Введение в ЦОС и распознавание речи

14.11.2024

Не забудьте отметиться и оставить отзыв!

О чем поговорим (в рамках 2ух занятий)

- Введение в ЦОС, как представлен звук в памяти компьютера
- Распознавание речи ASR (какую задачу решаем, как оцениваем качество решения)
- Архитектуры моделей для решения задачи ASR (рассмотрим основные подходы, подробнее поговорим про SOTA решения)
- SSL в задаче ASR
- Применение LM для улучшения качества распознавания

Чего не будет на занятиях:

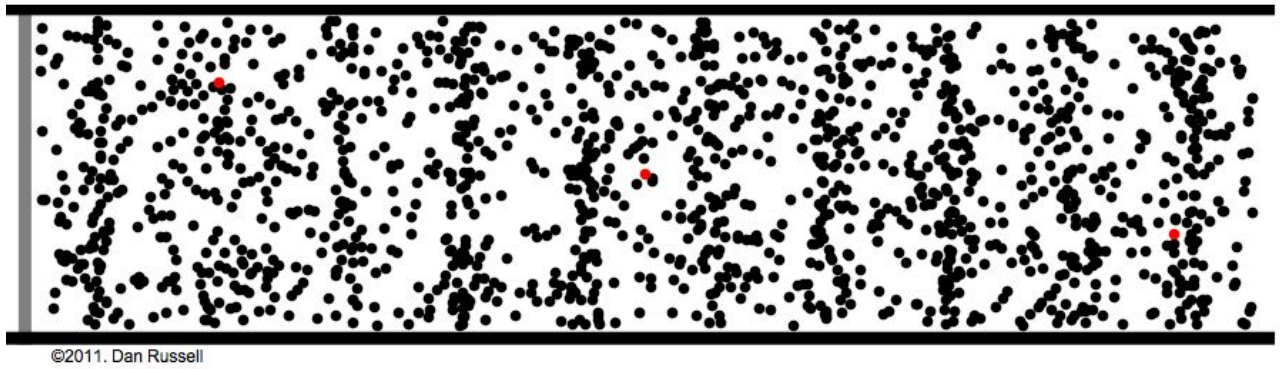
- TTS (text-to-speech)
- KWS (keyword spotting)
- VQE (voice quality enhancement)

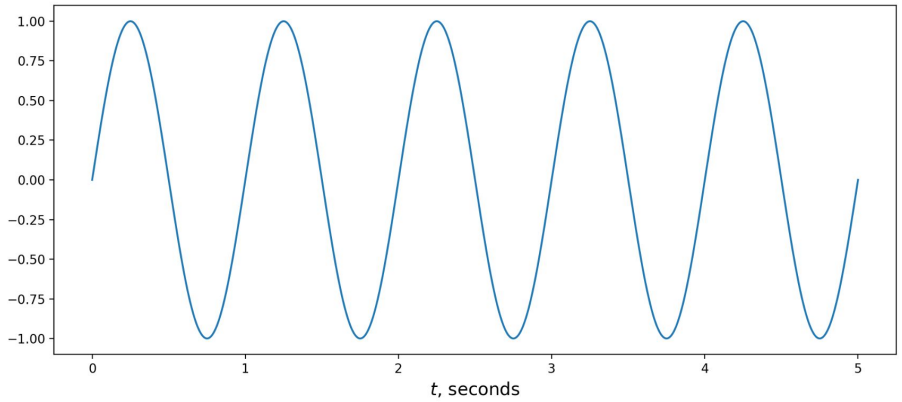
План занятия

- Представление звукового сигнала в памяти компьютера
- Введение в ЦОС
- Постановка задачи распознавания речи (ASR)
- Метрики оценки качества
- Подходы к решению задачи ASR
- CTC декодер
- Архитектуры энкодеров

Что такое звук?

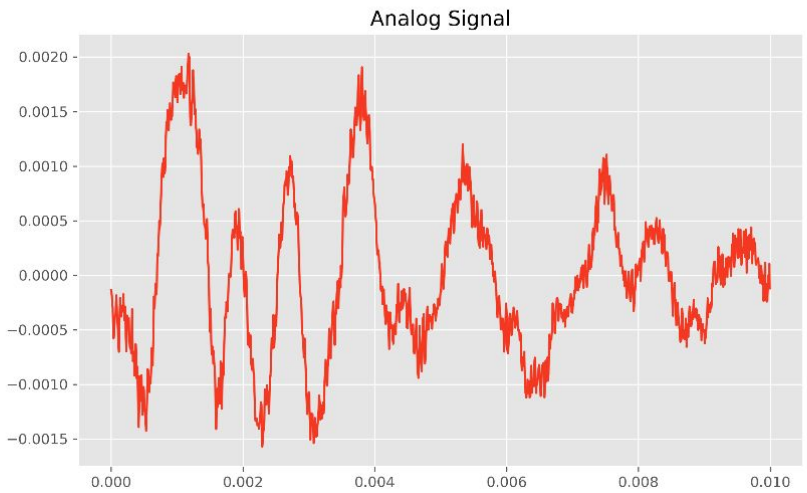
Звук (вспомним физику)



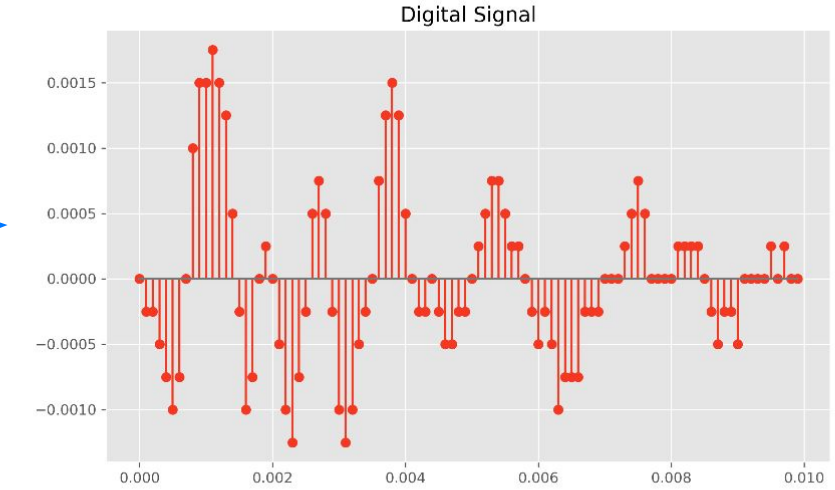


Звуковая волна (звуковые колебания) — это передающиеся в пространстве механические колебания молекул вещества (например, воздуха).

Дискретизация и квантование сигнала



АЦП



массив значений
(многомерный вектор)
сложно анализировать

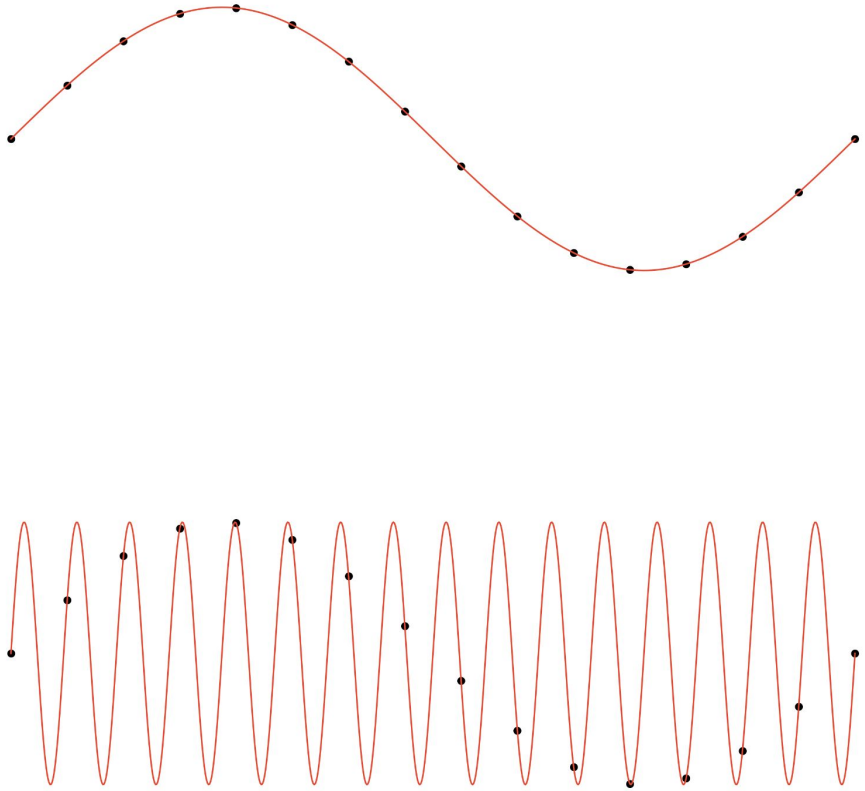
Преобразование Фурье

$$u(x) = \int_{-\infty}^{+\infty} U(f) e^{2\pi i f x}, \qquad U(f) = \int_{-\infty}^{+\infty} u(x) e^{-2\pi i f x} dx$$

