академия больших данных

mail.ru group

# Text to Speech

Калиновский И.А., к.т.н.
Just AI

# План курса

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


**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 |

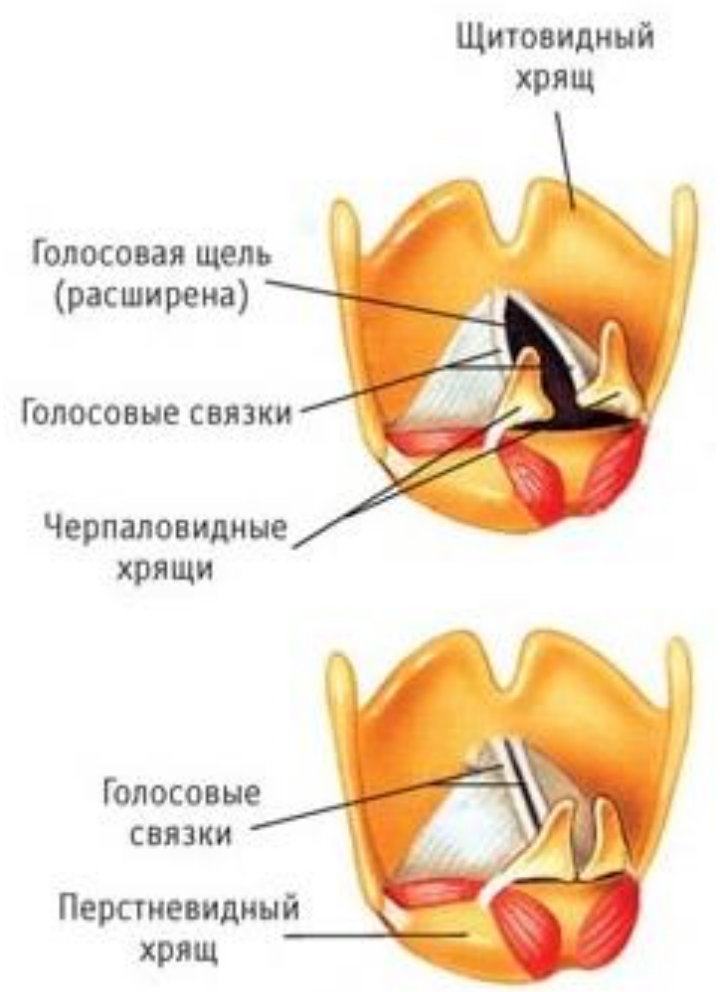
### 45%
в машинах

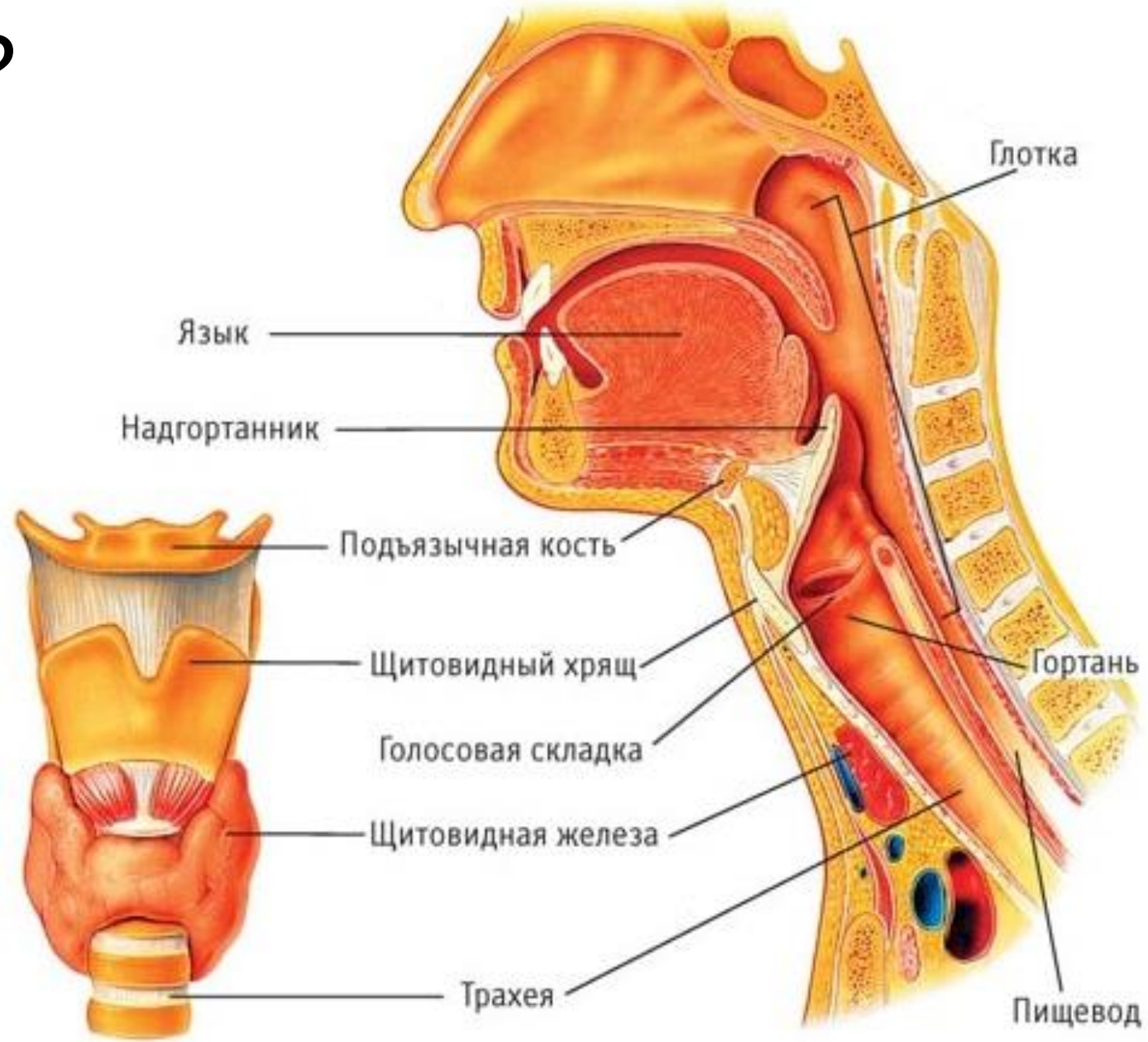
| | |
|---|---|
| 73,7% | звонок по телефону |
| 50,3% | маршруты и навигация |
| 41,2% | отправка смс |
| 28,7% | музыкальные сервисы |

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

https://techfusion.ru/just-ai-v-2020-godu-rossijskij-rynok-umnyh-kolonok-vyrastet-v-4-raza/

# Как мы говорим?




https://ladyle.ru/images/ladyle/2020/05/raspologenie-organov-cheloveka-v-kartinkah-gortan.jpg

# 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

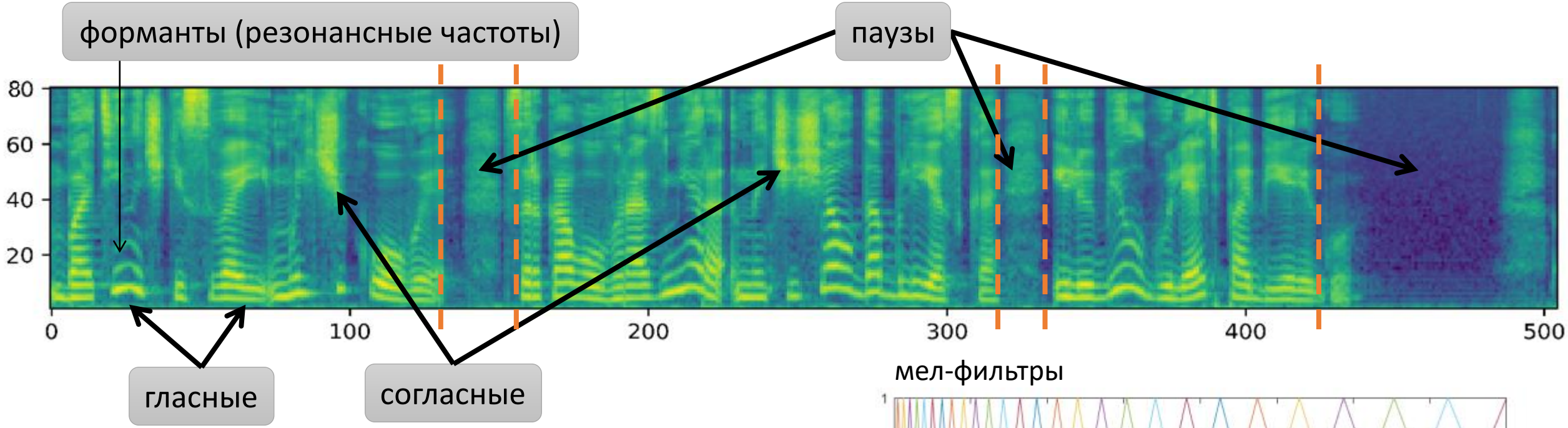
**Благозвучность голоса** — чистота его звучания, отсутствие неприятных призвуков (хрипоты, сиплости, гнусавости)

