



# Text to Speech

Калиновский И.А.  
Just AI, к.т.н.



# План блока

---

1. История создания говорящих машин
2. Системы синтеза речи на основе DNN
3. Нейронные вокодеры
4. Современные задачи и проблемы TTS

# Что такое синтез речи?

Синтез речи – восстановление формы речевого сигнала по его параметрам,  
в узком смысле – преобразование текста и другой информации в звучащую речь

## Область применения:

голосовые ассистенты



автоматизация дистанционного  
обслуживания (IVR)



автоматическое озвучивание книг,  
фильмов, игровых персонажей, рекламы

человеко-машинные интерфейсы



## Поставщики технологии:



# Голосовые ассистенты

## В России

50 млн

россиян пользуются  
голосовыми ассистентами  
минимум раз в месяц



Алиса, Яндекс — 45 млн активных пользователей в месяц (28% из них используют Алису в автомобилях)



Google Ассистент — около 7 млн активных пользователей в России



Siri, Apple — около 8 млн пользователей в России

По оценкам Just AI

## В мире

3,25 млрд

голосовых ассистентов используется в мире в начале 2019 года.

По данным Juniper Research

8 млрд

ассистентов будет использоваться в мире к 2023 году. По мнению аналитиков, на одного пользователя будет приходиться по 2-3 ассистента.



### Ассистенты в доме

22,9% умные колонки  
11,4% PC и ноутбуки  
11,3% наушники  
7,4% телевизор  
7,3% умные часы  
5,3% игровые консоли



### 58% на смартфонах

84% поиск в интернете  
70% маршруты  
66% звонок по телефону  
56% отправка смс  
55% поиск ресторана

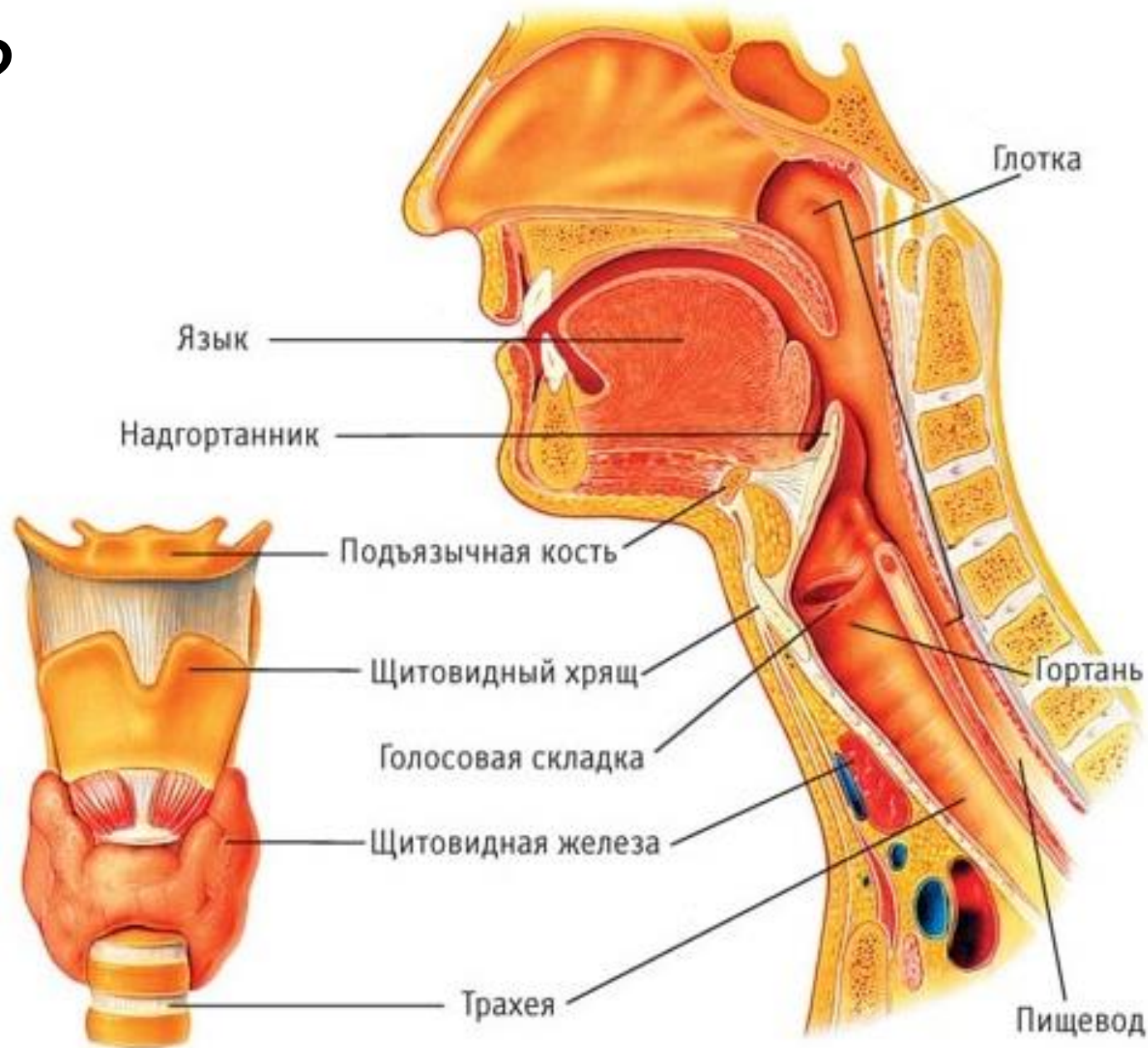
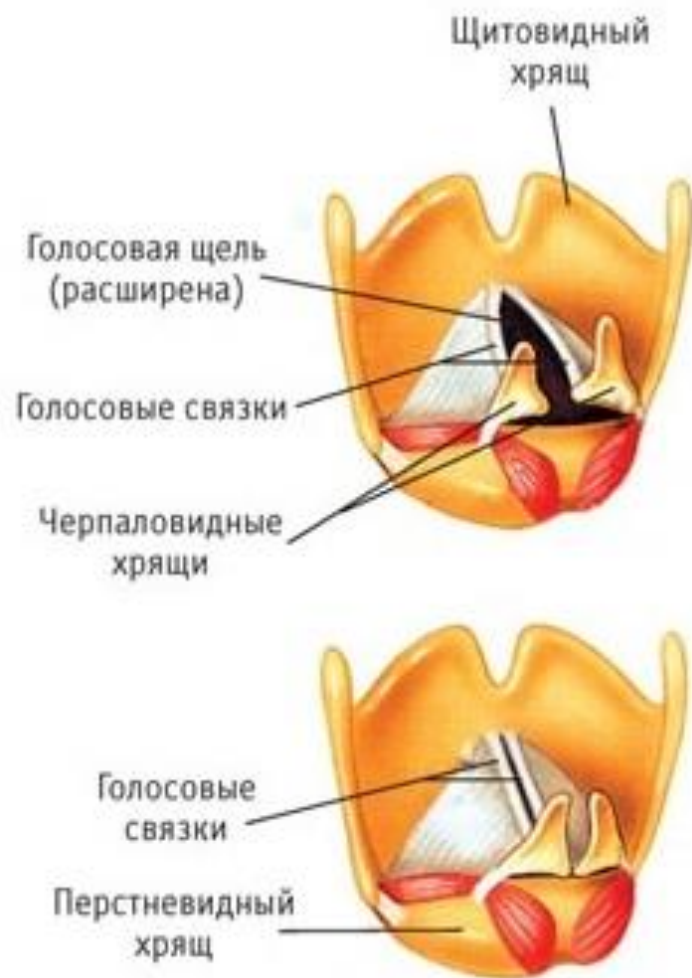


### 45% в машинах

73,7% звонок по телефону  
50,3% маршруты и навигация  
41,2% отправка смс  
28,7% музыкальные сервисы