обратное (f-t) прямое (t-f)

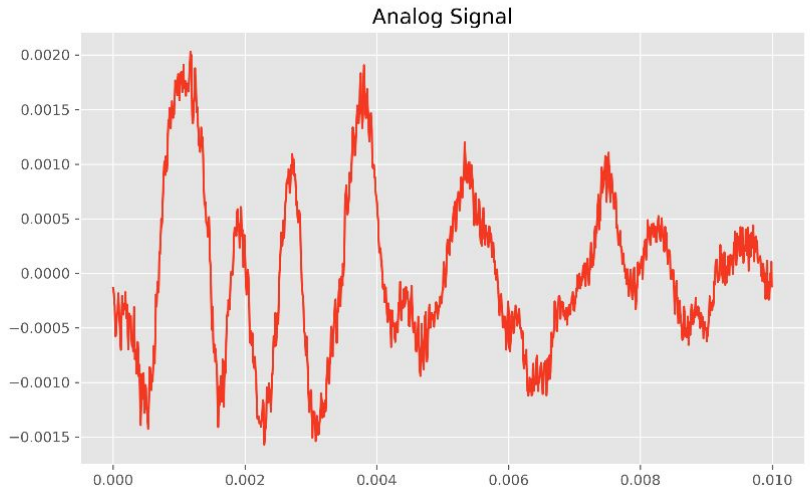
Основная идея: разложить сигнал на базисные функции для дальнейшего анализа спектра сигнала.
Подробнее тут и тут.

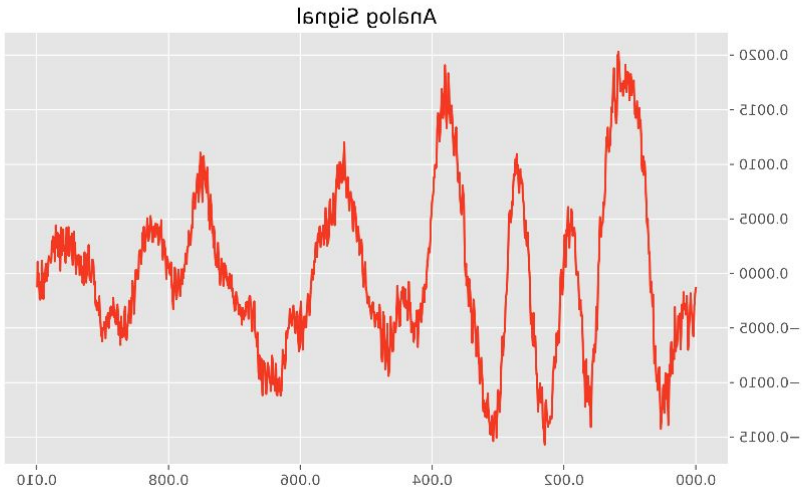
Теорема Котельникова (Найквиста-Шеннона)



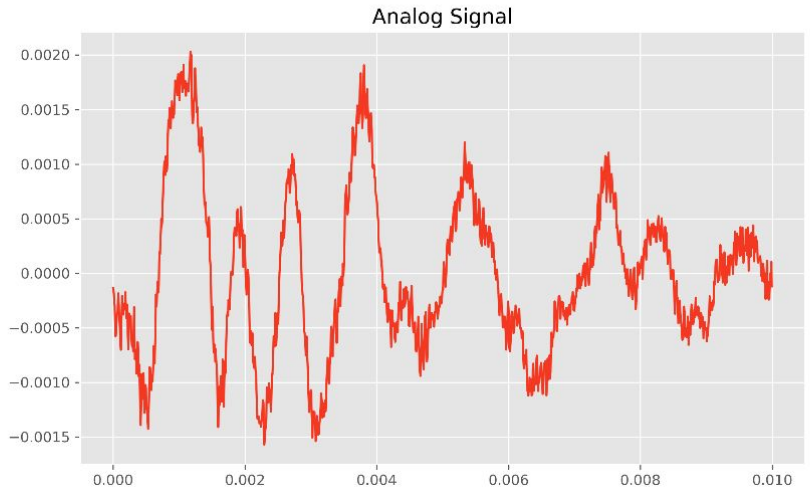
Любой аналоговый сигнал может быть восстановлен с какой угодно точностью по своим дискретным отсчётам, взятым с частотой F > 2Fc (Fc - частота семплирования)

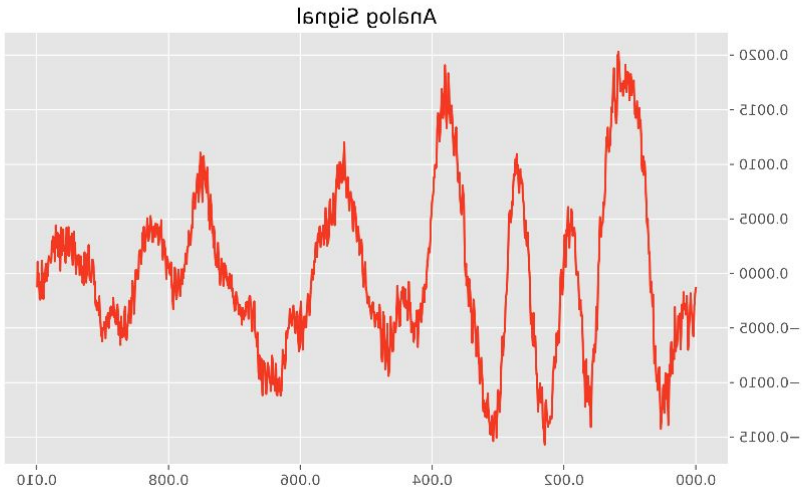
Достаточно ли просто знать спектр сигнала?





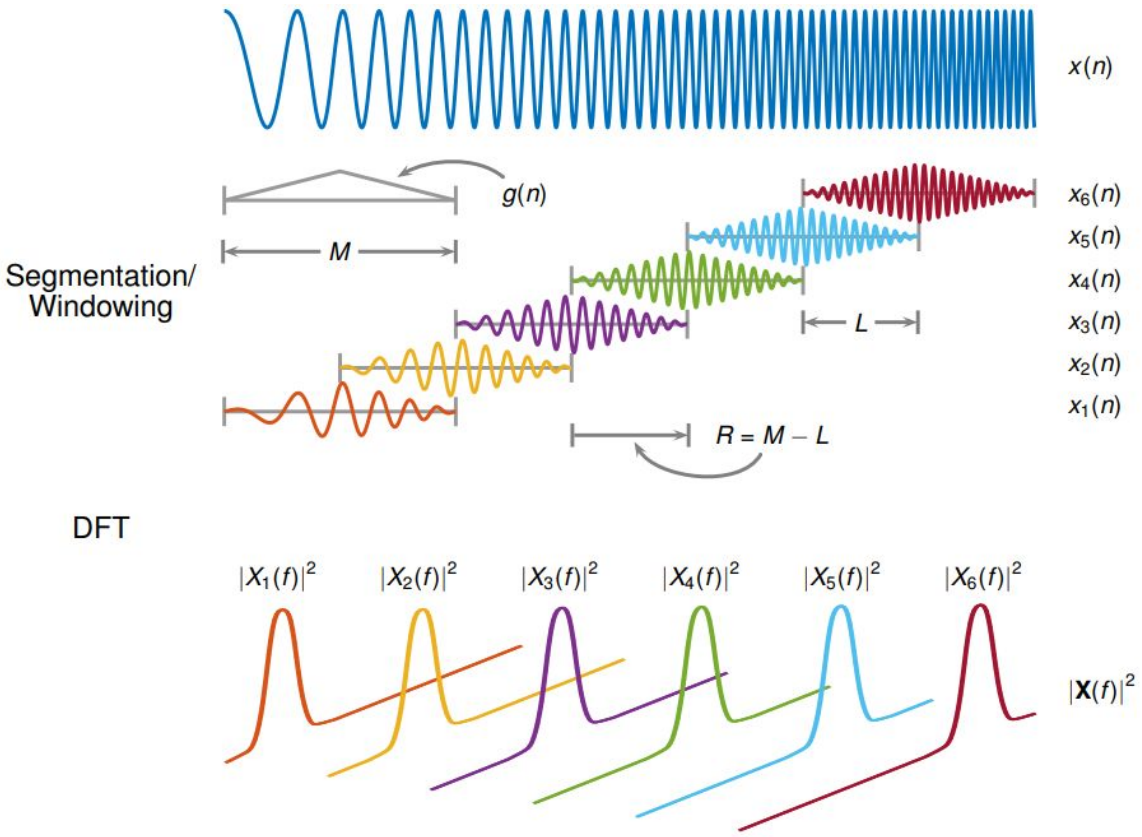
Достаточно ли просто знать спектр сигнала?