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

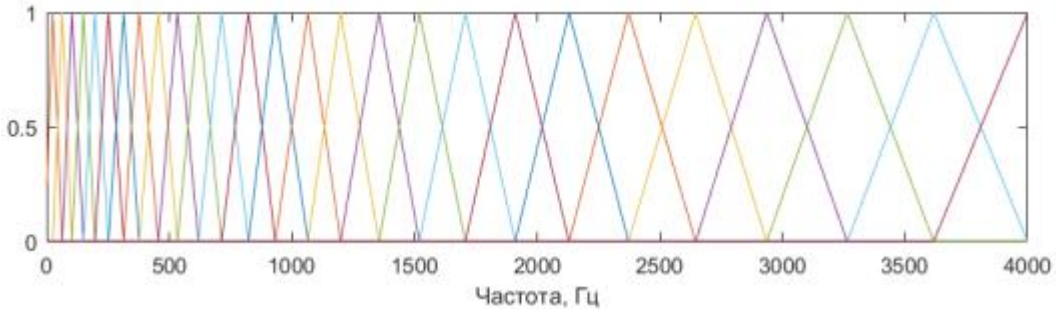
**Интонация** — это ритмомелодический строй речи. Интонация включает: высоту тона, силу звучания, мелодику, темп, ударения и паузы.

# Спектр речевого сигнала

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




мел-фильтры



# Механические синтезаторы

Гласный орган Х. Кратценштейна (1780 г.)

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

https://www.vocaltractlab.de/publications/hoffmann-2018-specom.pdf

# Механические синтезаторы

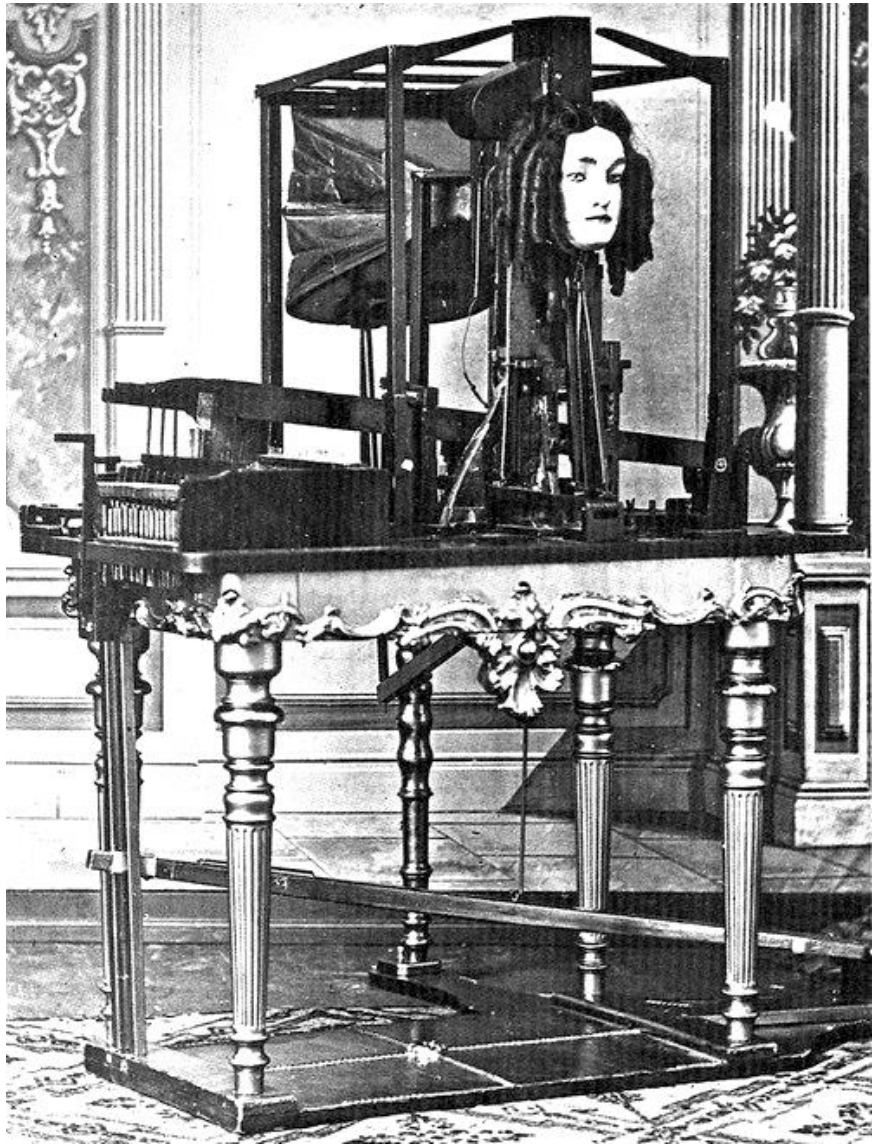
Машина фон Кемпелена  (1788 г.)