Исследование Voicebot.ai на примере рынка США, 2018—2019 год

# Как мы говорим?







# Характеристики голоса

---

**Тембр голоса** — неповторимая индивидуальная окраска, которая обусловлена строением речевого аппарата

**Сила голоса** — громкость, зависящая от активности работы органов дыхания и речи

**Высота голоса** — это его способность к тональным изменениям, то есть его диапазон

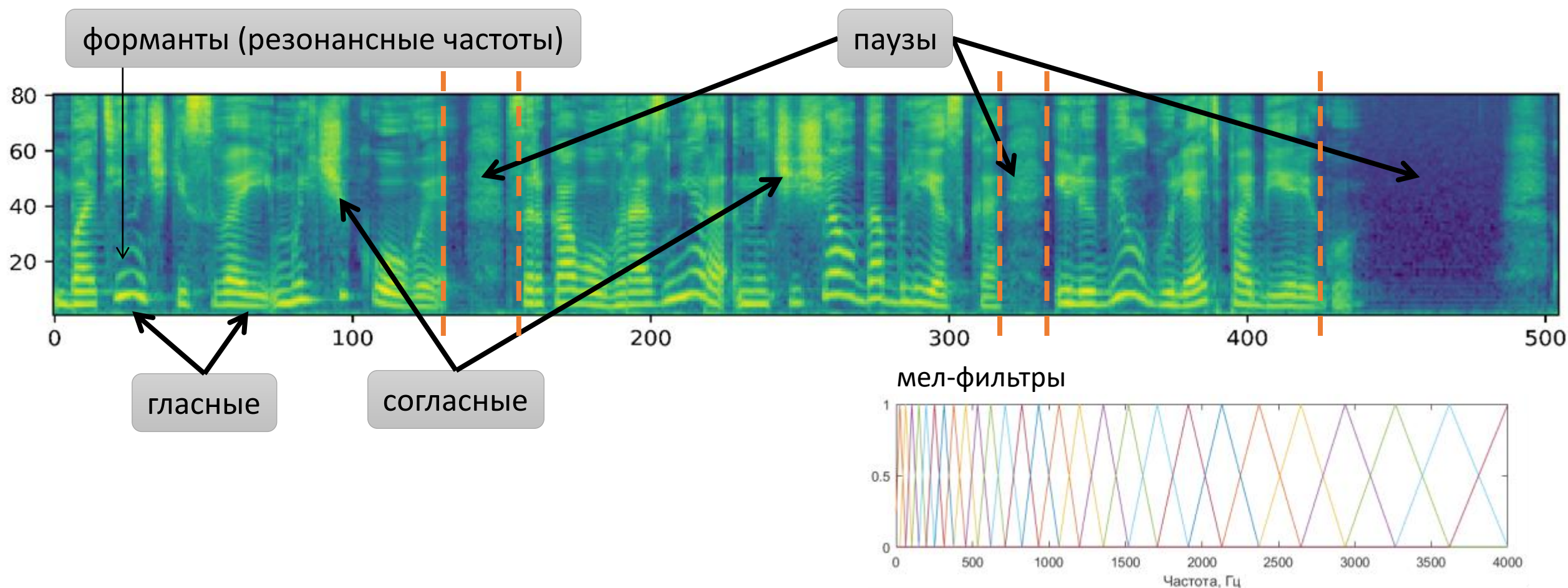
**Благозвучность голоса** — чистота его звучания, отсутствие неприятных призвуков (хрипоты, сиплости, гнусавости)

**Темп речи** — скорость произнесения элементов речи (звуков, слогов, слов)

**Интонация** — это ритмомелодический строй речи. Интонация включает: высоту тона, силу звучания, мелодику, темп, ударения и паузы.

# Спектр речевого сигнала

«Ботиночки свои мальчиковые, сорокового размера, начищал до блеска и любил гулять по берегу.»



# Механические синтезаторы

---

Гласный орган Х. Кратценштейна (1780 г.)



Резонаторы моделируют 5 звуков: а, э, и, о, у



# Механические синтезаторы

---

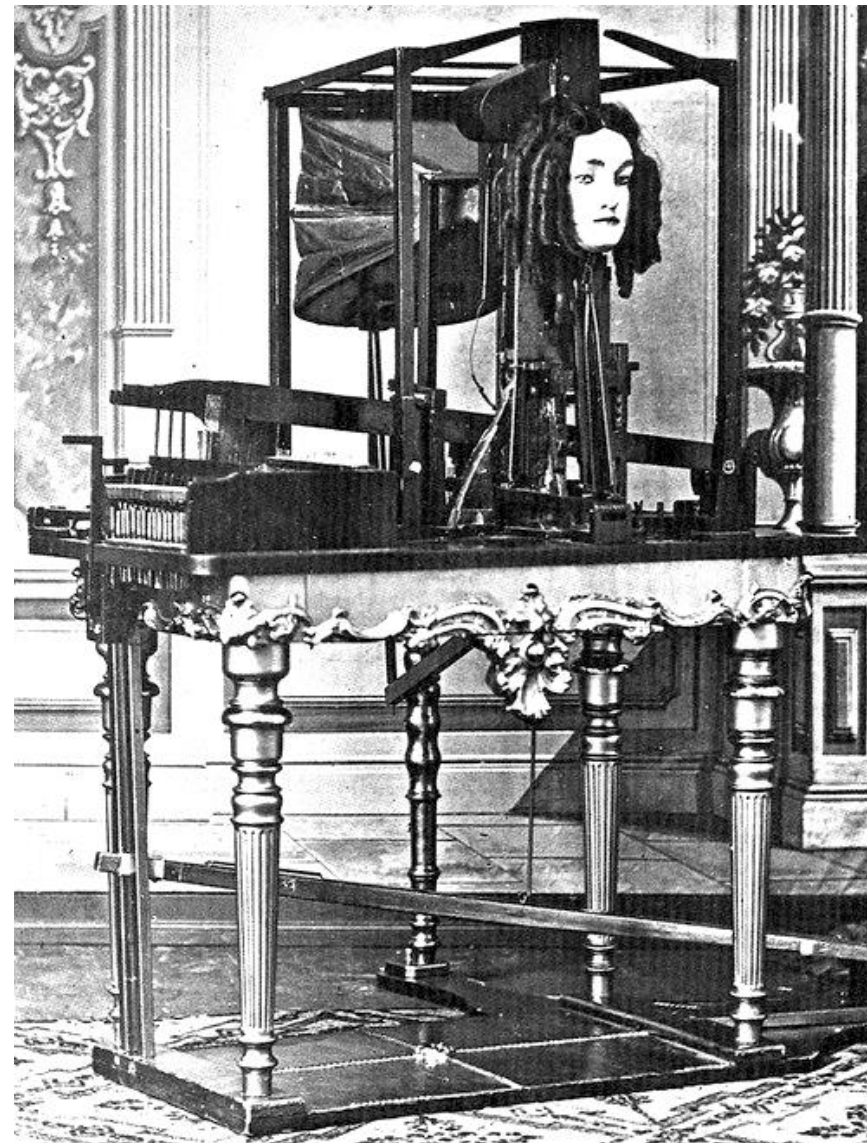
Машина фон Кемпелена (1788 г.)



# Механические синтезаторы

Говорящая машина Фабера «Эвфония» (1830 г.)

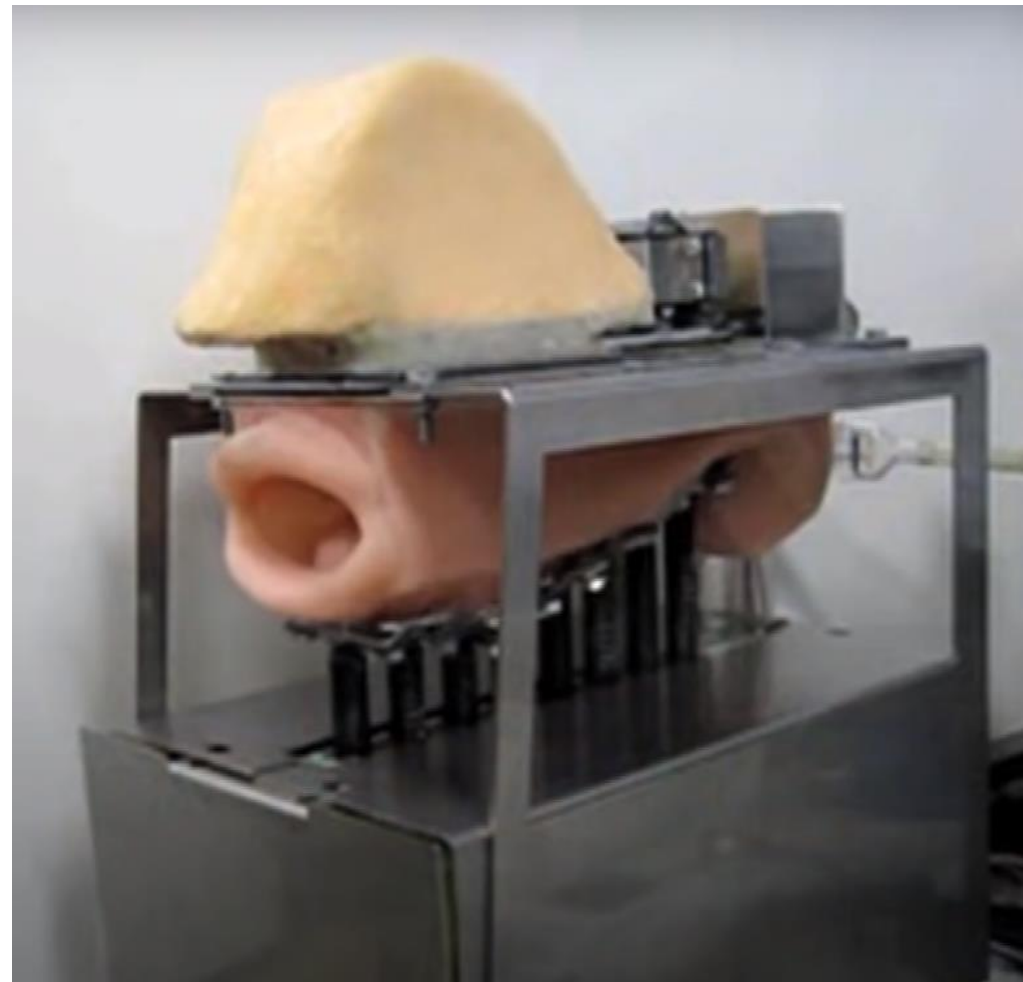
Вследствие неуклюжести и неточности устройства и сами звуки, извлекаемые из машины, грубы, крикливы, монотонны и не всегда схожи со звуками настоящей человеческой речи. Тем не менее из известных в науке говорящих приборов, машина Фабера являлась одним из наиболее удачных.