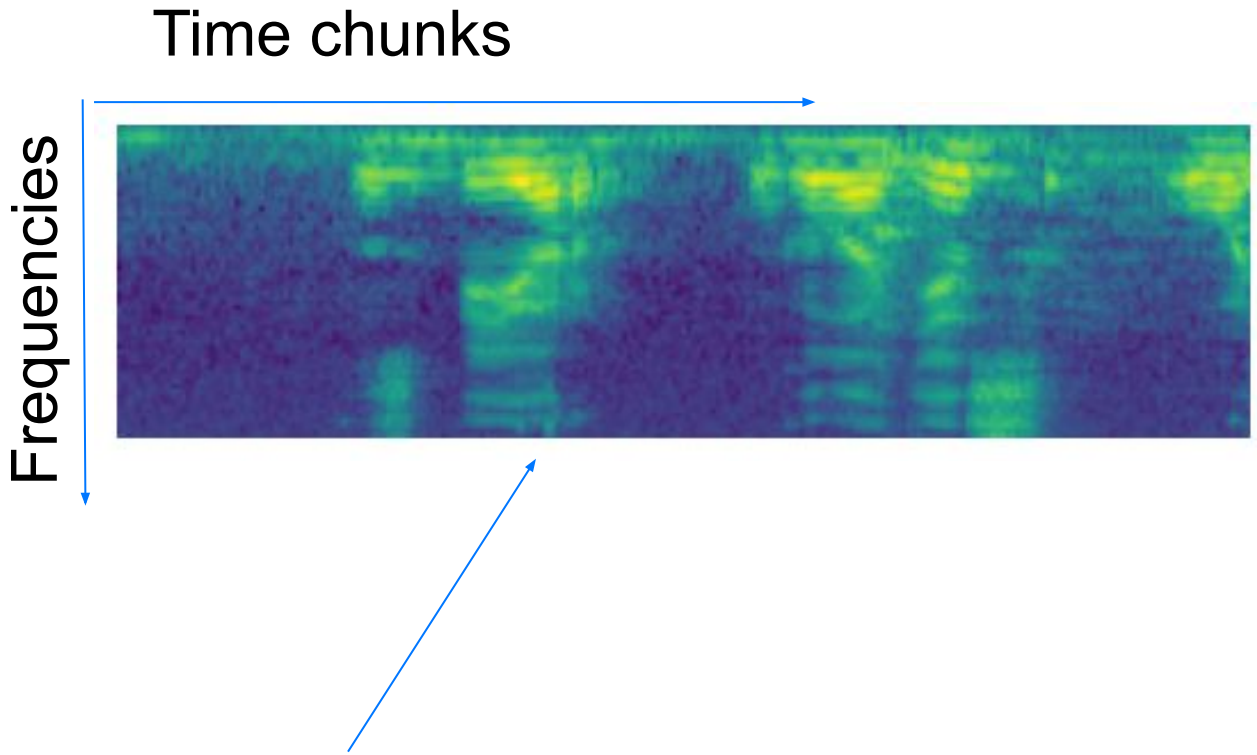




Спектр у этих двух сигналов - одинаковый!
Как учесть временную составляющую данных?

Спектрограмма (STFT)



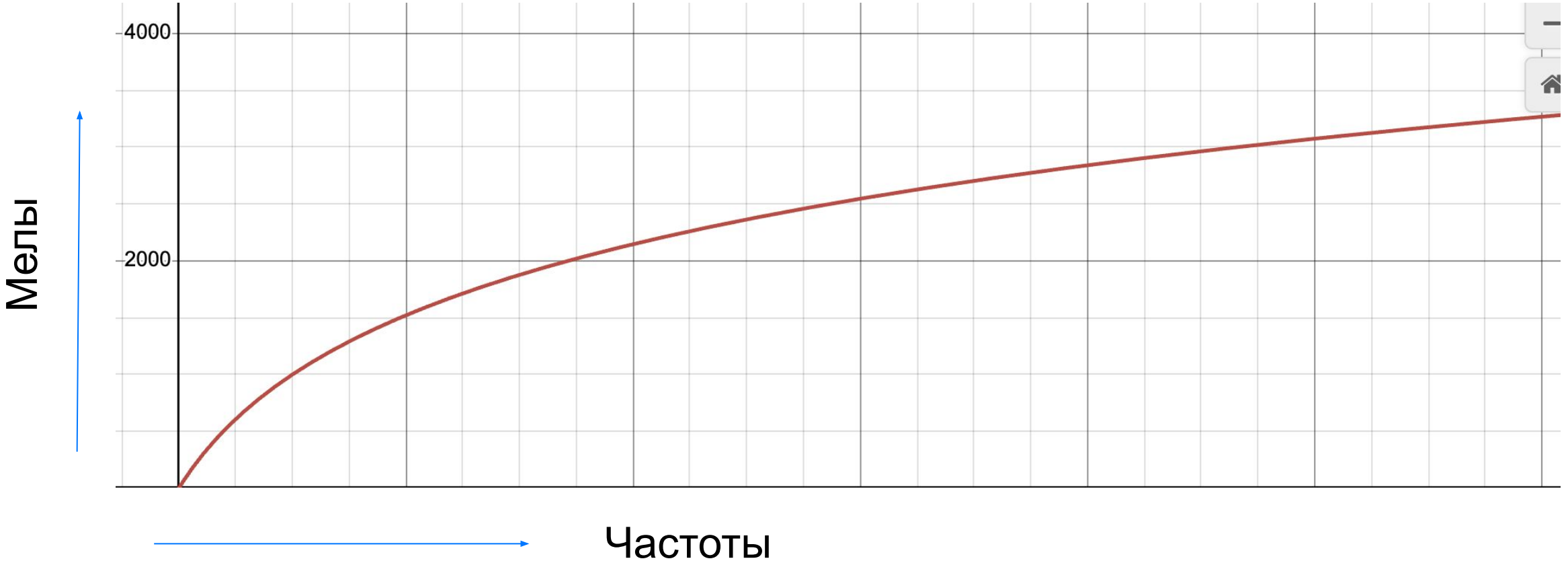


Энергия сигнала(f, T)

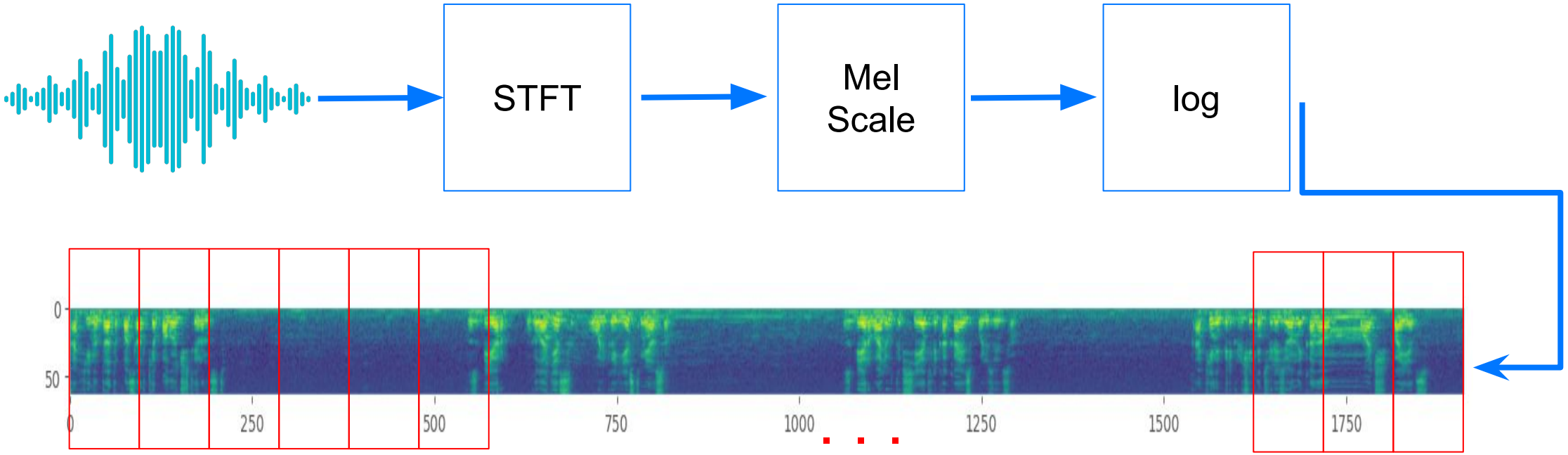
Спектрограмма (STFT). Мел шкала.

$$m[mel] = 1127 \ln(1 + \frac{f[Hz]}{700})$$

Шкала частот (имитирует чувствительность человеческого уха)в



Мел спектрограмма



Чанки спектрограммы (последовательность)

