Автоматическое распознавание речи. Введение

П. А. Холявин

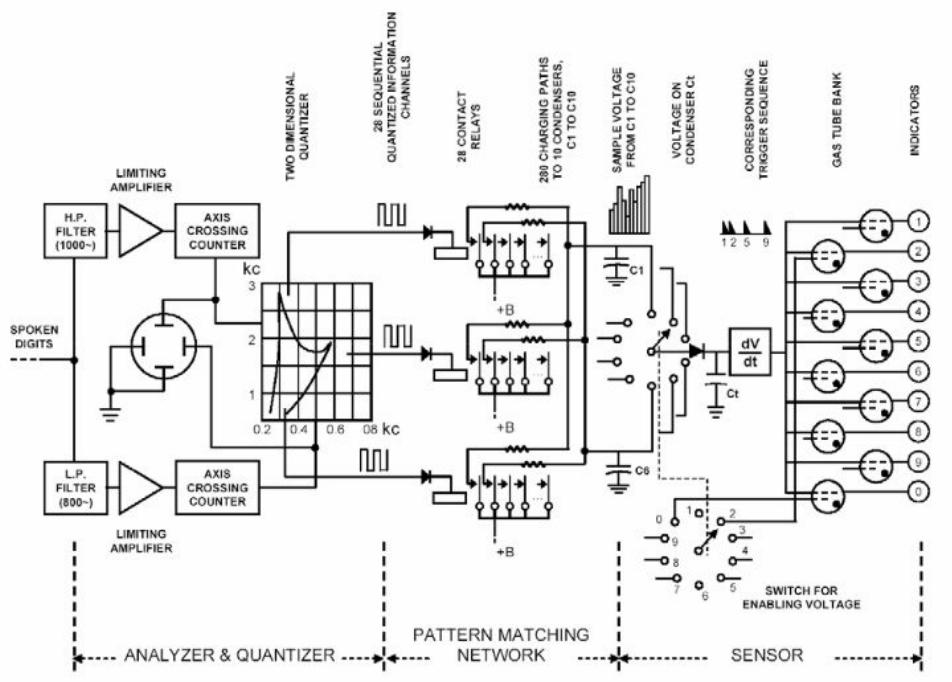
p.kholyavin@spbu.ru

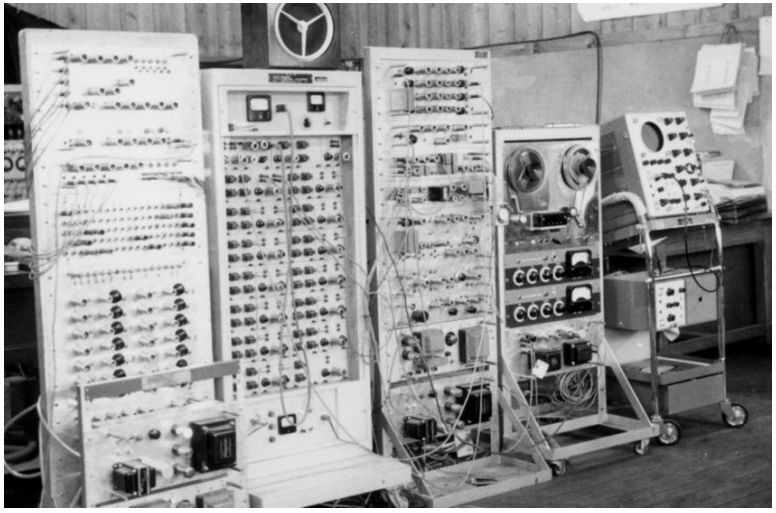
12.02.2025

«Распознавание» одного слова

1922

Распознавание отдельных слов





AUDREY, 1952
(Bell Laboratories)

Распознавание отдельных слов

IBM Shoebox, 1966

Whither speech recognition?

Received 20 June 1969 9.10, 9.1

Whither Speech Recognition?