# Механические синтезаторы

Современная вариация, Япония, 2011

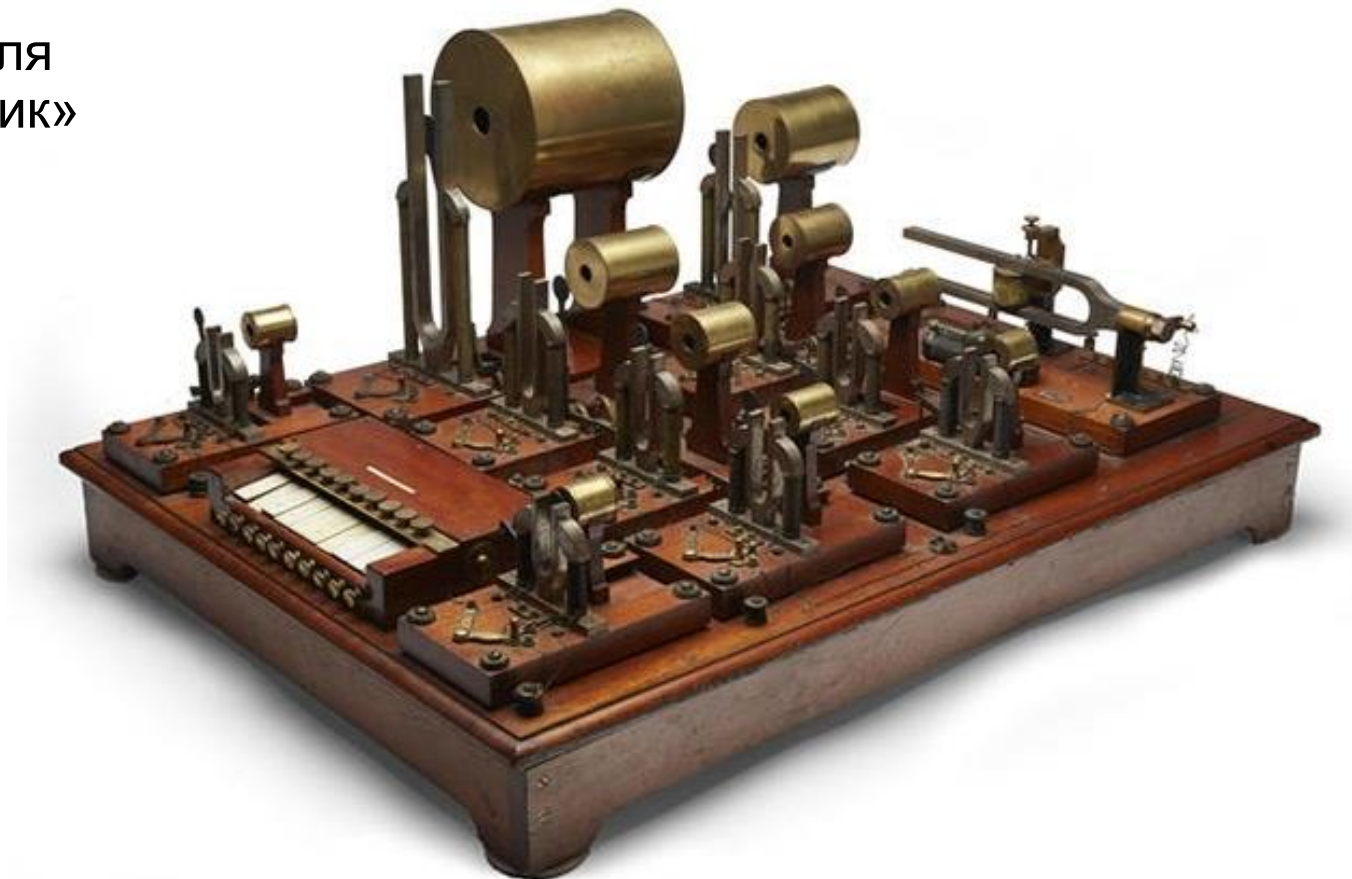
Устройство может имитировать любые звуки и позволяет синтезировать речь на редких языках



# Электрические синтезаторы

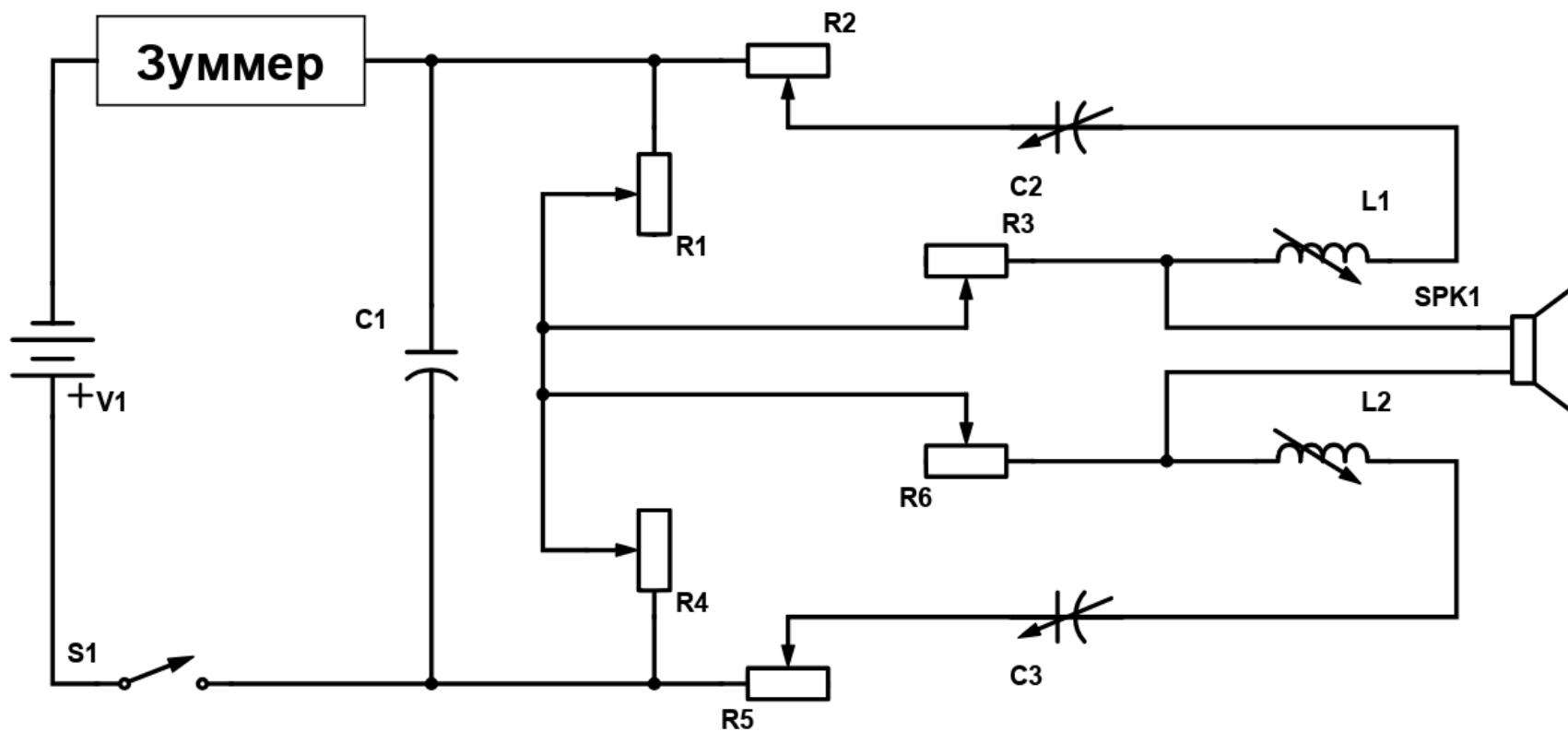
---

«Большой аппарат Гельмгольца для соединения тембров из 10 гармоник»  
или синтезатор Фурье –  
первый прообраз современного  
синтезатора