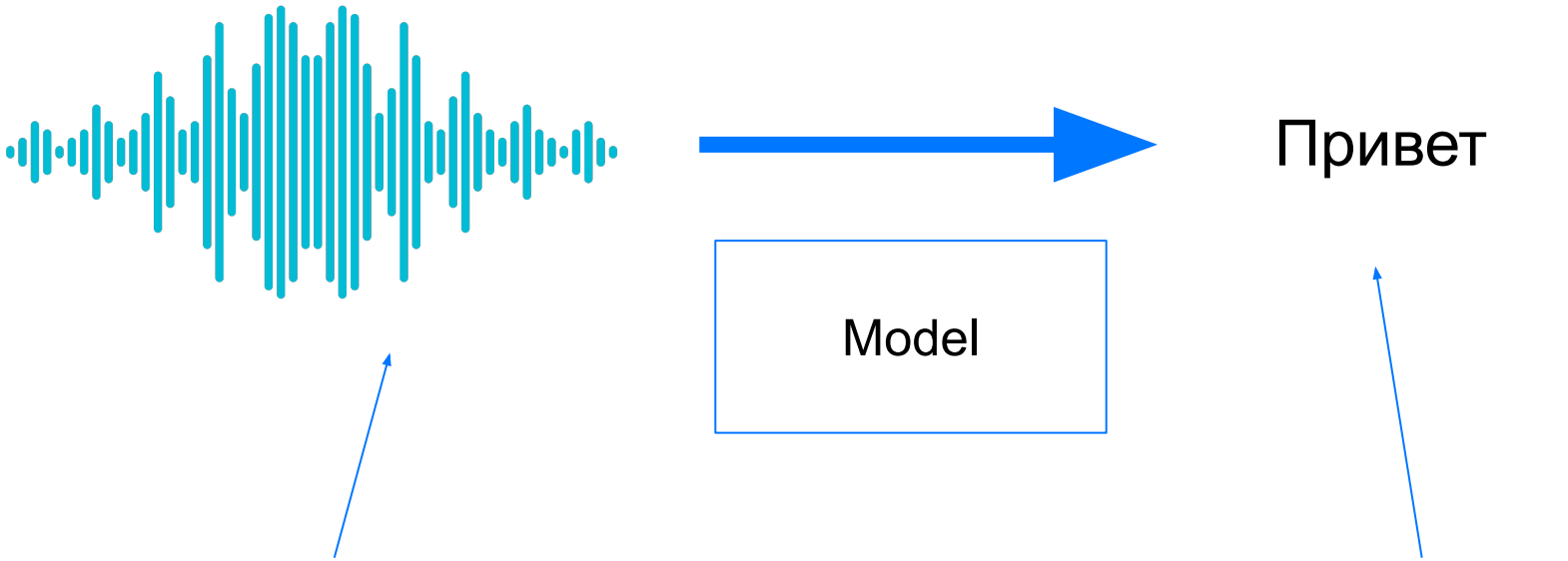
Дополнительно по теме:

- Room Impulse Response (RIR)
- Convolution theorem
- Bandpass filters

Постановка задачи ASR

Метрики оценки качества распознавания

Постановка задачи ASR



Как оценить качество “перевода”?

Метрики оценки качества транскрибации

- Word Error Rate (**WER**)
  - **S** – substitutions
  - **I** – insertions
  - **D** – deletions
  - **N** – reference words

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

```
ref. = "по дороге домой услышал скрип"
hypo = "а по дороге услышал стук"

| I | C  | C       | D     | C        | S     |
| % | по | дороге | домой | услышал | скрип |
| а | по | дороге | #     | услышал | стук  |
```

Метрики оценки качества транскрибации

- Word Error Rate (**WER**)
  - **S** – substitutions
  - **I** – insertions
  - **D** – deletions
  - **N** – reference words

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

Также, существуют:

- CER - character error rate;
- SER - sentence error rate;

По возрастанию строгости метрик: CER-WER-SER

Различаются только юнитами: символы-слова-предложения

Считаем с помощью алгоритма поиска редакционного расстояния

Метрики оценки качества транскрибации

- Word Error Rate (**WER**)
  - **S** – substitutions
  - **I** – insertions
  - **D** – deletions
  - **N** – reference words

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

Может ли WER быть > 1?

Также, существуют:

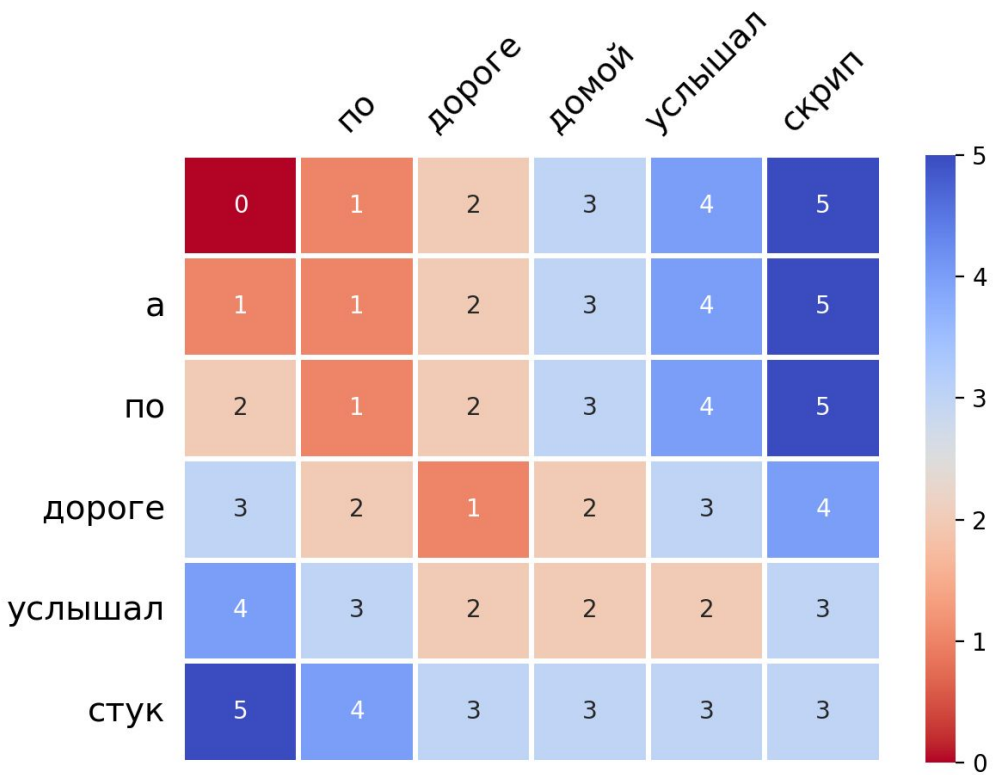
- CER - character error rate;
- SER - sentence error rate;

По возрастанию строгости метрик: CER-WER-SER

Различаются только юнитами: символы-слова-предложения

Считаем с помощью алгоритма поиска редакционного расстояния

Метрики оценки качества транскрибации

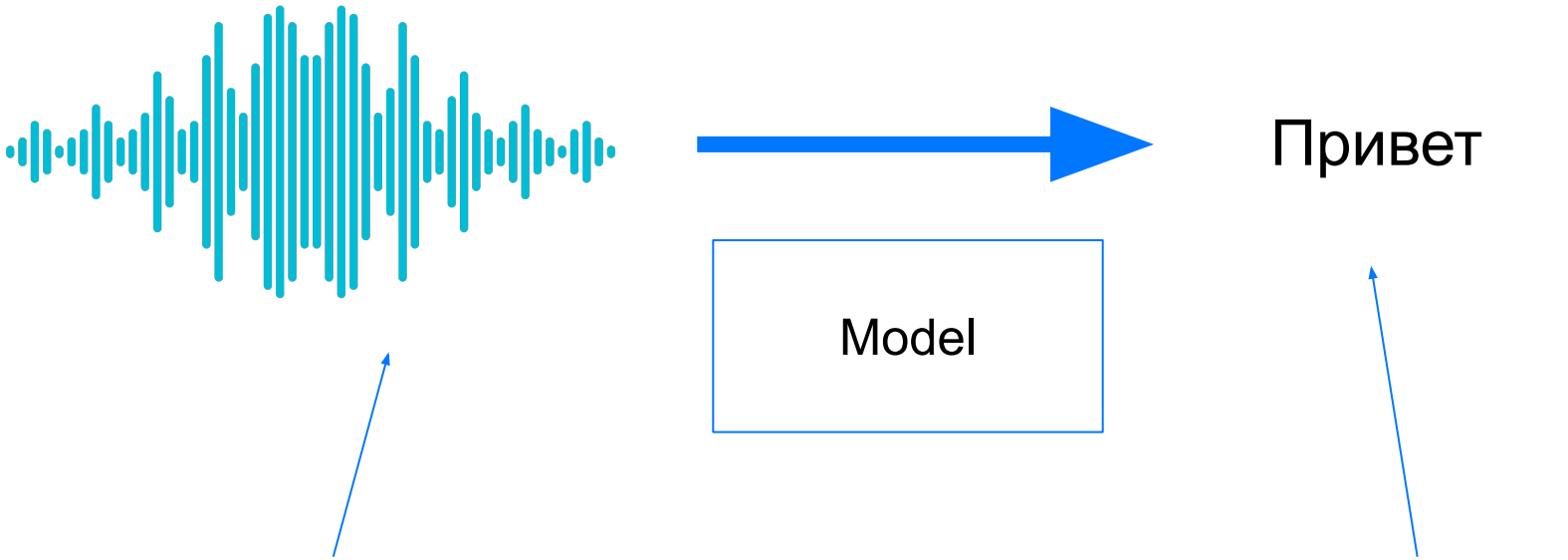


$$\mathrm{WER} = \frac{S + I + D}{N}$$

$d(S_1, S_2) = D(M, N)$, где

$$D(i,j) = \begin{cases} 0 & ; \ i = 0,\ j = 0 \\ i * \mathrm{deleteCost} & ; \ j = 0,\ i > 0 \\ j * \mathrm{insertCost} & ; \ i = 0,\ j > 0 \\ D(i-1, j-1) & ; \ S_1[i] = S_2[j] \\ \min (& \\ \quad D(i, j-1) + \mathrm{insertCost} & \\ \quad D(i-1, j) + \mathrm{deleteCost} & ; \ j > 0,\ i > 0,\ S_1[i] \neq S_2[j] \\ \quad D(i-1, j-1) + \mathrm{replaceCost} & \\ ) & \end{cases}$$