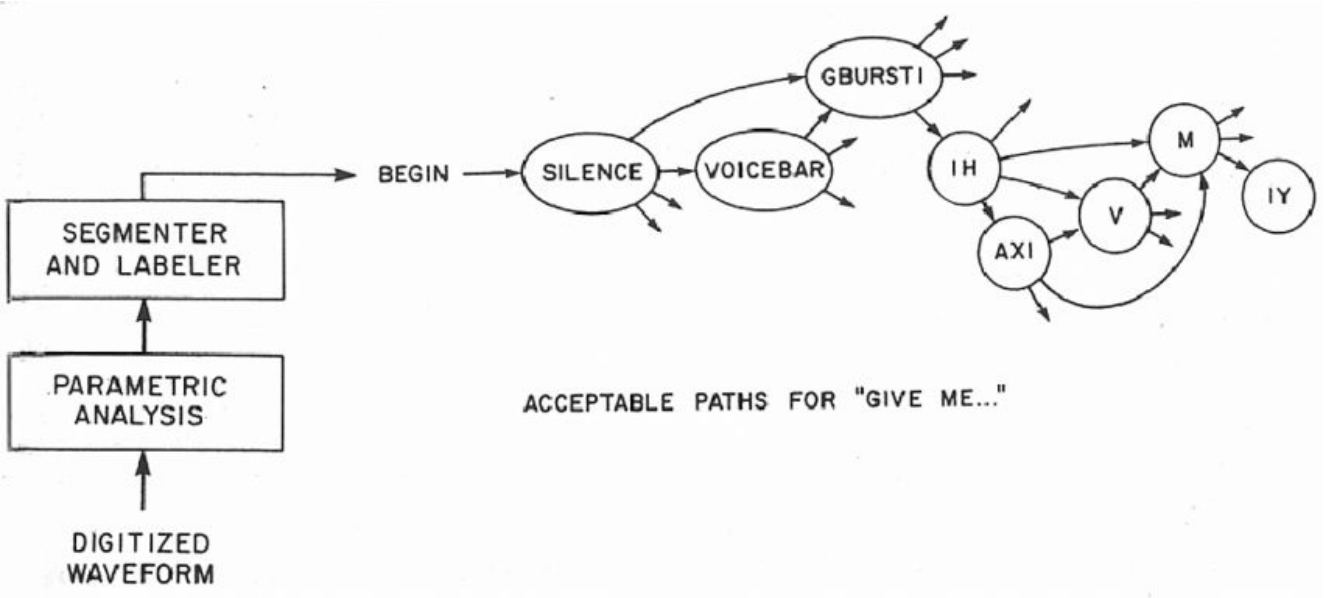
J.R. Pierce

Bell Telephone Laboratories, Inc., Murray Hill, New Jersey 07971

Speech recognition has glamor. Funds have been available. Results have been less glamorous. "When we listen to a person speaking—much of what we think we—hear is supplied from our memory. [W. James, *Talks to Teachers on Psychology and to Students on Some of Life's Ideals* (Holt, New York, 1889), p. 159]. General-purpose speech recognition seems far away Special-purpose speech recognition is severely limited. It would seem appropriate for people to ask themselves why they are working in the field and what they can expect to accomplish.

John R. Pierce, 1969 (Bell Labs)

Распознавание слитной речи



Carnegie-Mellon's HARPY
(1976)
-//- Hearsay-I (1976)

Распознавание слитной речи

1990-е:
- Доступ потребителей к системам распознавания речи
    - Dragon Dictate (1990)
    - IBM MedSpeak (1996)

- Работа над машинным пониманием речи

2010-е:
- Нейронные сети/глубокое обучение

Задача распознавания речи

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

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

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

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

Вариативность задачи

1. Количество слов в словаре:
   малый — единицы/десятки
   средний — сотни
   большой — тысячи/десятки тысяч
   сверхбольшой — сотни тысяч/миллионы