# Электрические синтезаторы

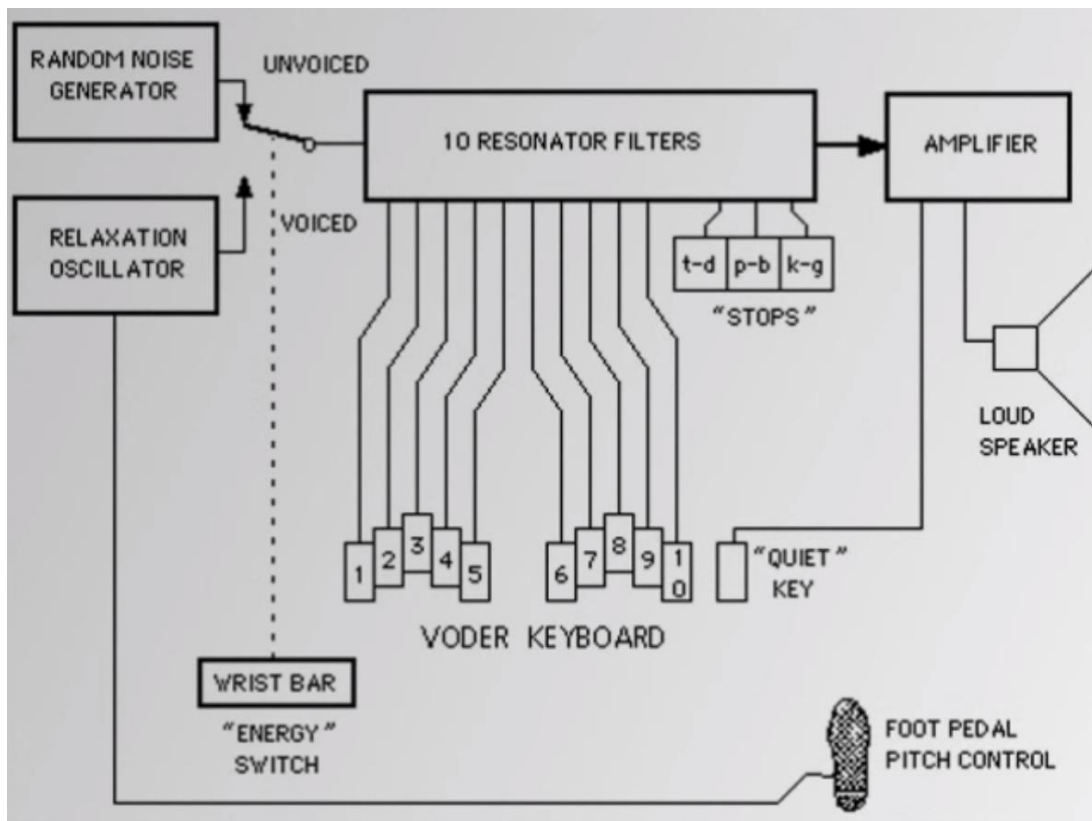
Синтезатор Д. Стюарта





# Электрические синтезаторы

Водер Г. Дадли, 1939



# Электрические синтезаторы

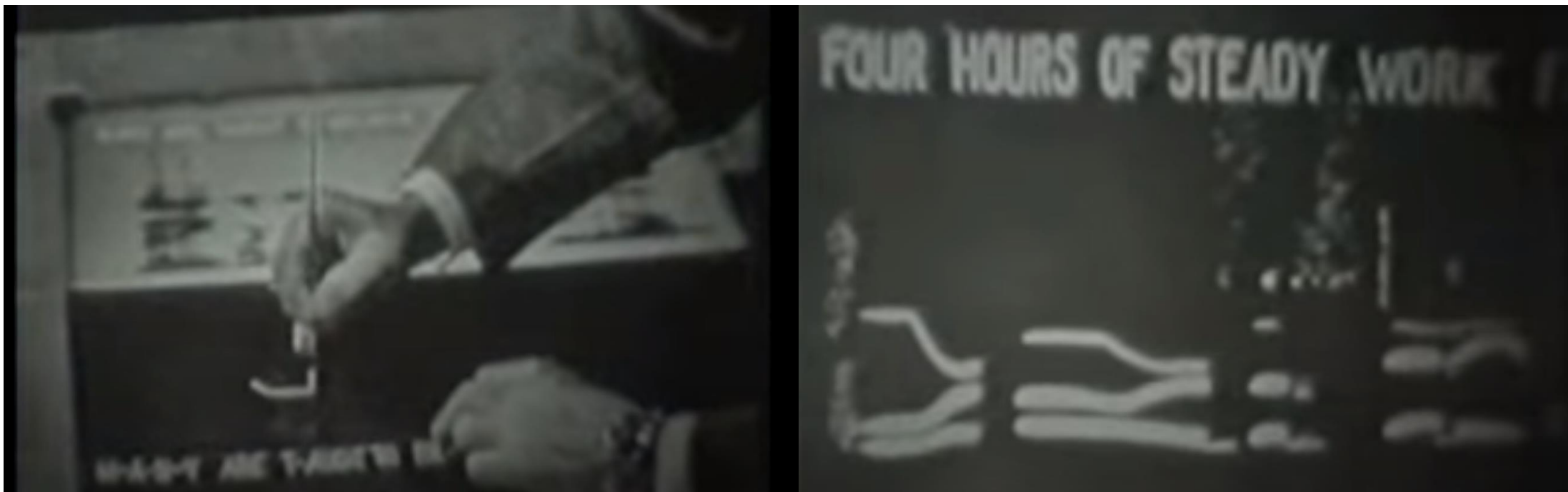
## Акустический спектрограф



# Электрические синтезаторы

---

Синтезатор Л. Шотта (Bell Labs), 1946

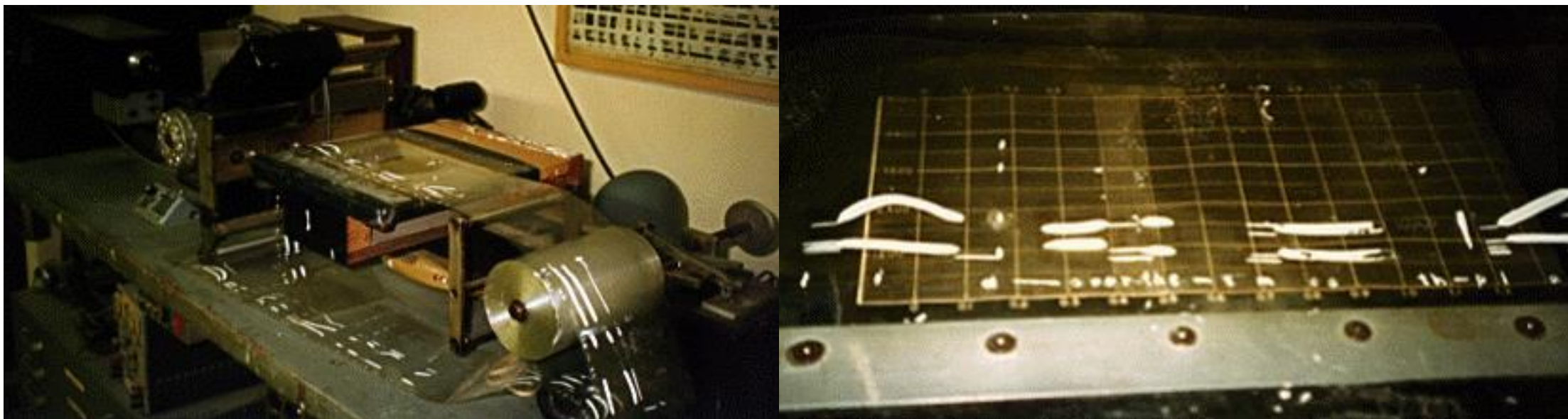




# Электрические синтезаторы

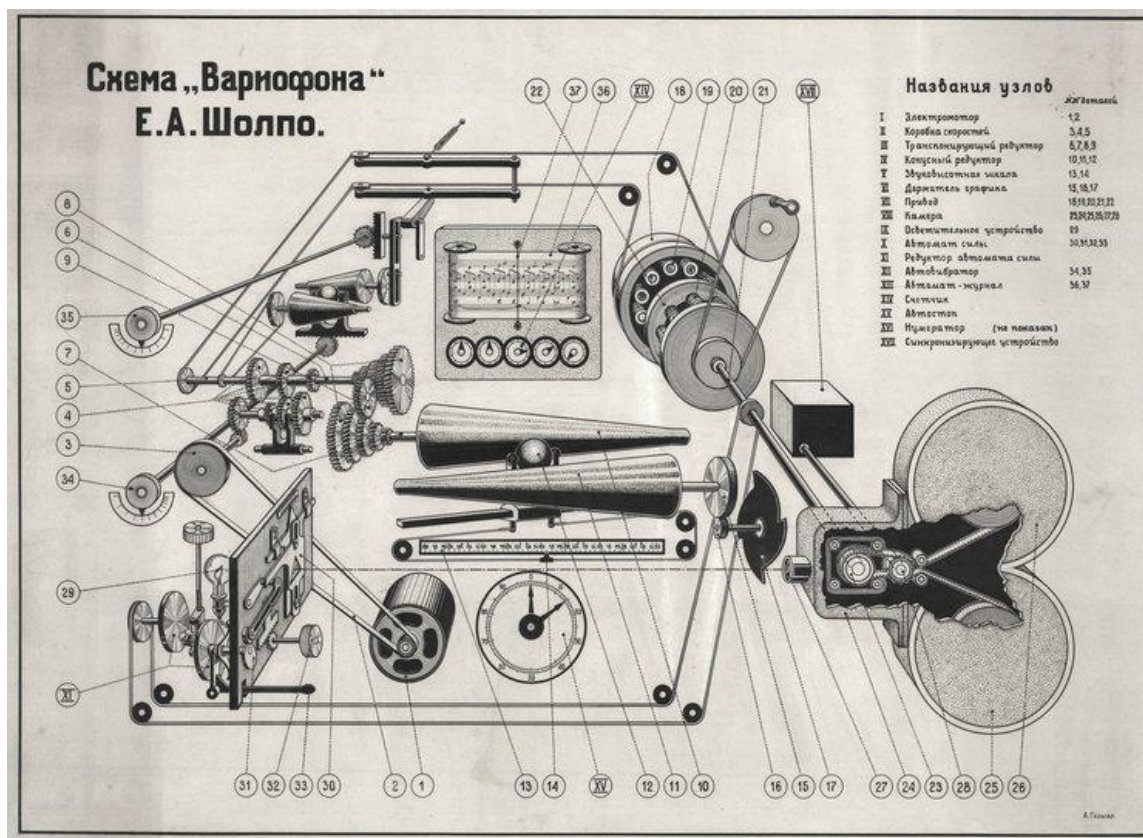
---

Pattern Playback Ф. Купера, 1949