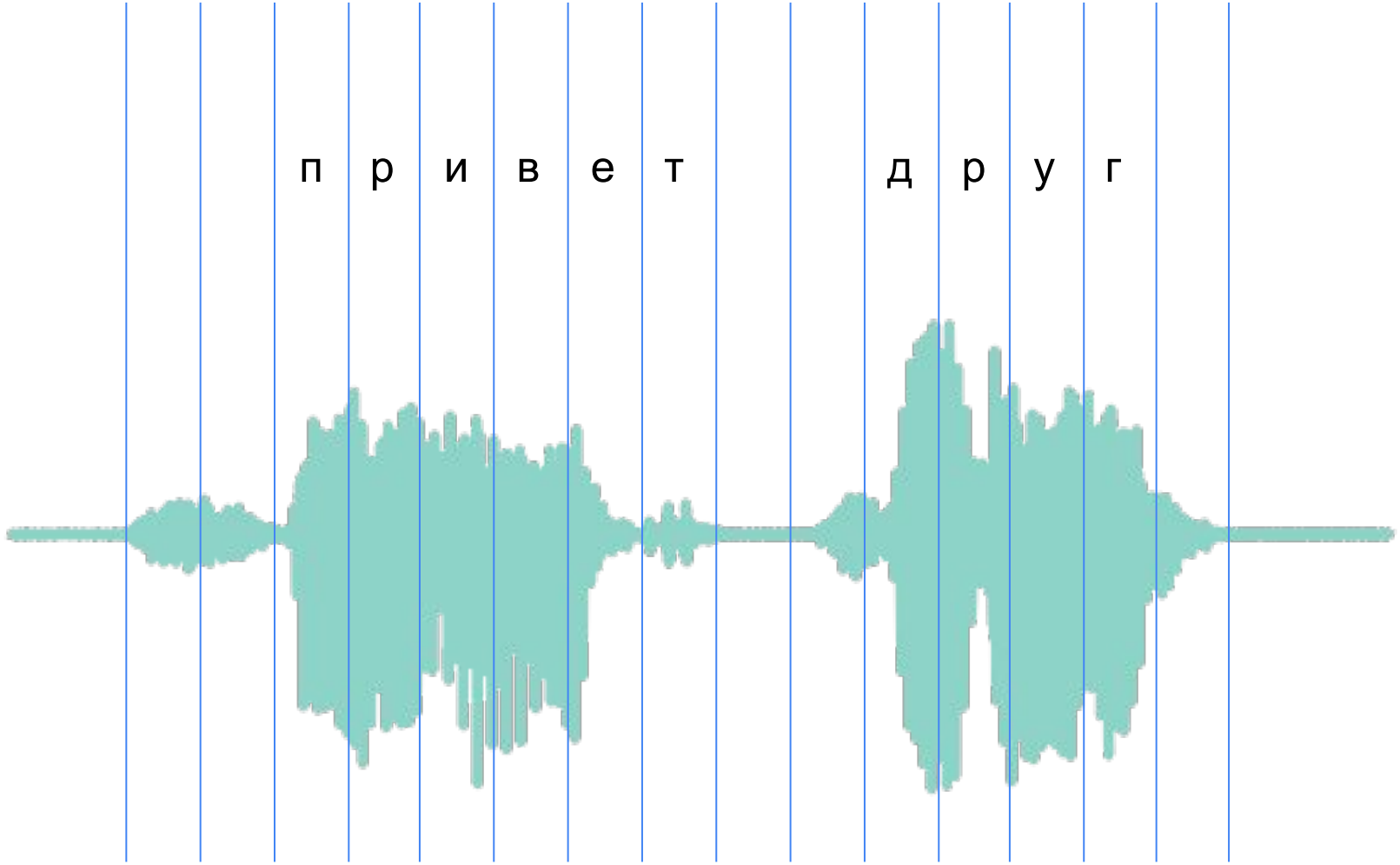
Постановка задачи ASR



Задача выравнивания последовательностей (АКА Seq2Seq)

Давайте сегментируем запись и отправим в разметку?

Будем классифицировать каждый чанк ??



Давайте сегментируем запись и отправим в разметку?

- Дорого
- Сложно (нужно очень много данных)
- Неустойчивый подход

Back in 2006 Connectionist Temporal Classification (CTC)

CTC

B-преобразование - удаление повторов и <BLANK>

привет

B

|прииивееет|

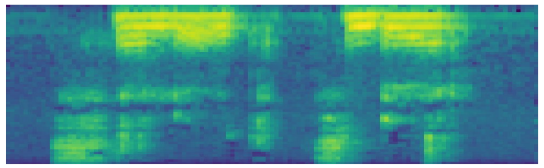
Логиты (|V| x seq_len)

контекстно независимый декодер

Декодер

хиддены энкодера (hi)

привет



Энкодер

Зачем нужен <BLANK> (|) токен

п	р	и	и	в	в	е	е	е	т			

→ привет

к	и	и	л	о	о	г	р	а	м	м	м	

→ килограмм

к	и	и	л	о	о	г	р	а	м	м	|	м

→ килограмм

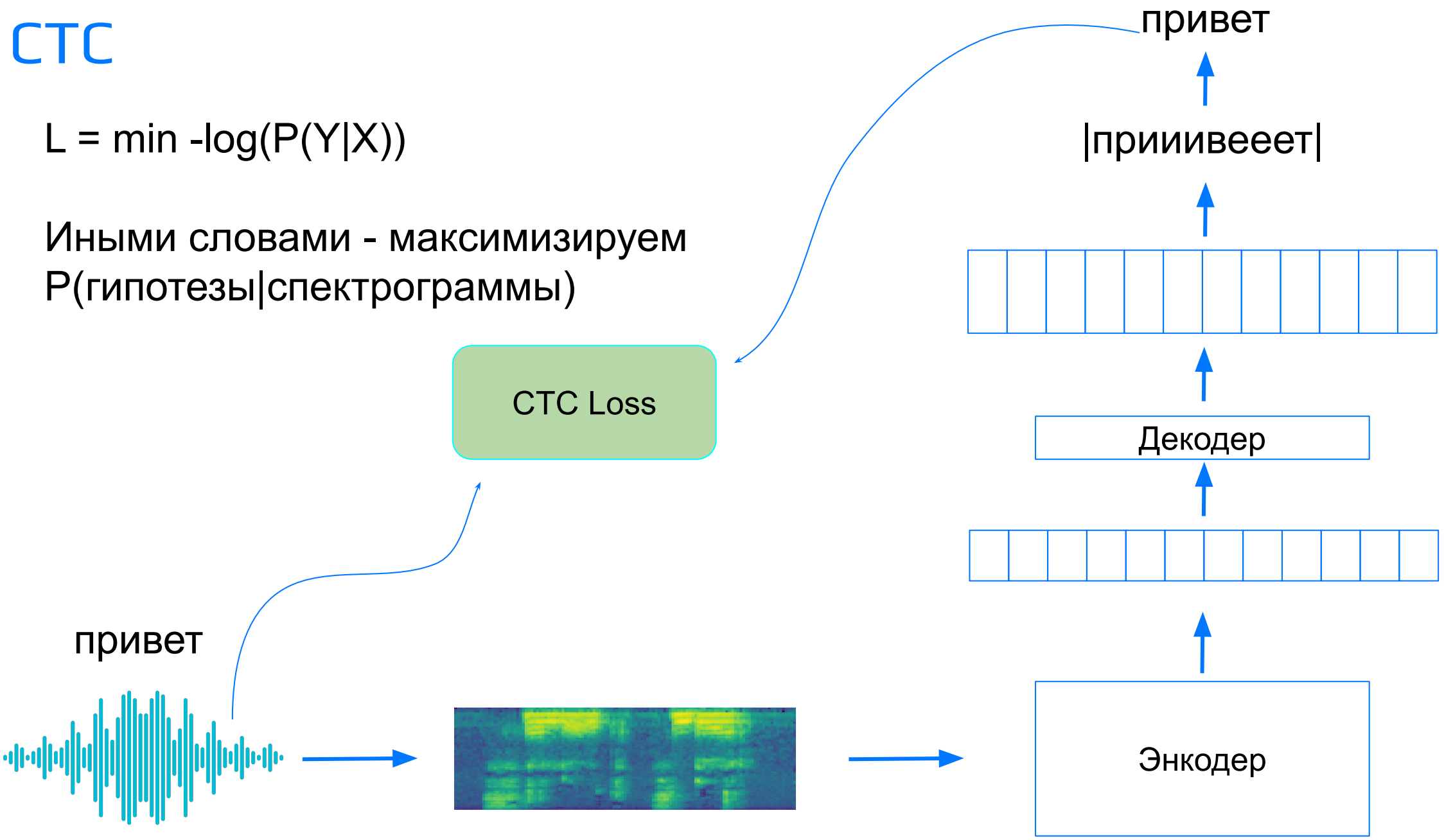
B-преобразование: сначала убираем дубликаты букв, затем удаляем <blank>.