https://www.youtube.com/watch?v=oIjkzZGe2I8

# 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.

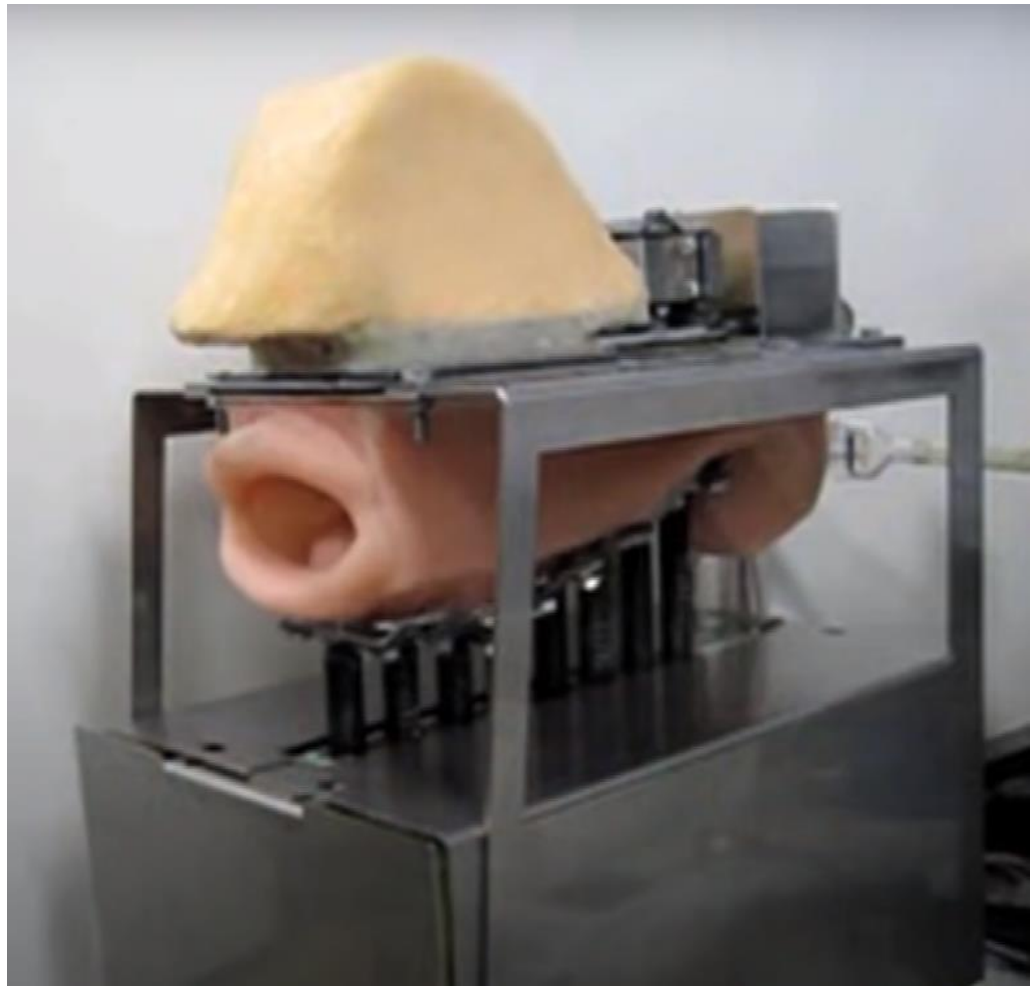


https://habr.com/ru/company/audiomania/blog/490596/

# Механические синтезаторы

Современная вариация, Япония, 2011

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



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

# Электрические синтезаторы

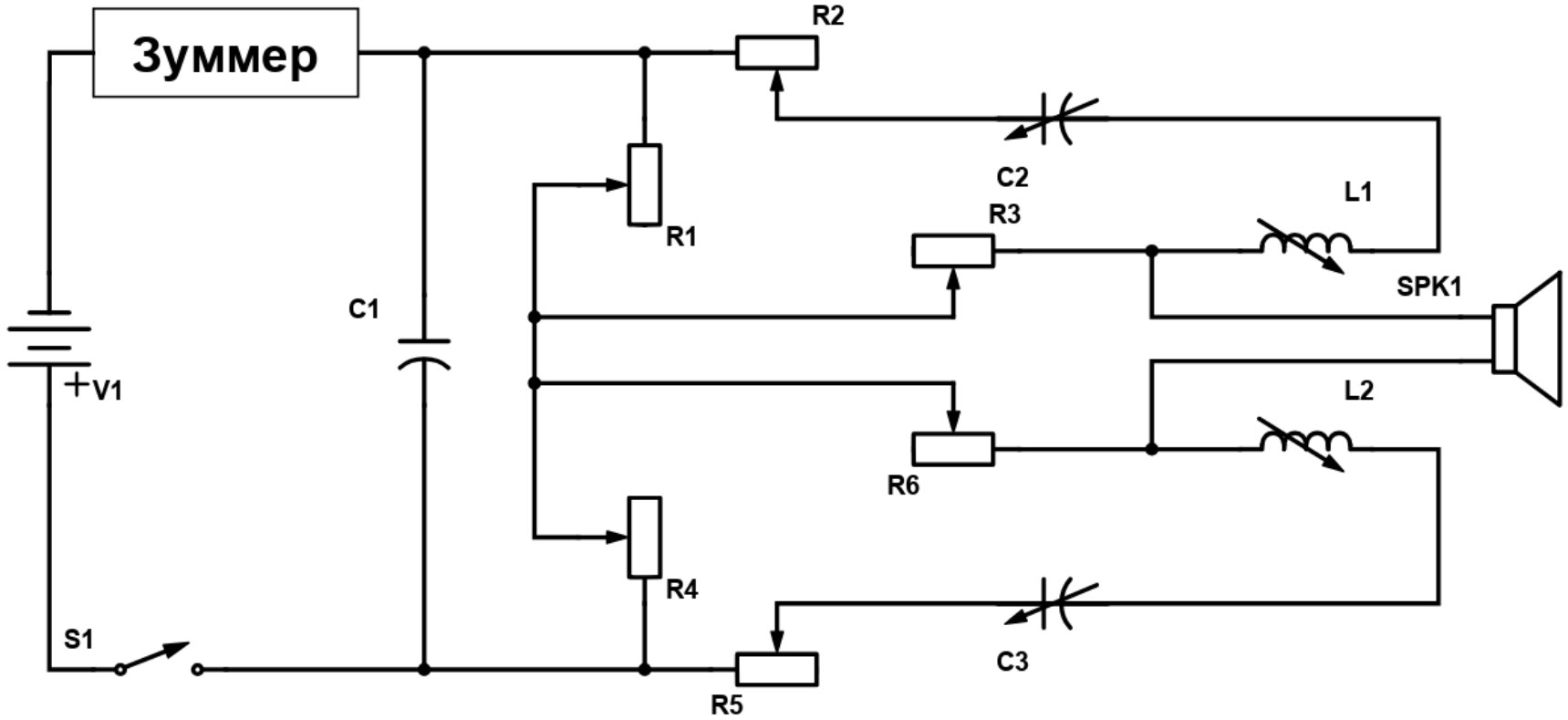
«Большой аппарат Гельмгольца для соединения тембров из 10 гармоник» или синтезатор Фурье –