# Электрические синтезаторы

Вариофон Е. Шолпо, 1942





# Синтезаторы I поколения

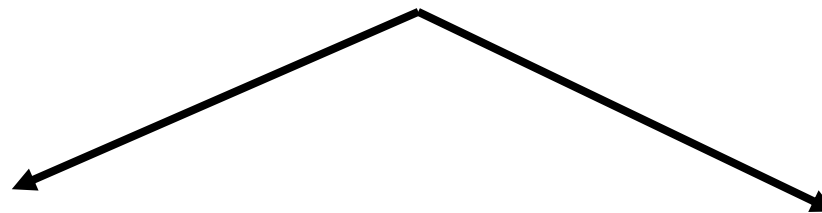
---

Артикуляционные

Акустические

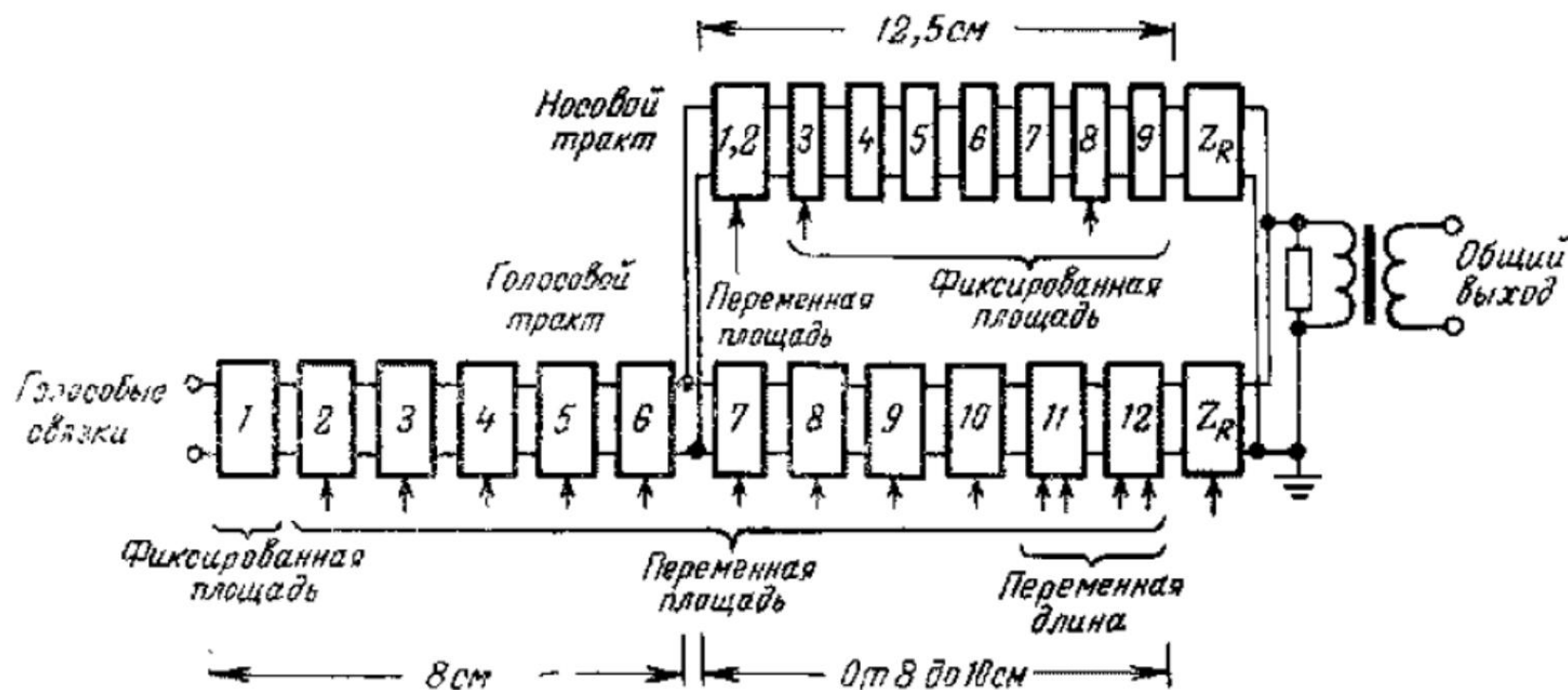
Формантный синтез

LPC-синтез



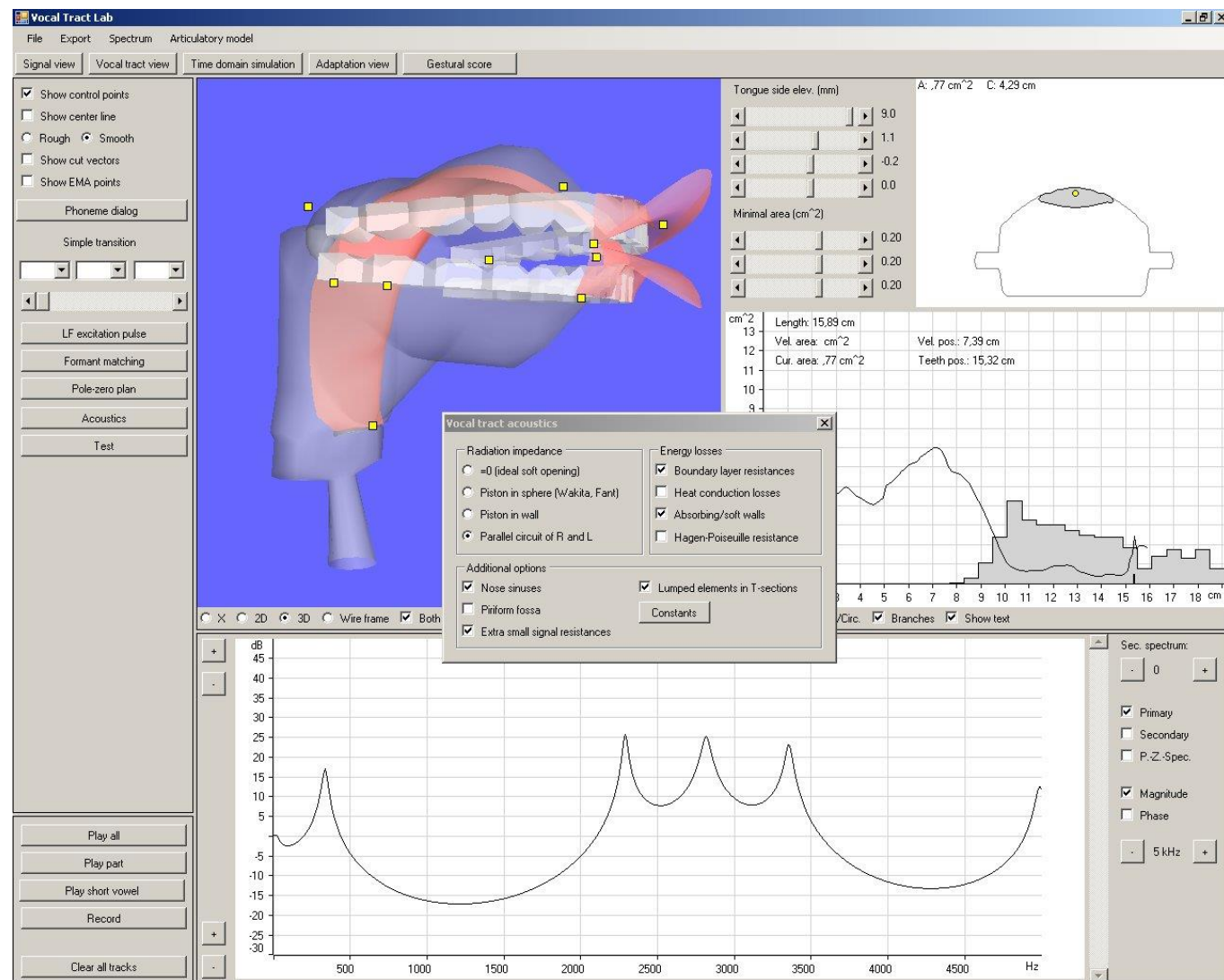
# Артикуляционные синтезаторы

DAVO (Dynamic Analog of the VOcal tract) Д. Розен, 1958



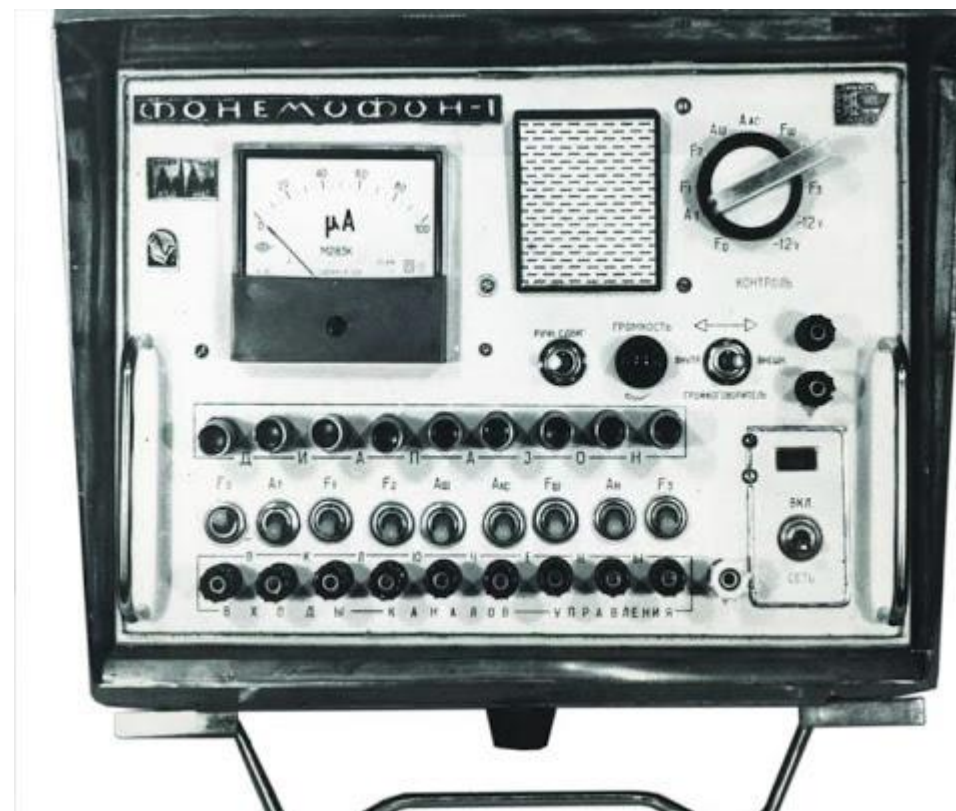
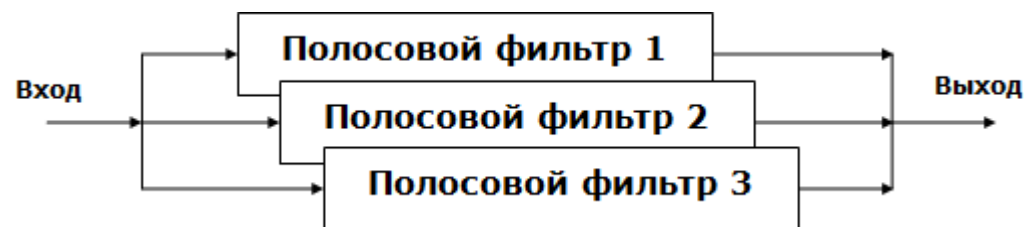