Как обучать?
Где взять GT?

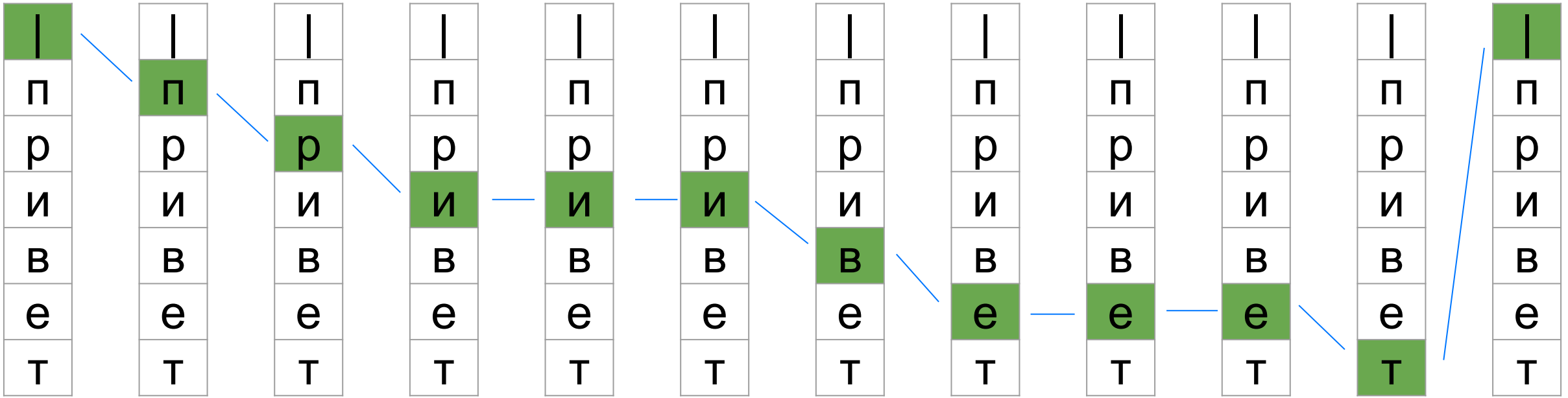
CTC

$L = \min -\log(P(Y|X))$

Иными словами - максимизируем
P(гипотезы|спектрограммы)

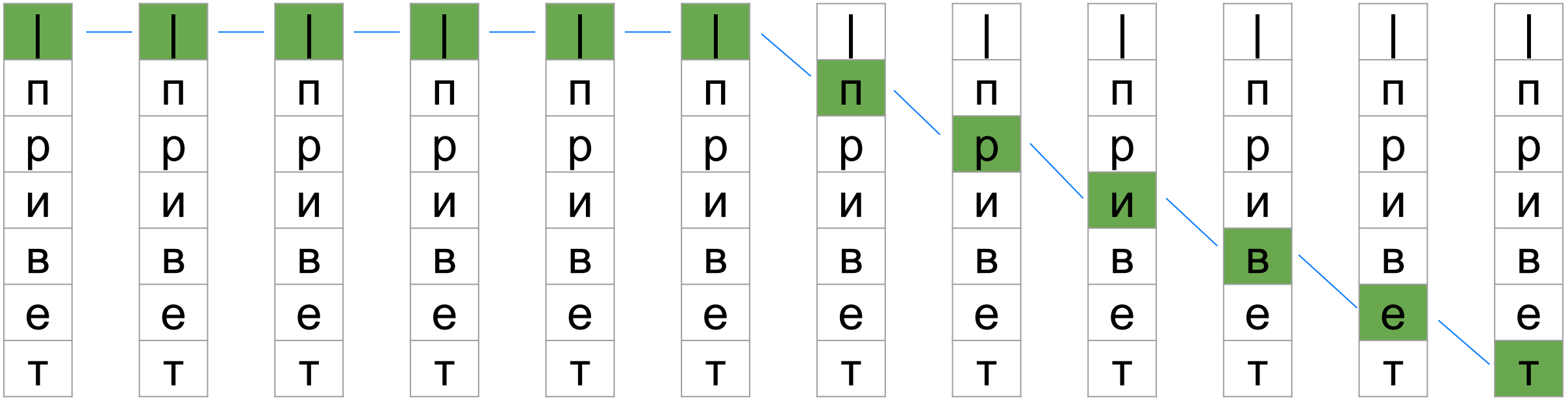


Матрица логитов (|V| x seq_len). Возможный правильный путь.



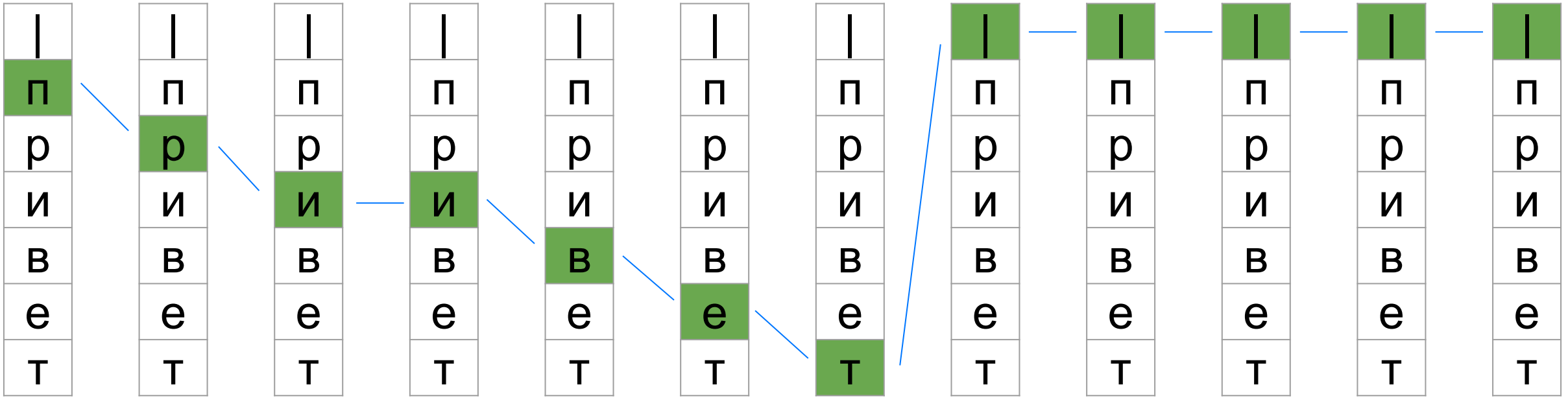
$$P(|\text{прииивееет}|) = y_{|}^{1} \cdot y_{\text{п}}^{2} \cdot y_{\text{р}}^{3} \cdot y_{\text{и}}^{4} \cdot y_{\text{и}}^{5} \cdot y_{\text{и}}^{6} \cdot y_{\text{в}}^{7} \cdot y_{\text{е}}^{8} \cdot y_{\text{е}}^{9} \cdot y_{\text{е}}^{10} \cdot y_{\text{т}}^{11} \cdot y_{|}^{12}$$

Матрица логитов (|V| x seq_len). Возможный правильный путь.



$$P(||||||\text{привет}) = y_{|}^{1} \cdot y_{|}^{2} \cdot y_{|}^{3} \cdot y_{|}^{4} \cdot y_{|}^{5} \cdot y_{|}^{6} \cdot y_{\text{п}}^{7} \cdot y_{\text{р}}^{8} \cdot y_{\text{и}}^{9} \cdot y_{\text{в}}^{10} \cdot y_{\text{е}}^{11} \cdot y_{\text{т}}^{12}$$