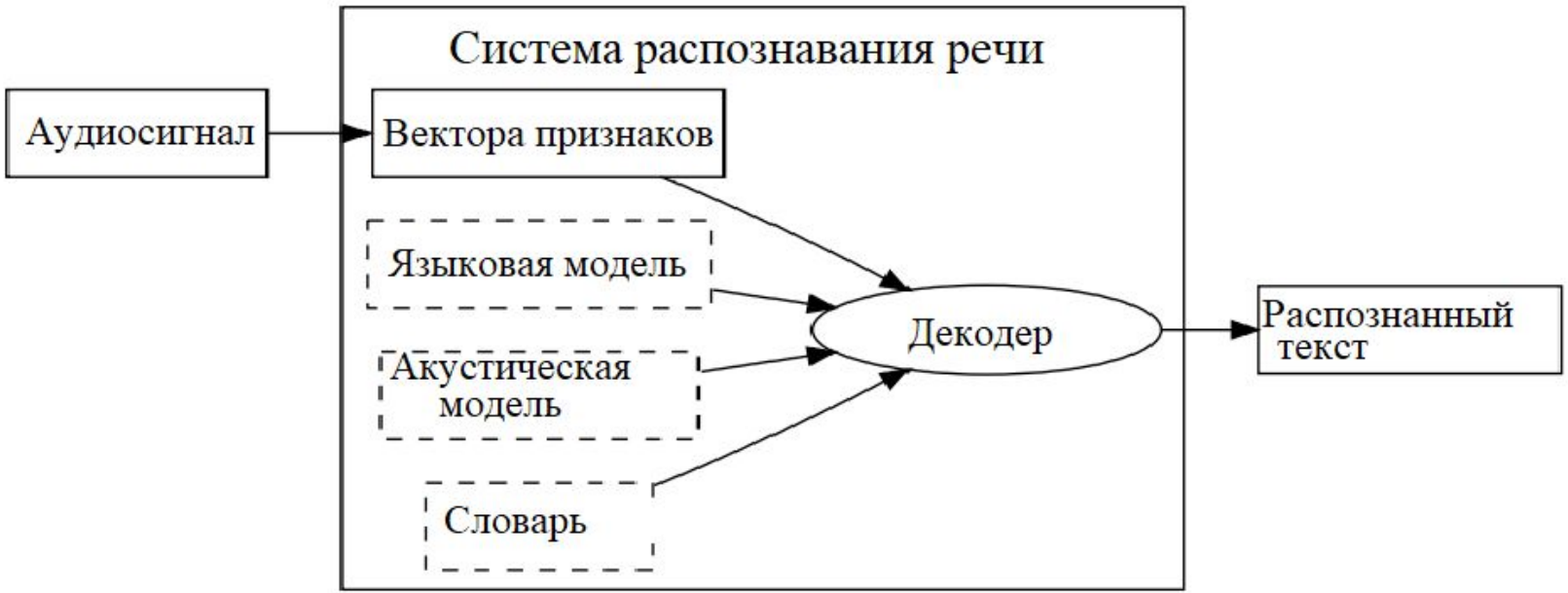
2. Дикторозависимость

3. Изолированные слова / слитная речь

4. Качество канала

5. Физиологические и лингвистические особенности

Части системы APP



Оценка работы APP

Word Error Rate

WER = 100 %

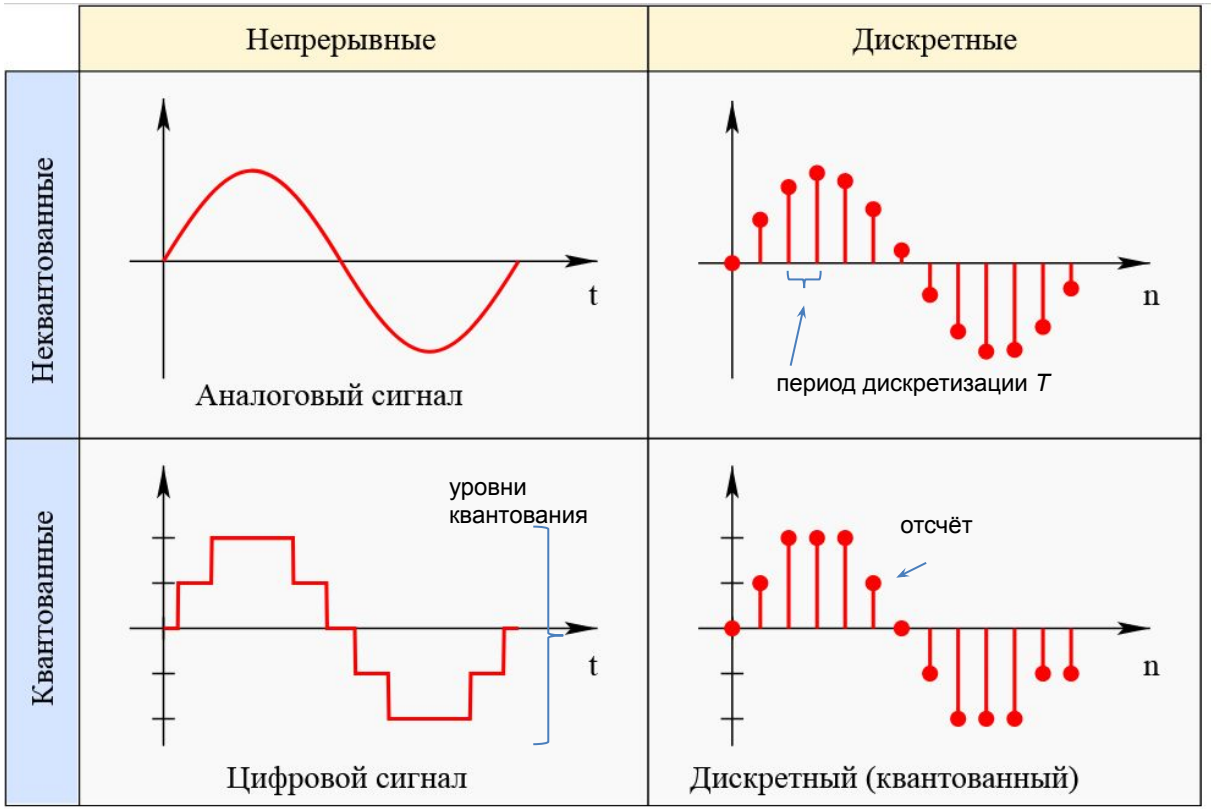
(Sentence Error Rate, Morpheme Error Rate, Phone Error Rate)

$$\mathrm{WER} = \frac{S + D + I}{N} \times 100$$