# Артикуляционные синтезаторы

VocalTractLab 2.3, 2020



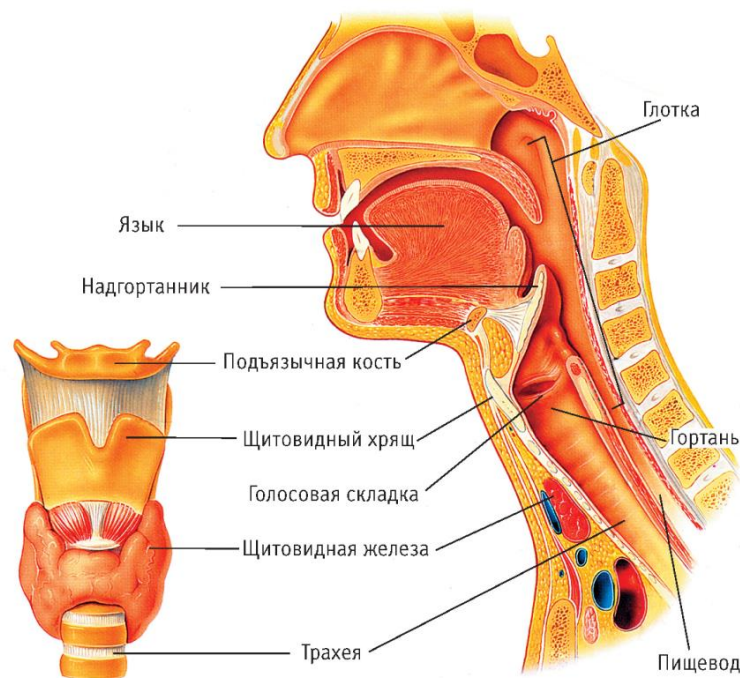
# Формантный синтез

Первая модель формантного синтезатора русской речи «Фонемофон-1», 1970

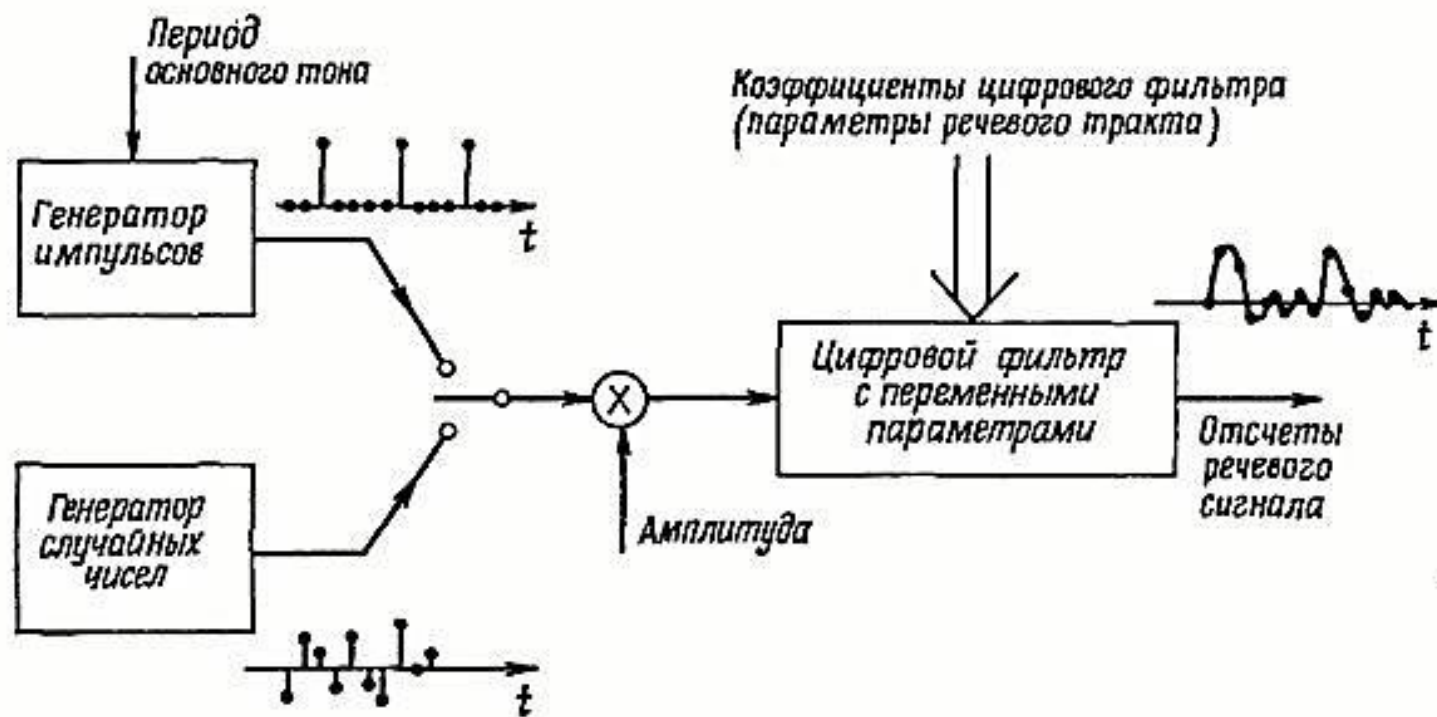


# КЛП-синтезаторы

Производство речи можно рассмотреть как процесс фильтрации: речевой тракт выступает в качестве фильтра, усиливающего только те частоты, порожденные голосовыми связками, которые совпадают с его собственной частотой.



