecutive F_0 period lengths.
- **Formant 1, 2, and 3 frequency**, centre frequency of first, second, and third formant
- **Formant 1**, bandwidth of first formant.

Energy/Amplitude related parameters:

- **Shimmer**, difference of the peak amplitudes of consecutive F_0 periods.
- **Loudness**, estimate of perceived signal intensity from an auditory spectrum.
- **Harmonics-to-Noise Ratio (HNR)**, relation of energy in harmonic components to energy in noise-like components.





Наборы акустических признаков

Spectral (balance) parameters:

- **Alpha Ratio**, ratio of the summed energy from 50–1000 Hz and 1–5 kHz
- **Hammarberg Index**, ratio of the strongest energy peak in the 0–2 kHz region to the strongest peak in the 2–5 kHz region.
- **Spectral Slope 0–500 Hz and 500–1500 Hz**, linear regression slope of the logarithmic power spectrum within the two given bands.
- **Formant 1, 2, and 3 relative energy**, as well as the ratio of the energy of the spectral harmonic peak at the first, second, third formant's centre frequency to the energy of the spectral peak at F_0 .
- **Harmonic difference H1–H2**, ratio of energy of the first F_0 harmonic (H1) to the energy of the second F_0 harmonic (H2).
- **Harmonic difference H1–A3**, ratio of energy of the first F_0 harmonic (H1) to the energy of the highest harmonic in the third formant range (A3).

6 temporal features are included:

- the **rate of loudness peaks**, i.e., the number of loudness peaks per second,
- the **mean length** and the **standard deviation** of continuously **voiced regions** ($F_0 > 0$),
- the **mean length** and the **standard deviation** of **unvoiced regions** ($F_0 = 0$; approximating pauses),
- the **number of continuous voiced regions per second** (pseudo syllable rate).

Спасибо за внимание!