$$\mathrm{Accuracy} = 100 - \mathrm{WER}$$

RTF (Real Time Factor)

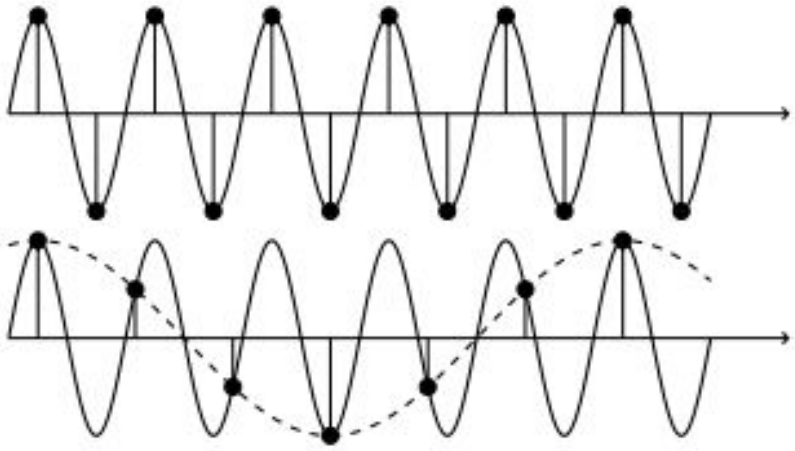
Аналого-цифровое преобразование



$F_{дискр} = 1 / T$

Теорема Котельникова

Любой сигнал s(t), спектр которого не содержит составляющих с частотами выше некоторого значения f, может быть без потерь представлен в виде дискретного сигнала с частотой дискретизации F >= 2f (частота Найквиста).