Строение речевого тракта



Цифровая модель образования речи





# Синтезаторы II поколения

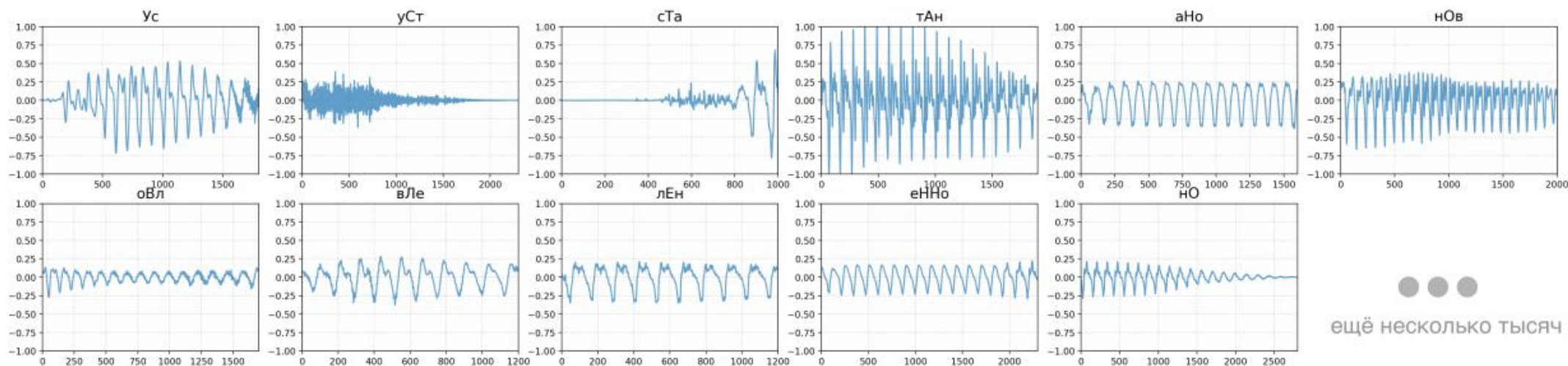
---

Синтезаторы 2-го поколения характеризуются появлением блока лингвистической обработки текста.

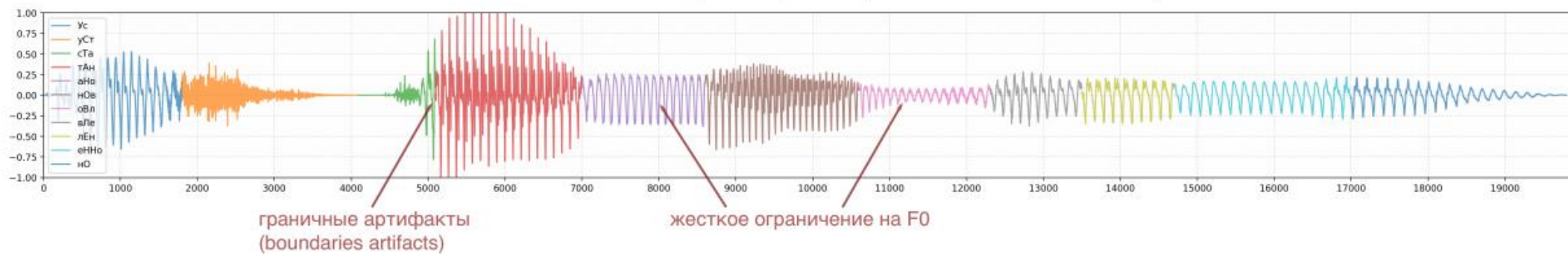
Первая полноценная модель TTS была создана в Японии в 1968 г. Норико Умеда

# Синтезаторы III поколения

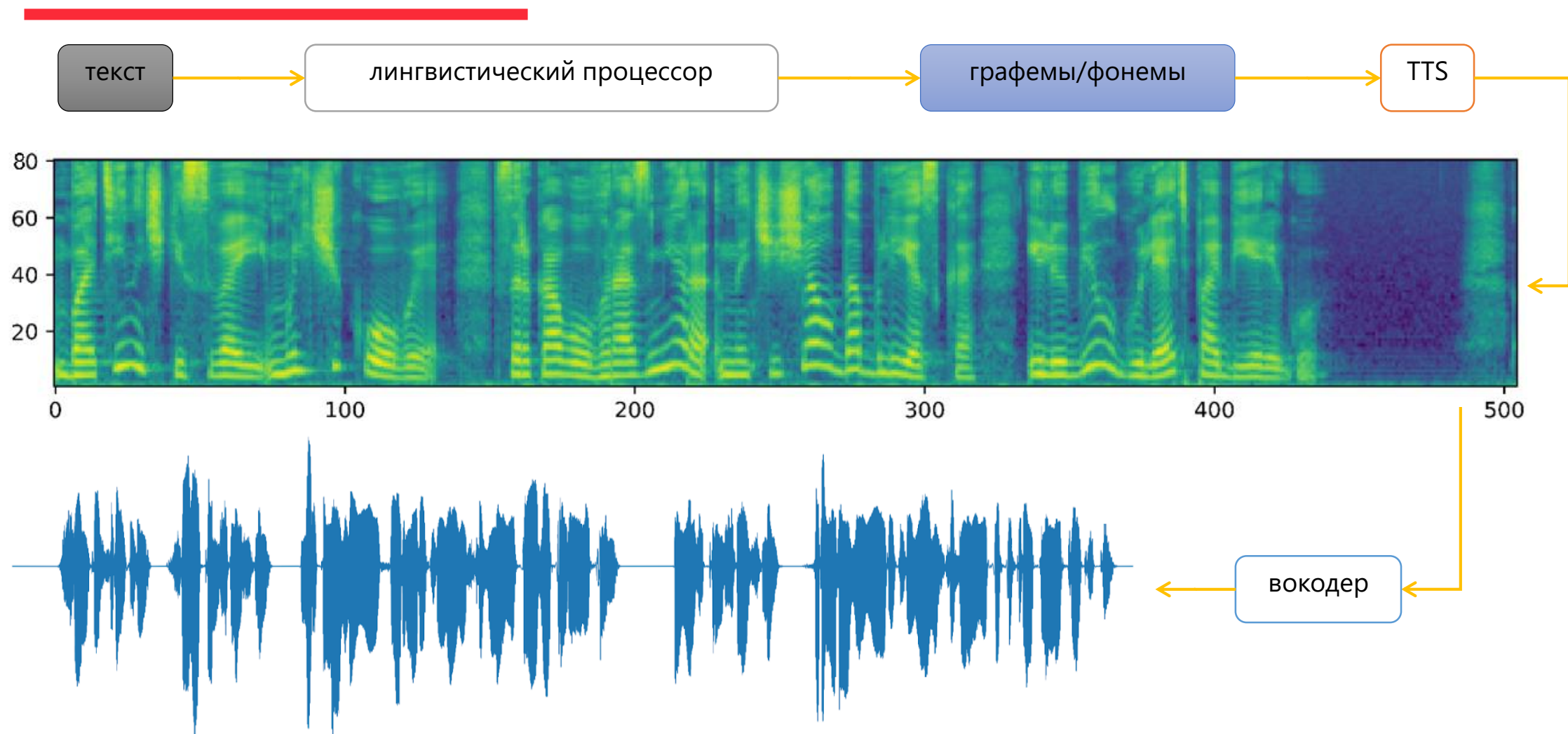
Коллекция звуков (units database)



Конкатенация звуков (units concatenation)