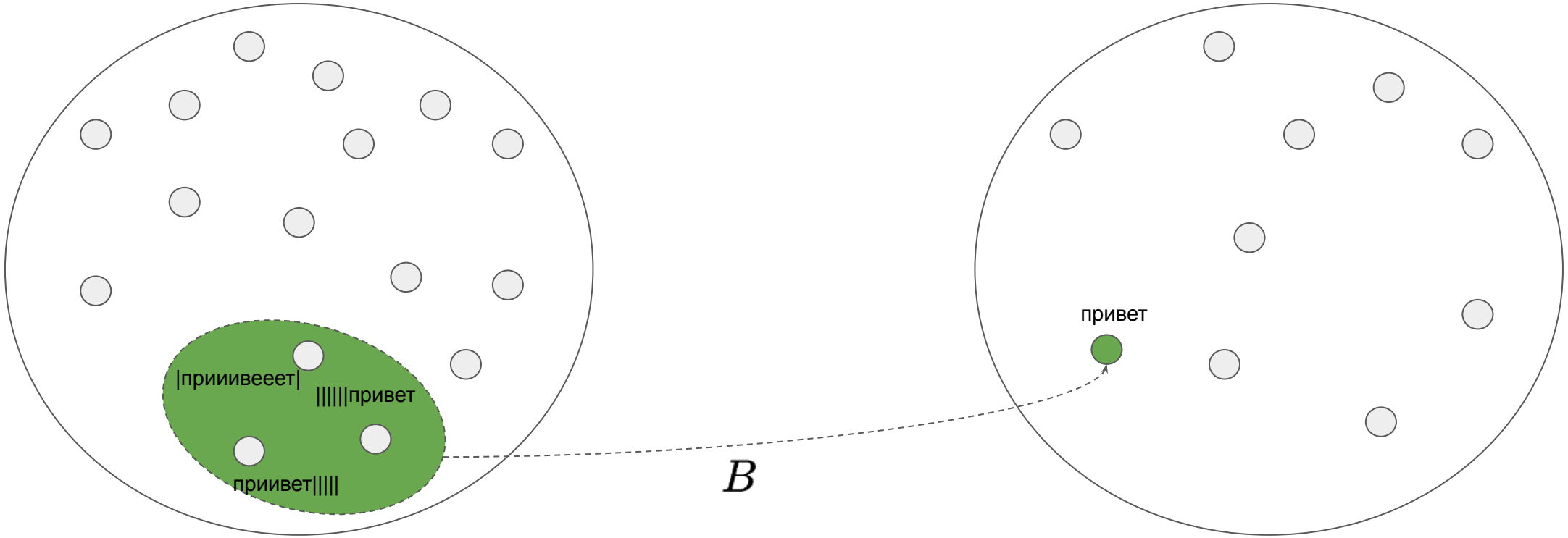
Матрица логитов (|V| x seq_len). Возможный правильный путь.



$$P(\text{приивет}|||||) = y_{\text{п}}^{1} \cdot y_{\text{р}}^{2} \cdot y_{\text{и}}^{3} \cdot y_{\text{и}}^{4} \cdot y_{\text{в}}^{5} \cdot y_{\text{е}}^{6} \cdot y_{\text{т}}^{7} \cdot y_{|}^{8} \cdot y_{|}^{9} \cdot y_{|}^{10} \cdot y_{|}^{11} \cdot y_{|}^{12}$$

CTC Loss

$$P(\text{привет}) = P(|\text{прииивееет}|) + P(|||||\text{привет}) + P(\text{приивет}|||||) + \ldots$$



Множество всех возможных путей

Множество всех возможных гипотез

CTC Loss

$$\mathcal{L}(X, R) = -\log \sum_{C \in B^{-1}(R)} P(C|X) = -\log \sum_{C \in B^{-1}(R)} \prod_{t=1}^{T} p(c_t|X)$$

спека

GT гипотеза

гипотезы для которых R = B(C)

$$|V| = 7, seq\ len = 12, possible\ paths = 7^{12} \sim 14B$$

Как считать CTC лосс эффективно

Дополнительно про CTC Loss:

- CTC paper
- Connectionist Temporal Classification Loss
- Sequence Modeling With CTC

Encoders

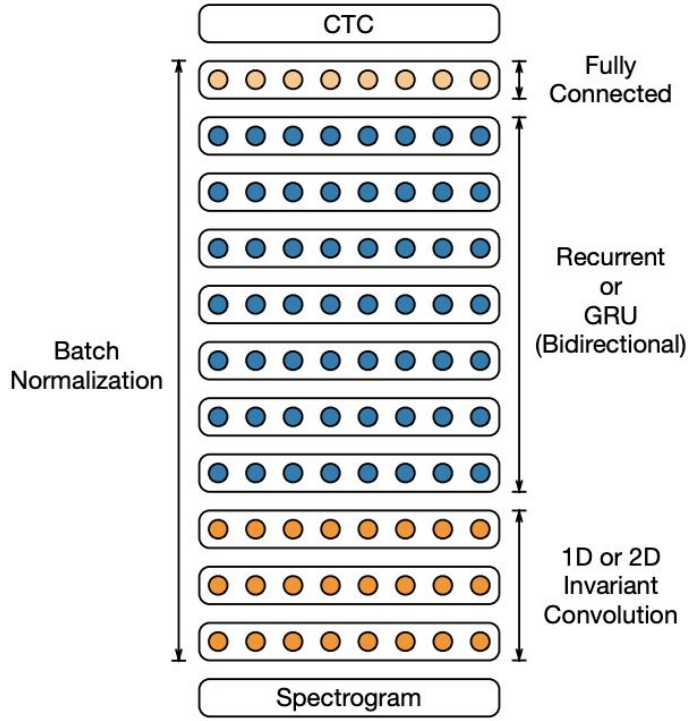
Энкодеры. Бенчмарки для сравнения.

- Librispeech ~1k часов аудиокниг
- WSJ ~80 часов (чтение текста из Wall Street Journal)
- Лидерборд моделей на HF

Read Speech			
Test set	DS1	DS2	Human
WSJ eval'92	4.94	3.60	5.03
WSJ eval'93	6.94	4.98	8.08
LibriSpeech test-clean	7.89	5.33	5.83
LibriSpeech test-other	21.74	13.25	12.69

Deep Speech 2

Read Speech			
Test set	DS1	DS2	Human
WSJ eval'92	4.94	3.60	5.03
WSJ eval'93	6.94	4.98	8.08
LibriSpeech test-clean	7.89	5.33	5.83
LibriSpeech test-other	21.74	13.25	12.69



54	**Deep Speech 2**	5.33	×	Deep Speech 2: End-to-End Speech Recognition in English and Mandarin			2015

Jasper

Доклад ODS

Table 5: *LibriSpeech, WER (%)*

Model	E2E	LM	dev-clean	dev-other	test-clean	test-other
CAPIO (single) [23]	N	RNN	3.02	8.28	3.56	8.58
pFSMN-Chain [25]	N	RNN	2.56	7.47	2.97	**7.5**
DeepSpeech2 [26]	Y	5-gram	-	-	5.33	13.25
Deep bLSTM w/ attention [21]	Y	LSTM	3.54	11.52	3.82	12.76
wav2letter++ [27]	Y	ConvLM	3.16	10.05	3.44	11.24
LAS + SpecAugment [4] [28]	Y	RNN	-	-	*2.5*	*5.8*
Jasper DR 10x5	Y	-	3.64	11.89	3.86	11.95
Jasper DR 10x5	Y	6-gram	2.89	9.53	3.34	9.62
Jasper DR 10x5	Y	Transformer-XL	2.68	8.62	**2.95**	8.79
Jasper DR 10x5 + Time/Freq Masks [4]	Y	Transformer-XL	2.62	7.61	2.84	7.84