первый прообраз современного синтезатора

https://habr.com/ru/company/audiomania/blog/491746/

# Электрические синтезаторы

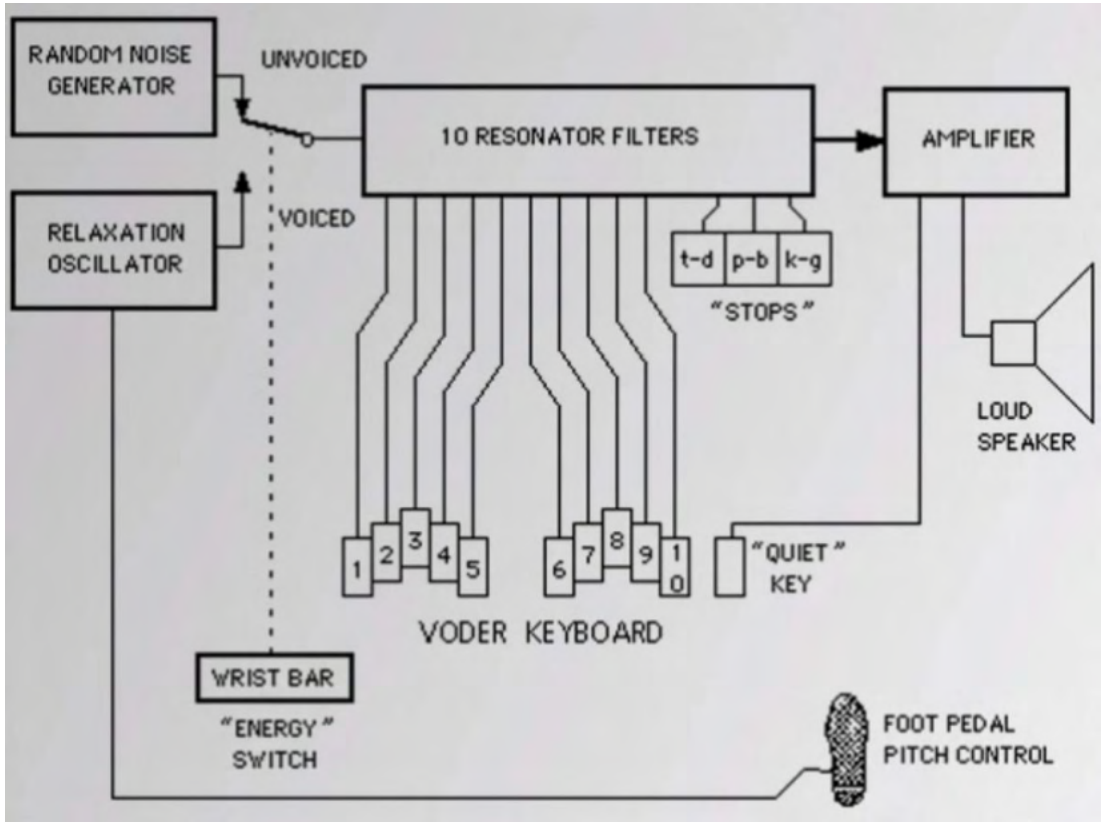
Синтезатор Д. Стюарта




https://habr.com/ru/company/audiomania/blog/491746/

# Электрические синтезаторы

Водер Г. Дадли, 1939

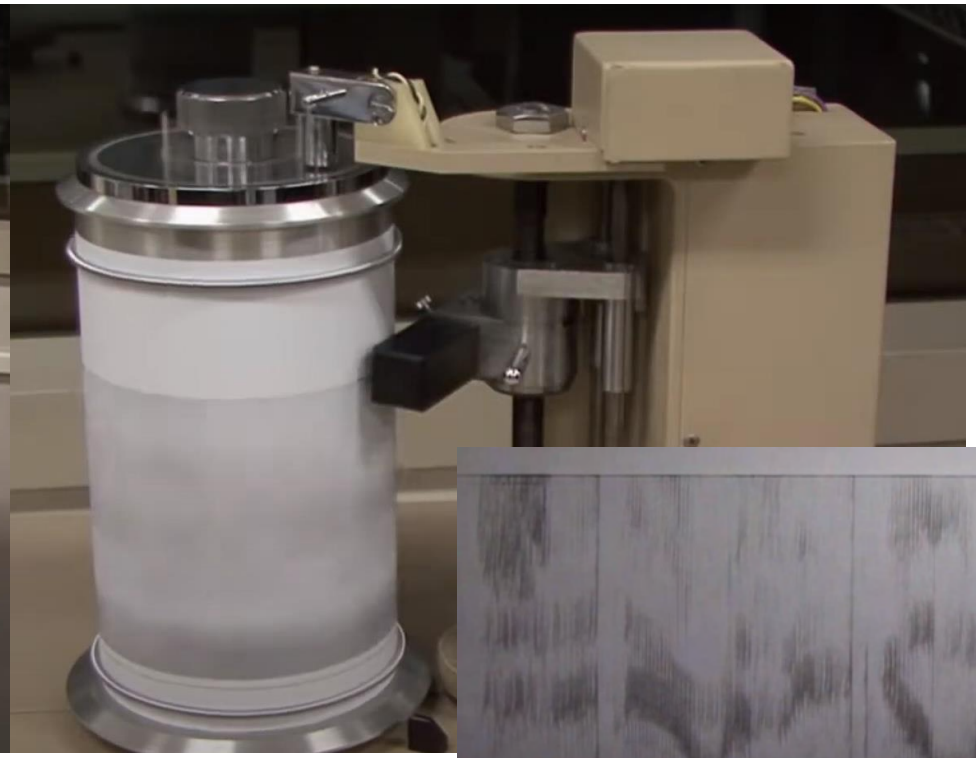




https://www.youtube.com/watch?v=5hyI_dM5cGo

# Электрические синтезаторы

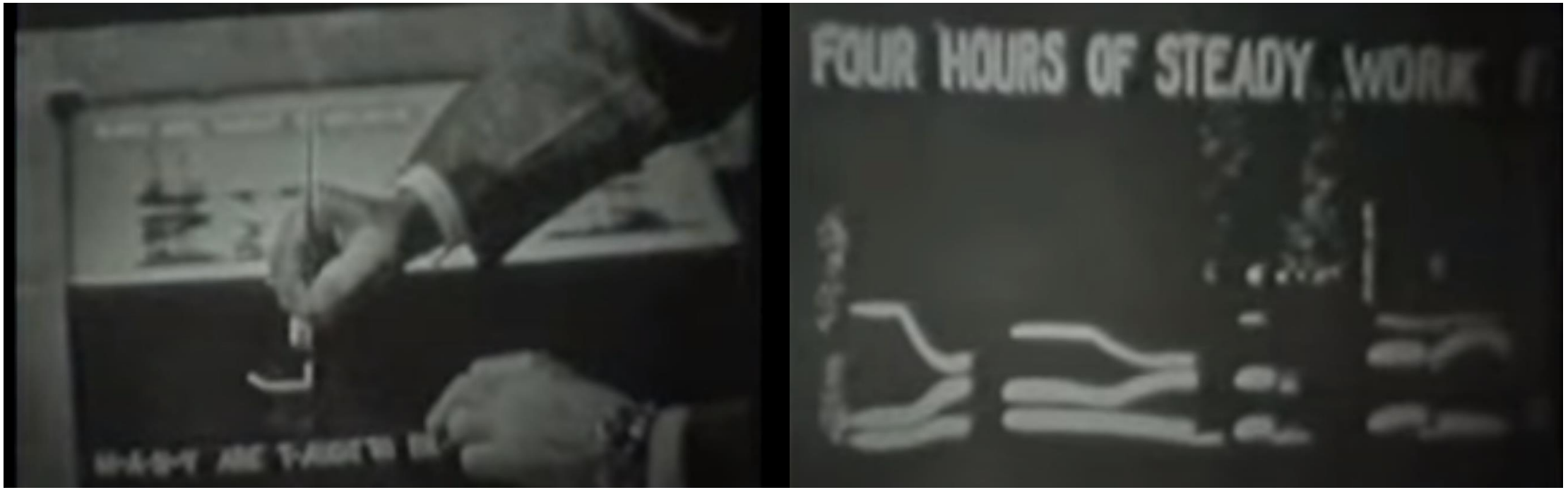
Акустический спектрограф



https://www.youtube.com/watch?v=6LzSIQbSX1U

# Электрические синтезаторы

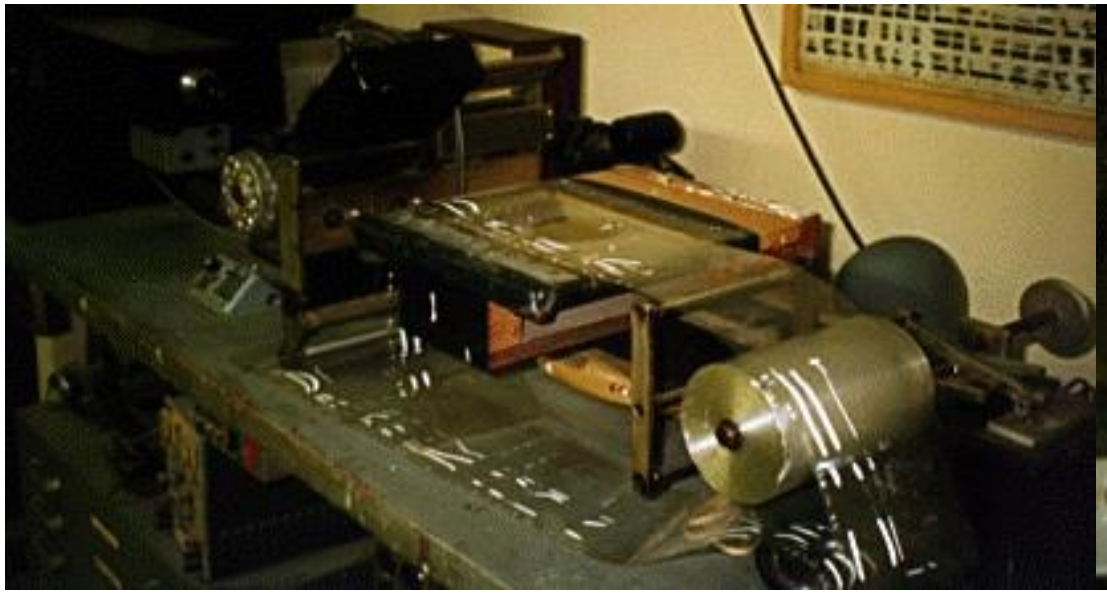