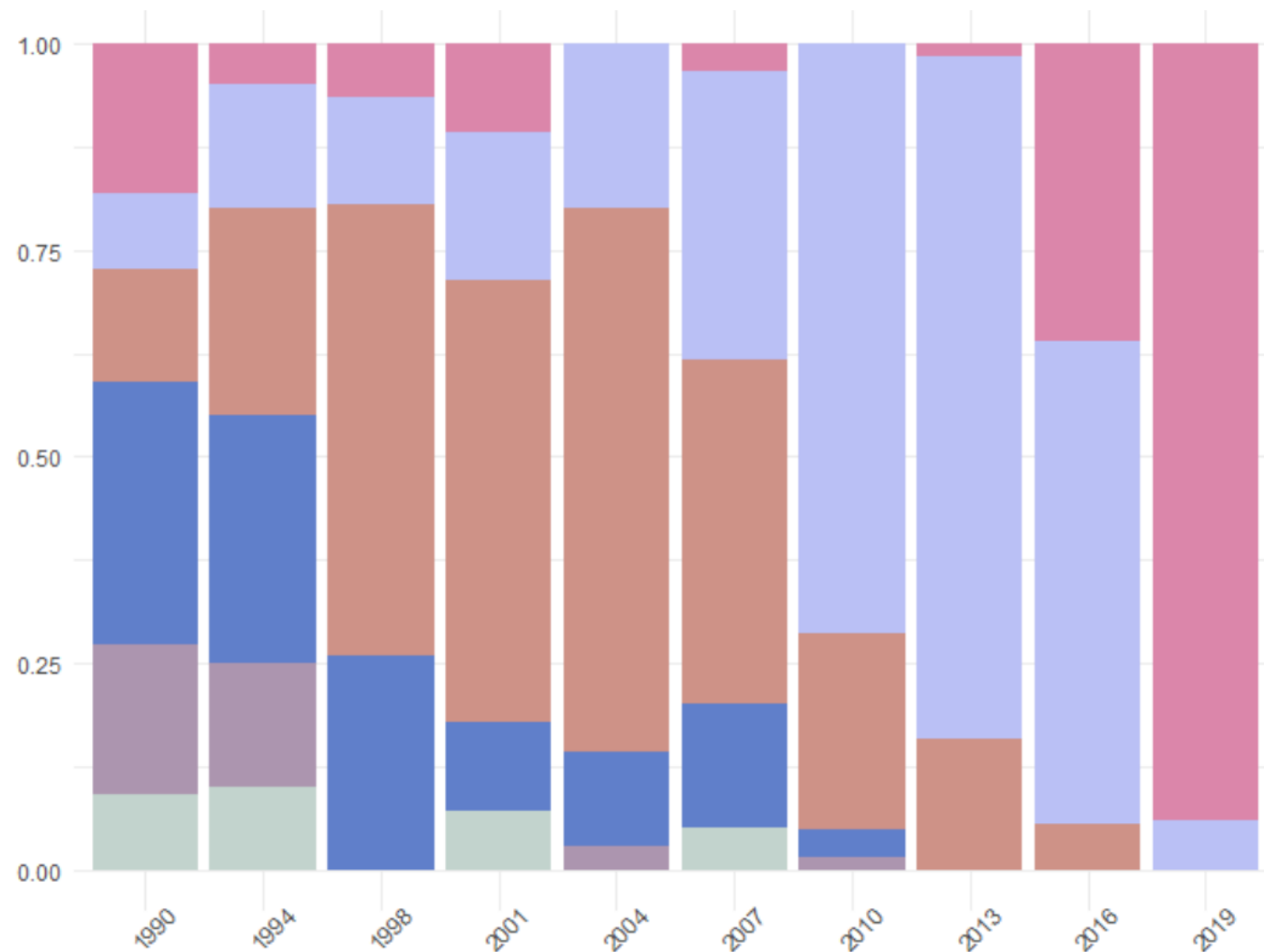
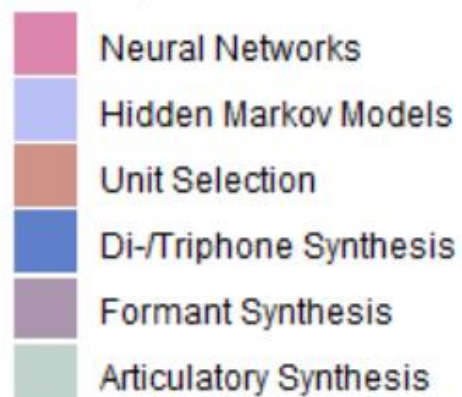
# Синтезаторы IV поколения



# Технологии синтеза речи

Обзор статей, принятых на  
Speech Synthesis Workshop

## Concepts



# Фонетическое членение речи

Непрерывный поток звучащей речи является структурированным, он членится на разные по объему отрезки, фонетические единицы, которые образуют иерархическую систему.

*Сегодня светит солнце, но дует ветер.*

<b>синтагмы</b>	сегодня светит солнце	но дует ветер
<b>слова</b>	сегодня   светит   солнце	но   дует   ветер
<b>слоги</b>	се-го-дня   све-тит   сол-нце	но   ду-ет   ве-тер
<b>фонемы</b>	s'/i <sub>1</sub> -v/o <sub>0</sub> -d'/n'/a <sub>4</sub>   s/v'/e <sub>0</sub> -t'/i <sub>4</sub> /t   s/o <sub>0</sub> -n/c/y <sub>4</sub>	n/o <sub>1</sub>   d/u <sub>0</sub> -j/i <sub>4</sub> /t   v'/e <sub>0</sub> -t'/i <sub>4</sub> /r



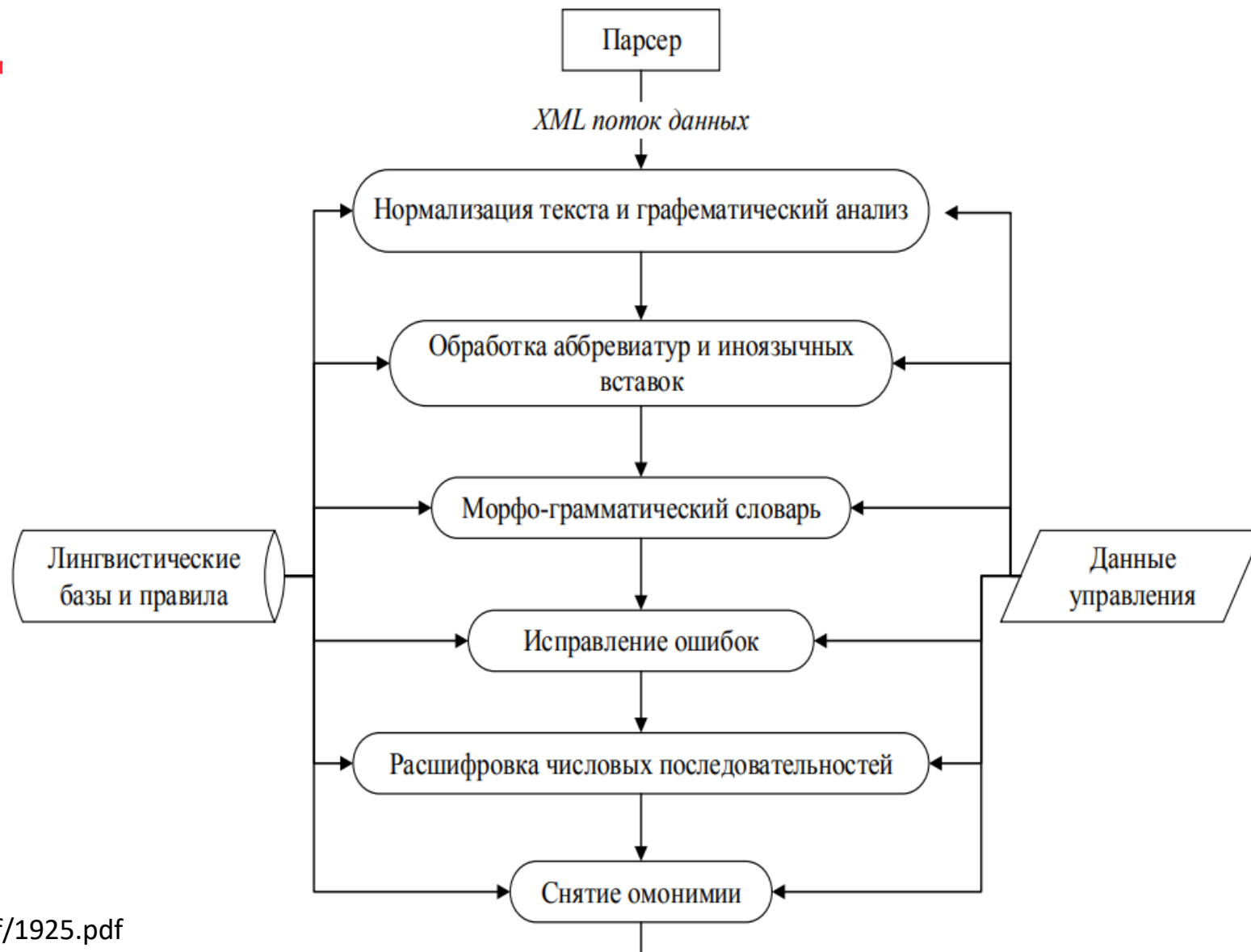
# Проблемы фонемизации текста

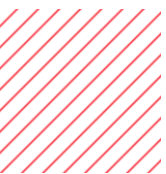
	Орфографическая запись	Фонетическая транскрипция
1) между буквой и звуком нет однозначного соответствия	хоккей	[хакэй']
2) проблема расстановки ударений	удáлся	[удáлс'а]
3) проблема снятия омографии	дорóга / дорогá зáмок / замóк	[дарóга] / [дарагá] [зáмак] / [замóк]

графема – минимальная единица письменности

фонема – минимальная смысловозначительная единица языка,  
реализуется в виде конкретных звуков (аллофонов)

# Лингвистический процессор





# Спасибо за внимание!