Частотный анализ сигнала

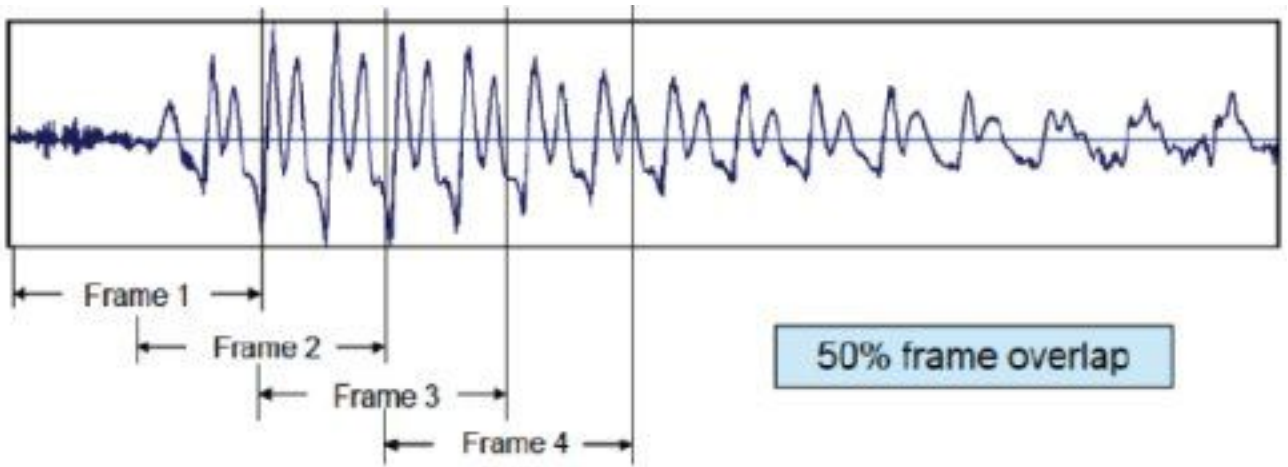
1. Дискретное преобразование Фурье (частотный спектр дискретного сигнала)