Синтезатор Л. Шотта (Bell Labs), 1946

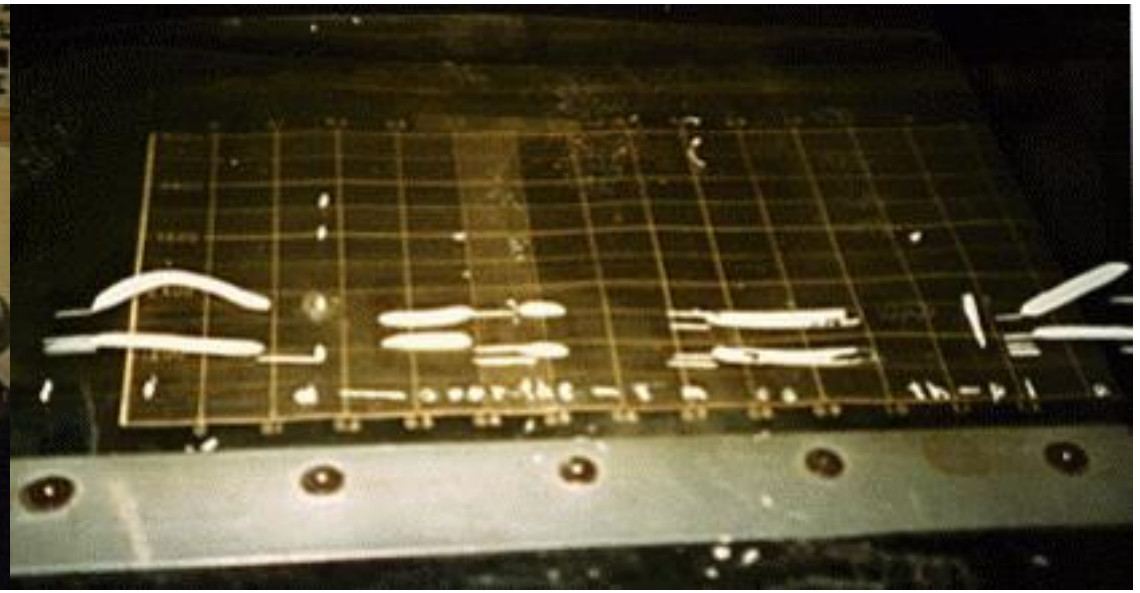



https://www.youtube.com/watch?v=TZh6ZYYqLJc

# Электрические синтезаторы

Pattern Playback Ф. Купера, 1949

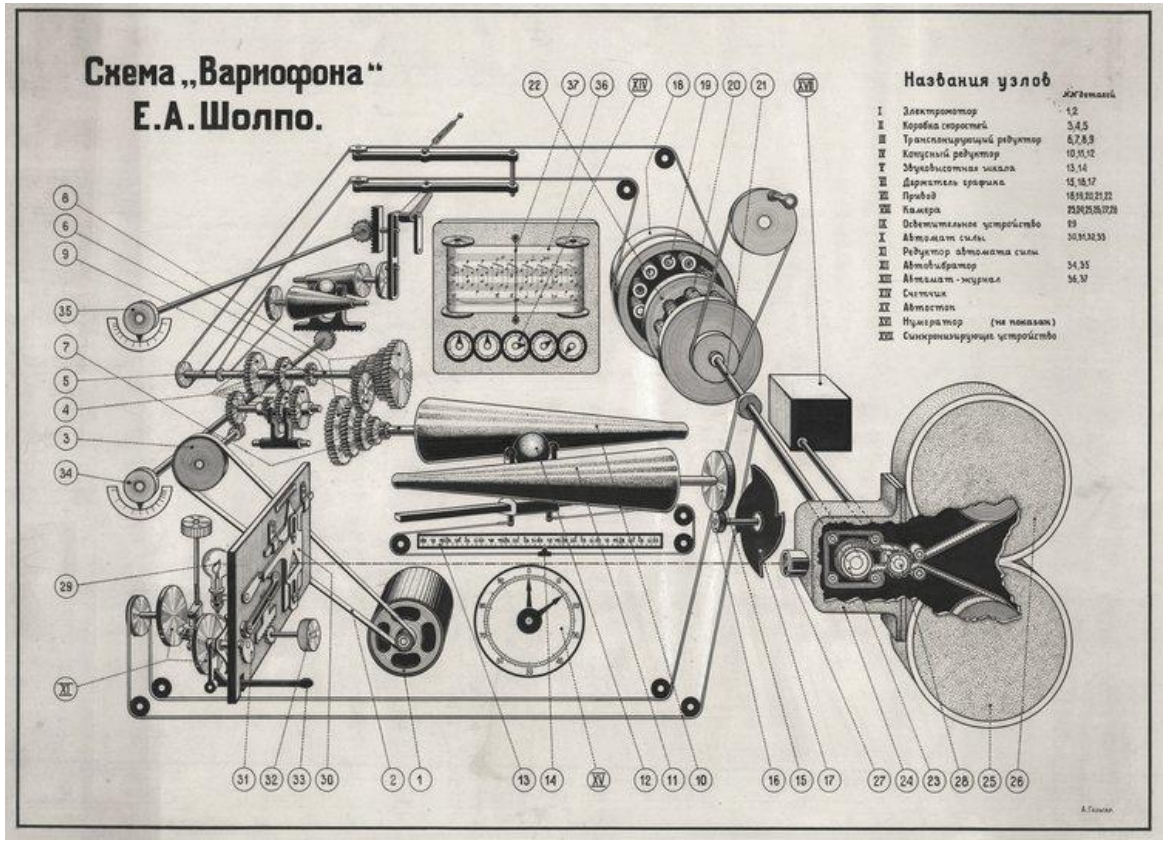




https://www.youtube.com/watch?v=TZh6ZYYqLJc

# Электрические синтезаторы

Вариофон Е. Шолпо, 1942

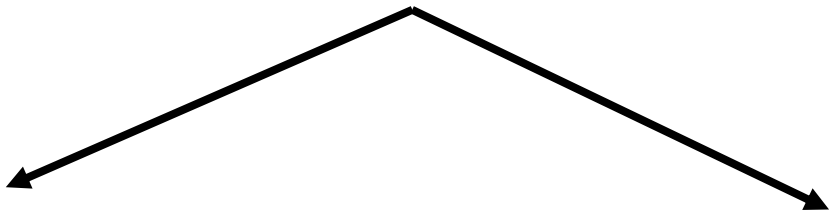



https://www.digitalmusicacademy.ru/lesson-variofon

# Синтезаторы I поколения

Артикуляционные

Акустические

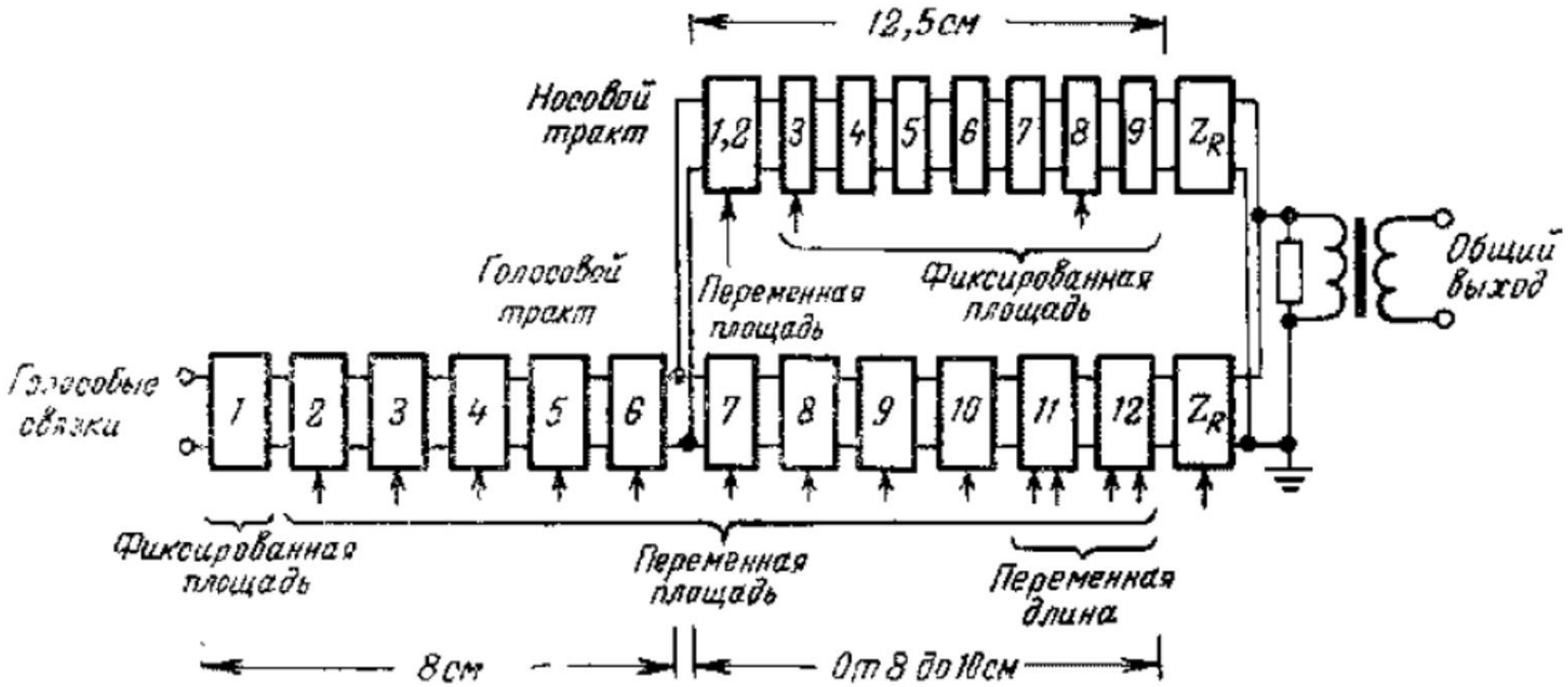


Формантный синтез

LPC-синтез