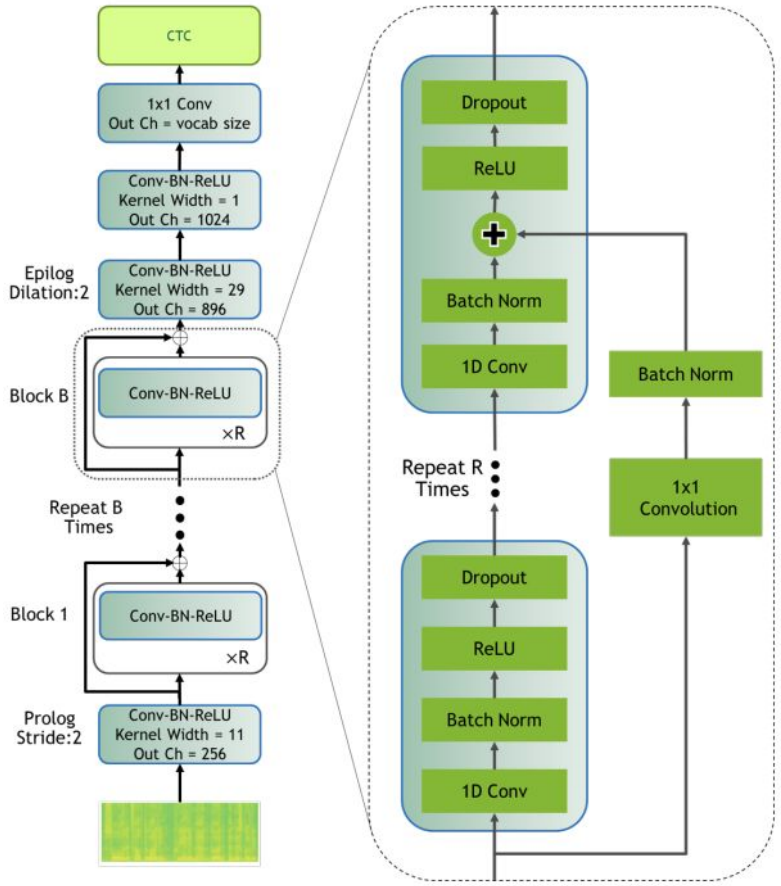
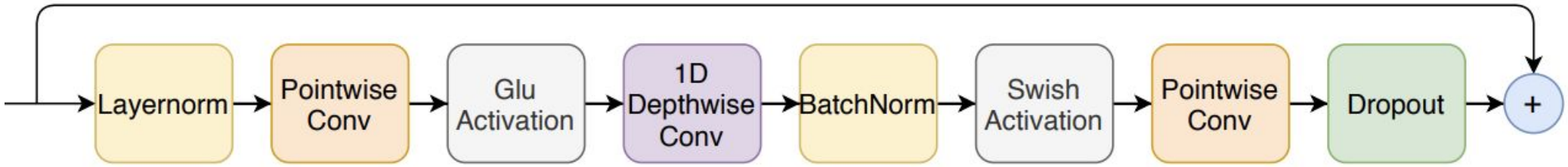
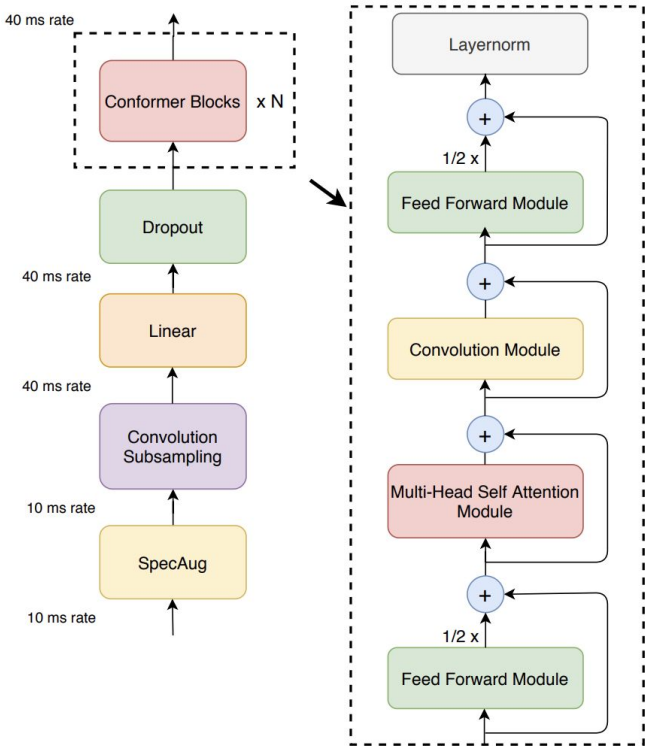


Figure 1: *Jasper BxR model: B - number of blocks, R - number of sub-blocks.*

44	**Jasper DR 10x5**	2.95	×	Jasper: An End-to-End Convolutional Neural Acoustic Model		2019

Conformer





Method	#Params (M)	WER Without LM		WER With LM	
		testclean	testother	testclean	testother
Hybrid					
Transformer [33]	-	-	-	2.26	4.85
CTC					
QuartzNet [9]	19	3.90	11.28	2.69	7.25
LAS					
Transformer [34]	270	2.89	6.98	2.33	5.17
Transformer [19]	-	2.2	5.6	2.6	5.7
LSTM	360	2.6	6.0	2.2	5.2
Transducer					
Transformer [7]	139	2.4	5.6	2.0	4.6
ContextNet(S) [10]	10.8	2.9	7.0	2.3	5.5
ContextNet(M) [10]	31.4	2.4	5.4	**2.0**	4.5
ContextNet(L) [10]	112.7	**2.1**	4.6	**1.9**	4.1
Conformer (Ours)					
Conformer(S)	10.3	**2.7**	**6.3**	**2.1**	**5.0**
Conformer(M)	30.7	**2.3**	**5.0**	**2.0**	**4.3**
Conformer(L)	118.8	**2.1**	**4.3**	**1.9**	**3.9**

Дополнительно по теме:

- Golos - датасет бенчмарк на русском
- OpenSTT - русскоязычные датасеты и бенчмарки
- Swish activation
- GLU activation
- Macaron net
- Depthwise convolution
- Librispeech LB

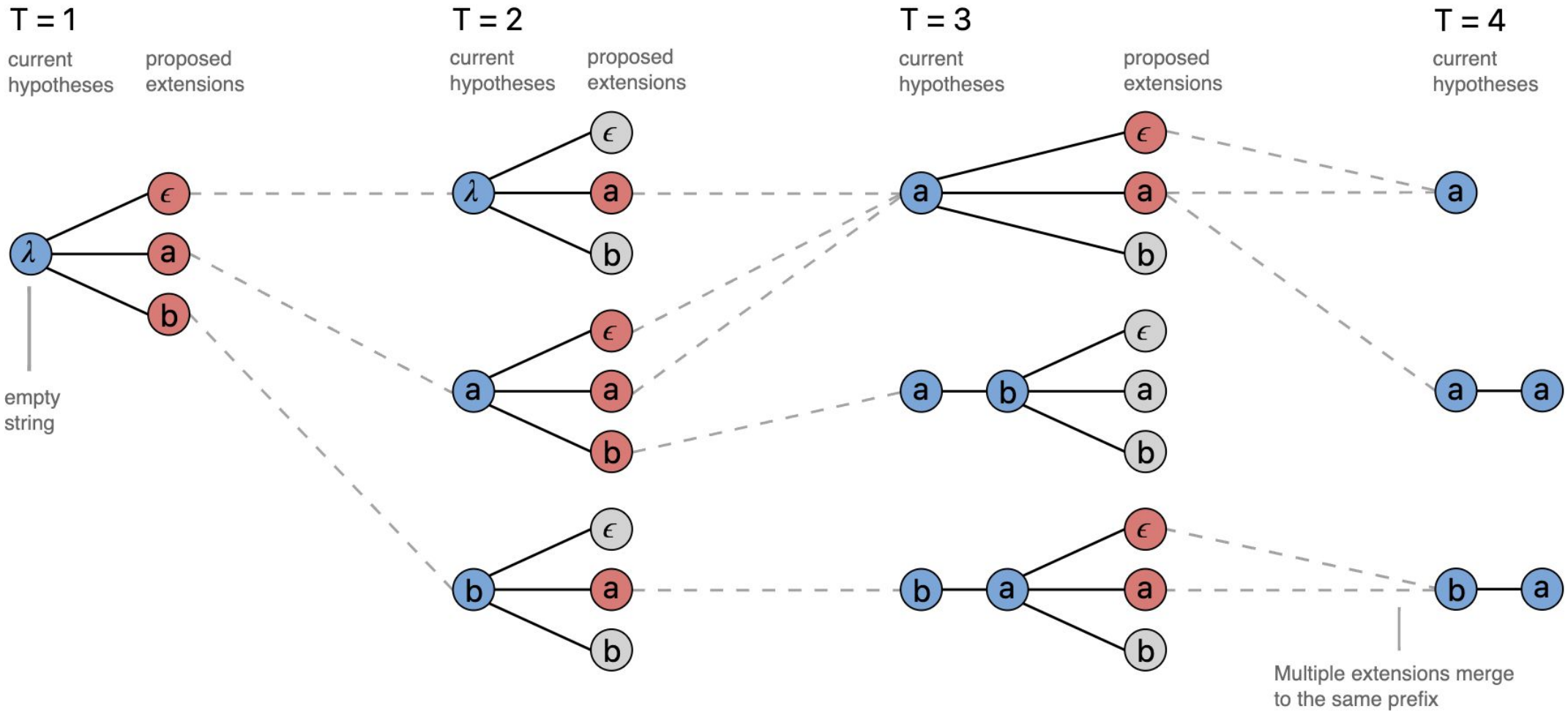
Decoding

Greedy decoding

<blank>	0.6	0.7
a	0.4	0.3
b	0.4	0.5

P(“”) = 0.7 * 0.6 = 0.42
P(a) = P(aa) + P(a|) + P(a|) = 0.4 * 0.3 + 0.6 * 0.3 + 0.2 * 0.7 = 0.44
P(a) > P(“”), но при жадном декодировании мы этого никогда не узнаем

Beam Search Decoding



Спасибо
за внимание!