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn}$$

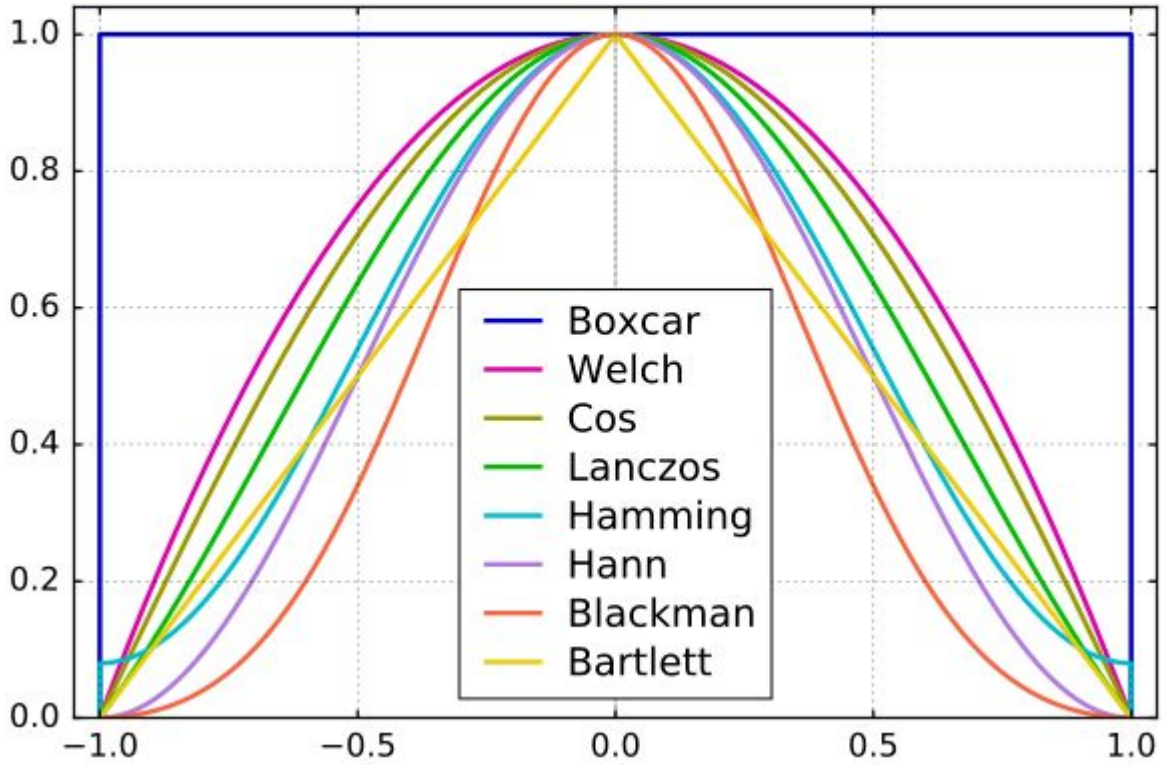
```
from scipy.fft import fft
```

Оконный метод

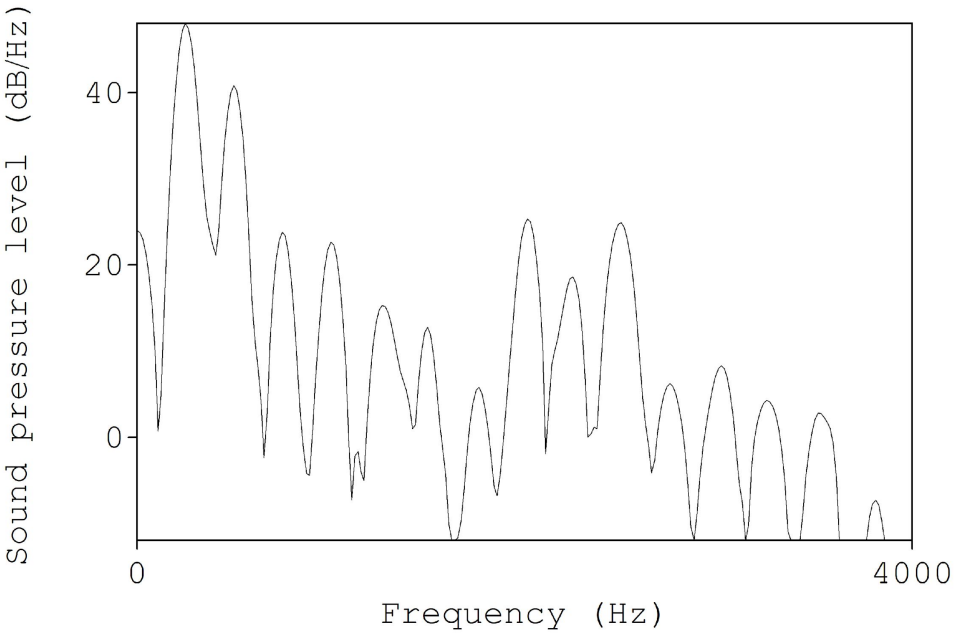
1. Параметры: длина окна и шаг окна

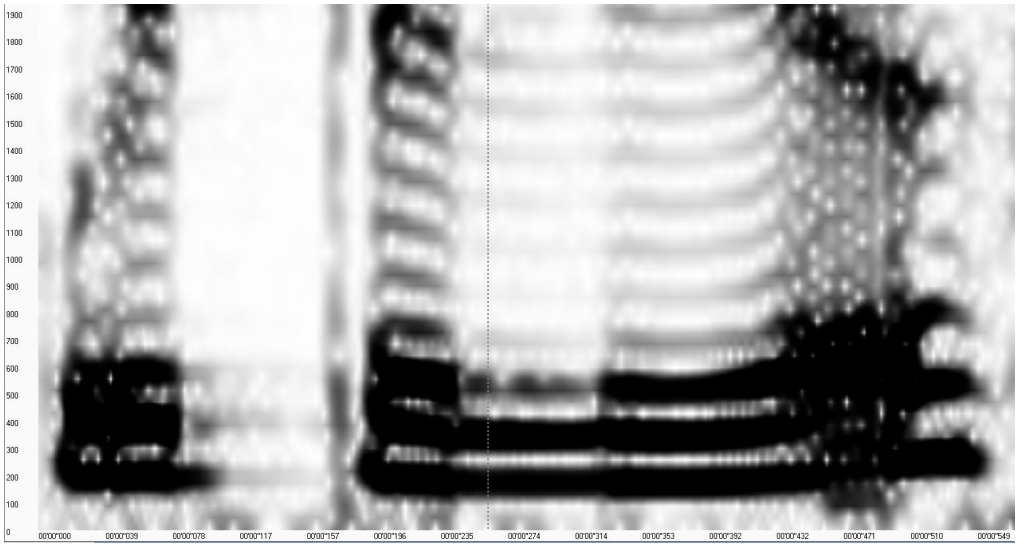


Оконные функции



Спектрографический анализ





Спасибо за внимание!