# Артикуляционные синтезаторы

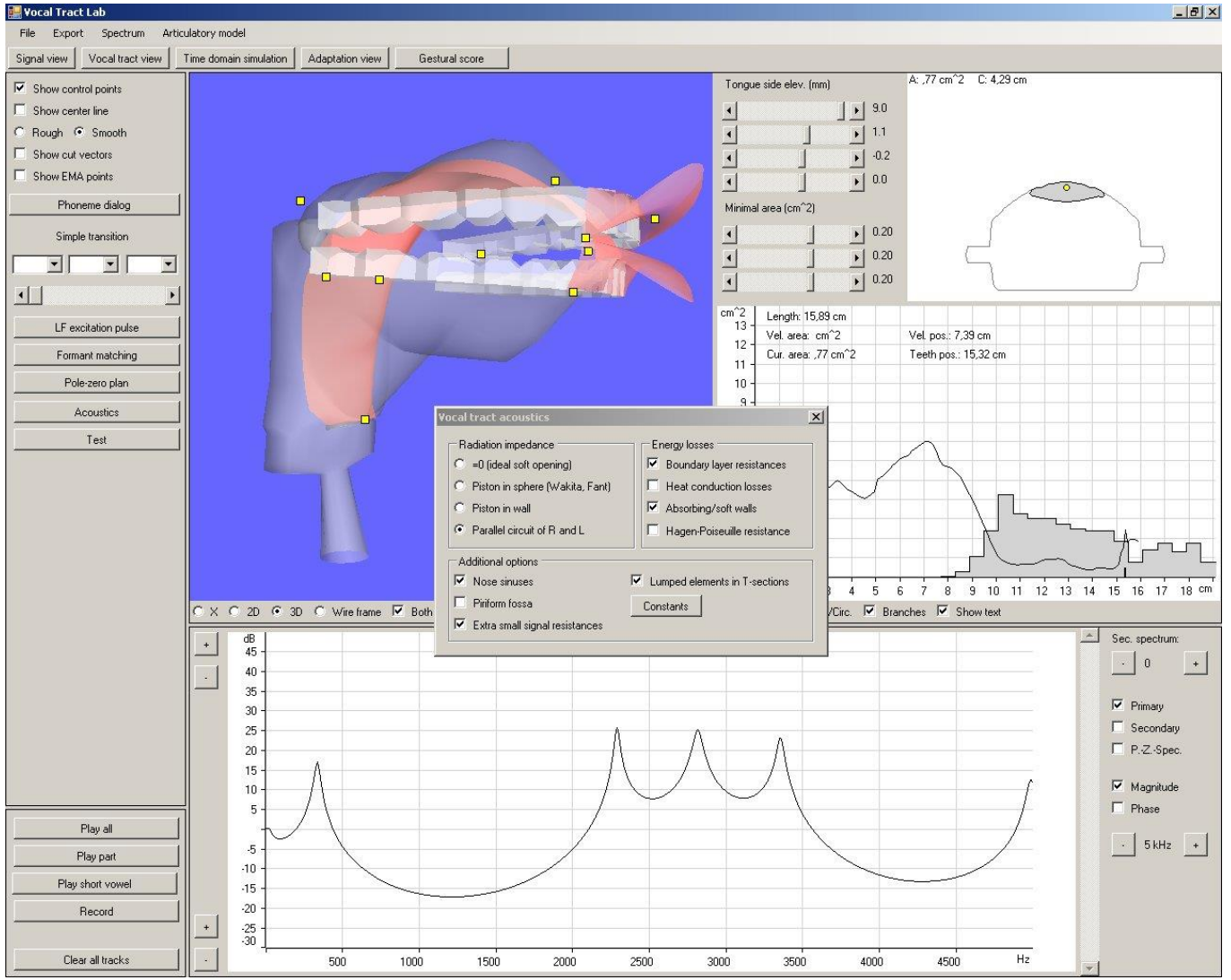
DAVO (Dynamic Analog of the VOcal tract) Д. Розен, 1958




https://books.ifmo.ru/file/pdf/1925.pdf

# Артикуляционные синтезаторы

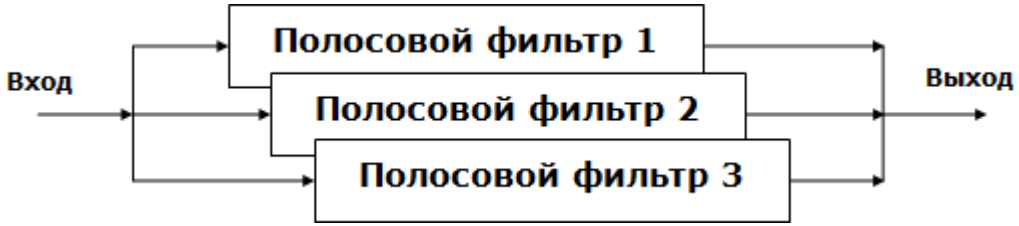
VocalTractLab 2.3, 2020




https://www.vocaltractlab.de/index.php?page=start

# Формантный синтез

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

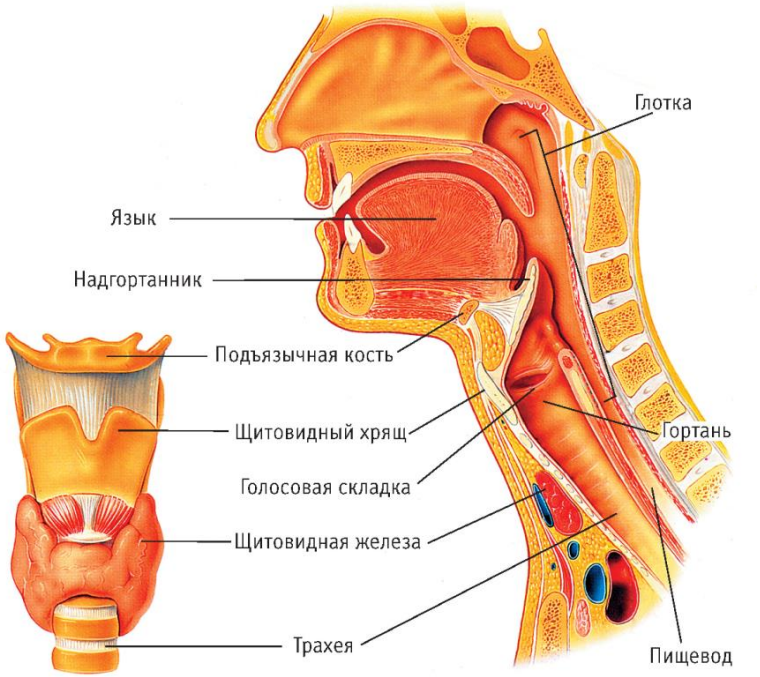




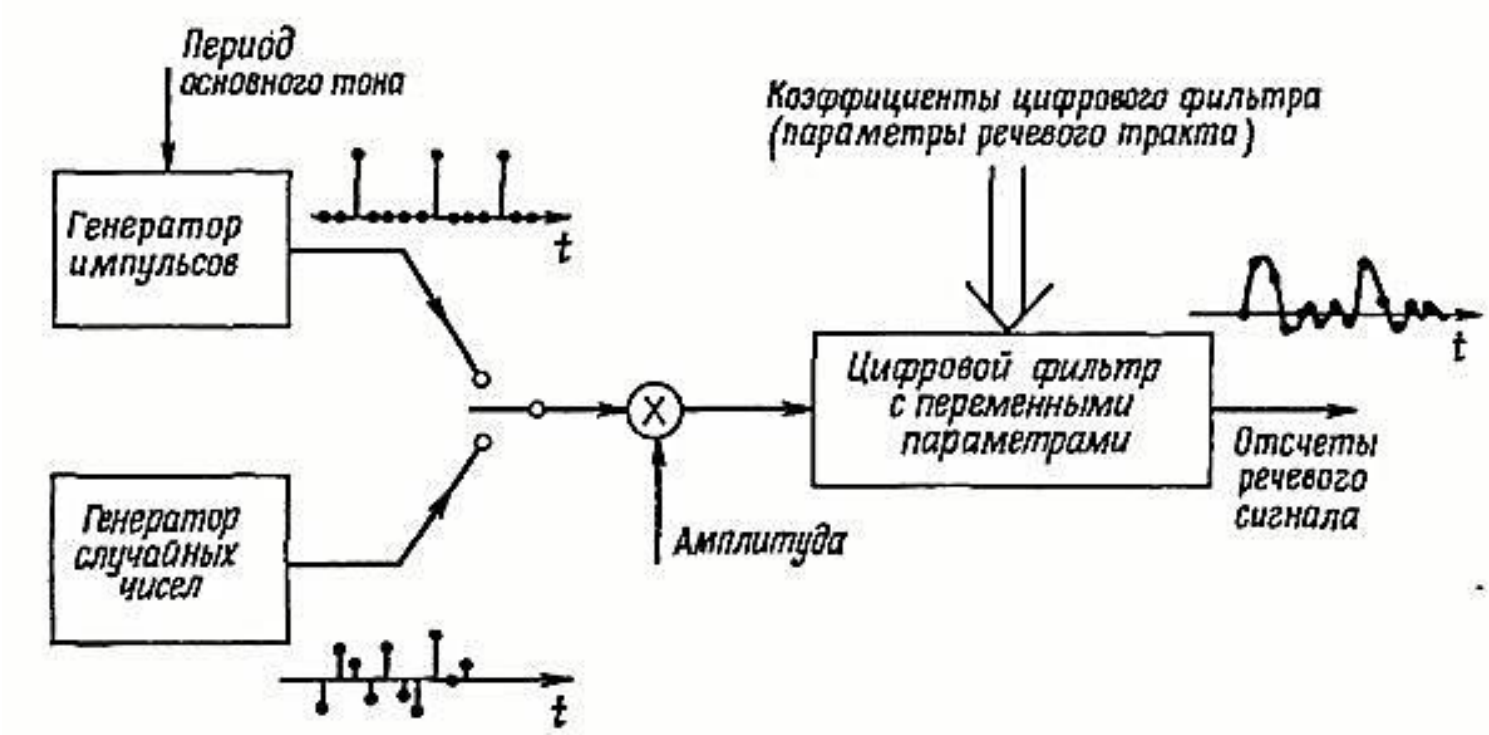
https://ssrlab.by/wp-content/uploads/2003/12/2003_BMLabanau_K_istorii_russkogovor_mashin.pdf

# 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*

https://lcav.gitbook.io/dsp-labs/linear-prediction/theory

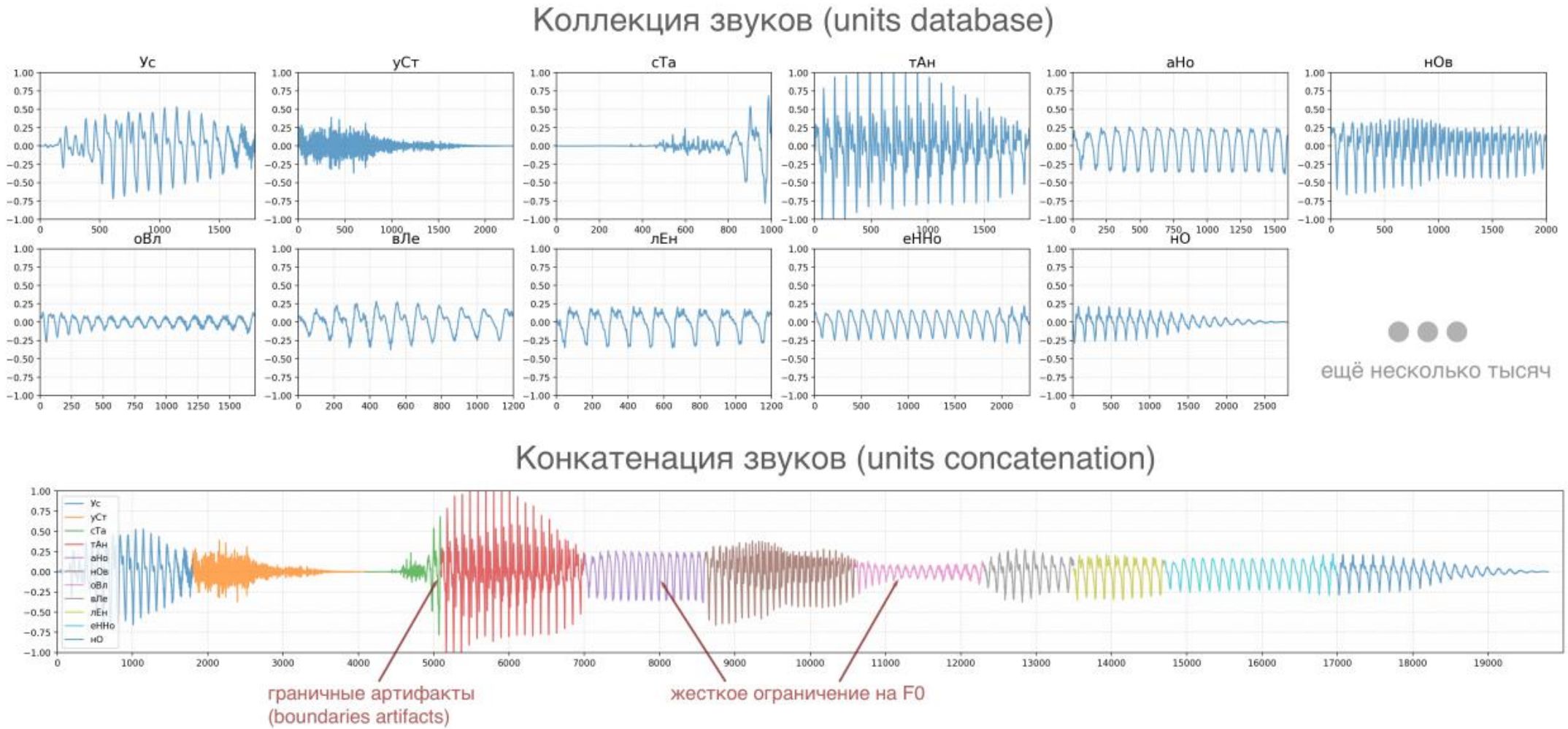
# Синтезаторы II поколения

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

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

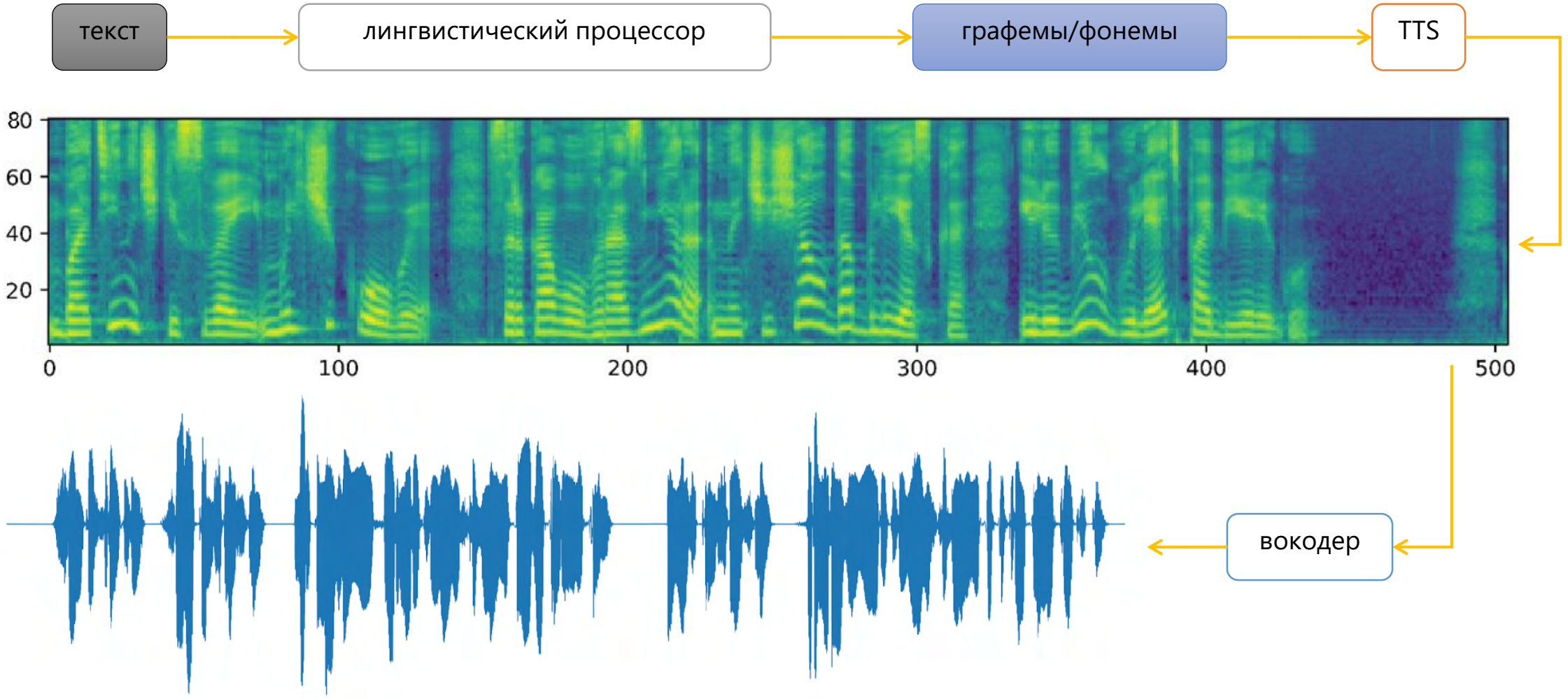
https://acousticstoday.org/klatts-speech-synthesis-d/

# Синтезаторы III поколения




https://www.youtube.com/watch?v=6LzSIQbSX1U

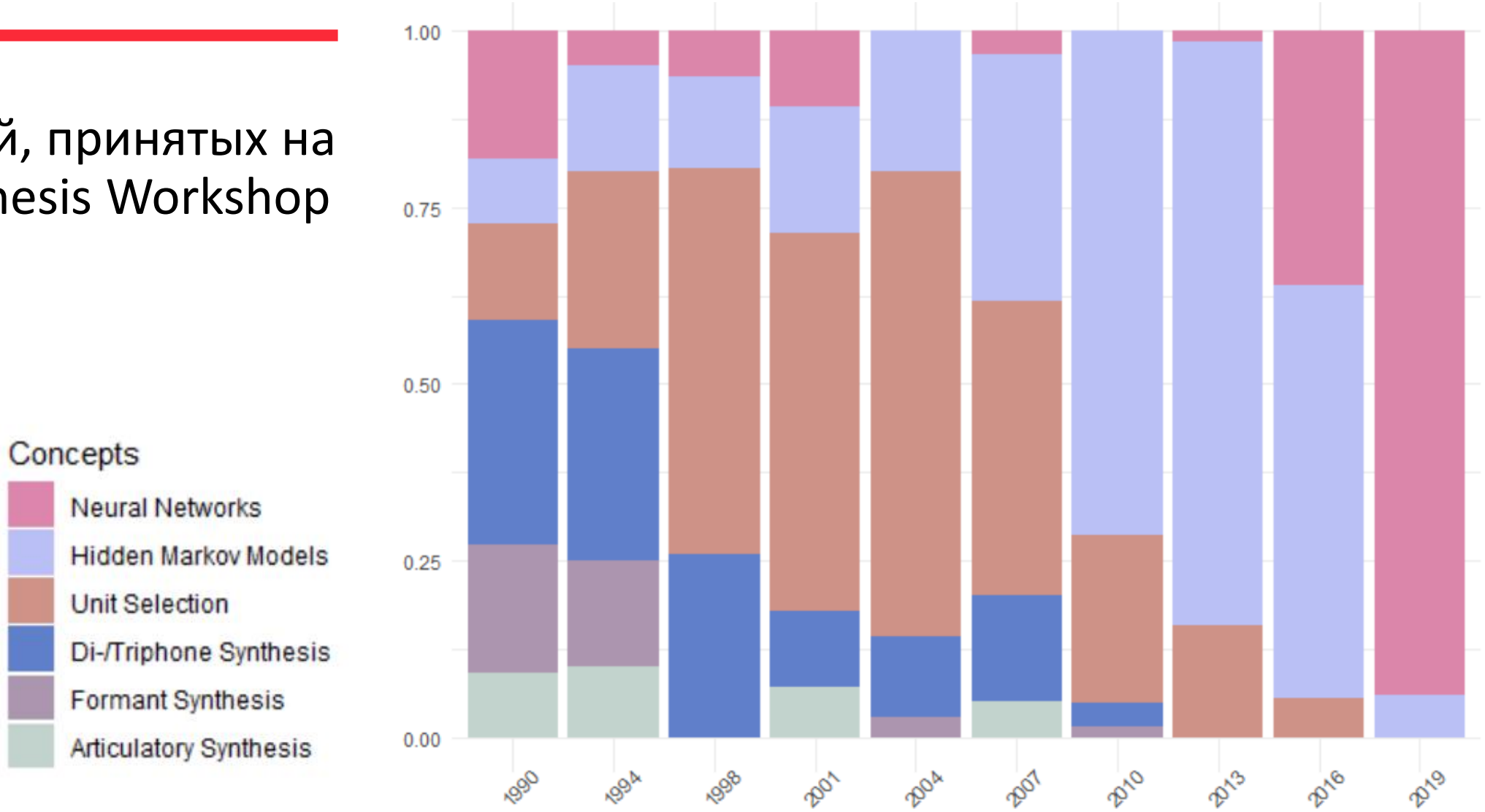
# Синтезаторы IV поколения




https://www.youtube.com/watch?v=6LzSIQbSX1U

# Технологии синтеза речи

Обзор статей, принятых на Speech Synthesis Workshop



# 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.*

| | | |
|---|---|---|
| **синтагмы** | *сегодня светит солнце* | *но дует ветер* |
| **слова** | *сегодня* \| *светит* \| *солнце* | *но* \| *дует* \| *ветер* |
| **слоги** | *се-го-дня* \| *све-тит* \| *сол-нце* | *но* \| *ду-ет* \| *ве-тер* |
| **фонемы** | s‘/$i_1$-v/$o_0$-d‘/n‘/$a_4$ \| s/v'/$e_0$-t'/$i_4$/t \| s/$o_0$-n/c/$y_4$ | *n*/$o_1$ \| *d*/$u_0$-*j*/$i_4$/*t* \| *v*'/$e_0$-*t*'/$i_4$/*r* |

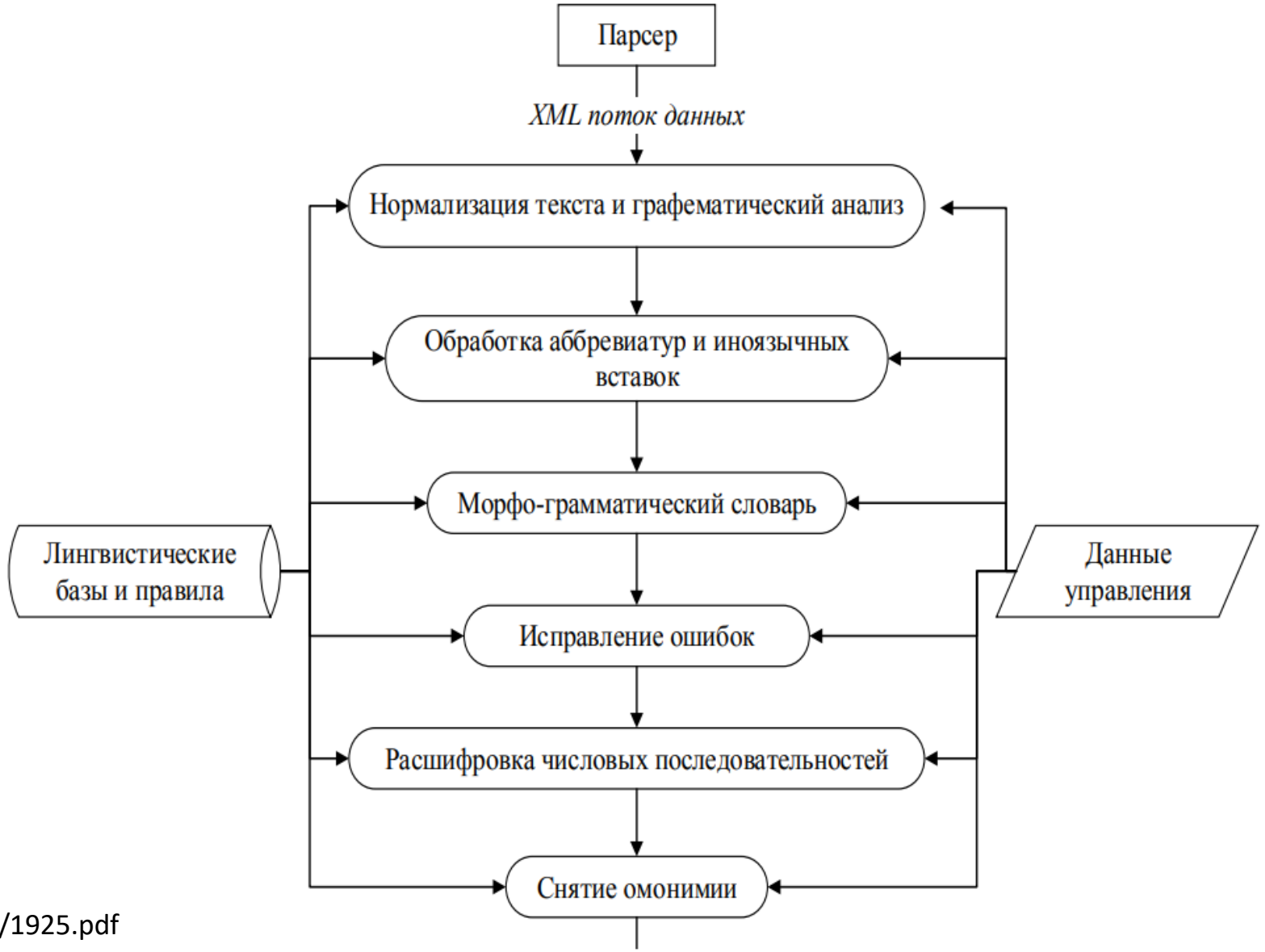
# Проблемы фонемизации текста

| | *Орфографическая запись* | *Фонетическая транскрипция* |
|---|---|---|
| 1) между буковой и звуком нет однозначного соответствия | хоккей | [хакэ́й’] |
| 2) проблема расстановки ударений | уда́лся | [уда́лс’а] |
| 3) проблема снятия омографии | доро́га / дорога́<br>за́мок / замо́к | [даро́га] / [дарага́]<br>[за́мак] / [замо́к] |

графема – минимальная единица письменности

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

# Лингвистический процессор




https://books.ifmo.ru/file/pdf/1925.pdf

# Спасибо за внимание!