



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ  
УНИВЕРСИТЕТ

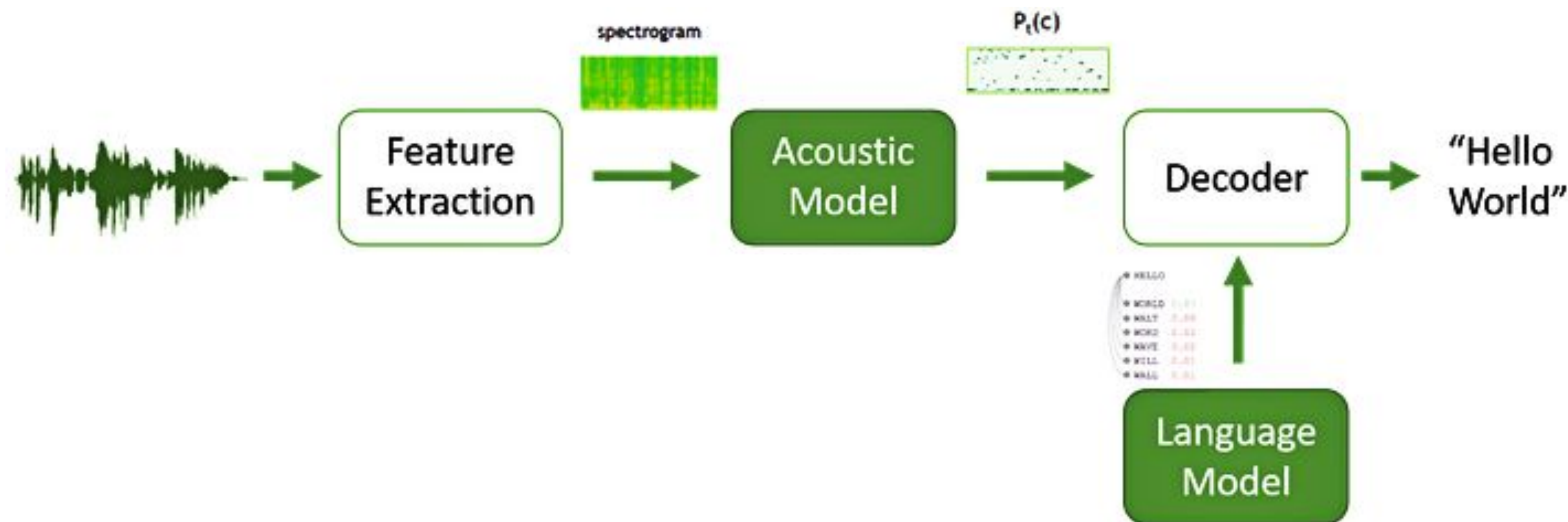
Факультет компьютерных наук

# On Generative Spoken Language Modeling from Raw Audio

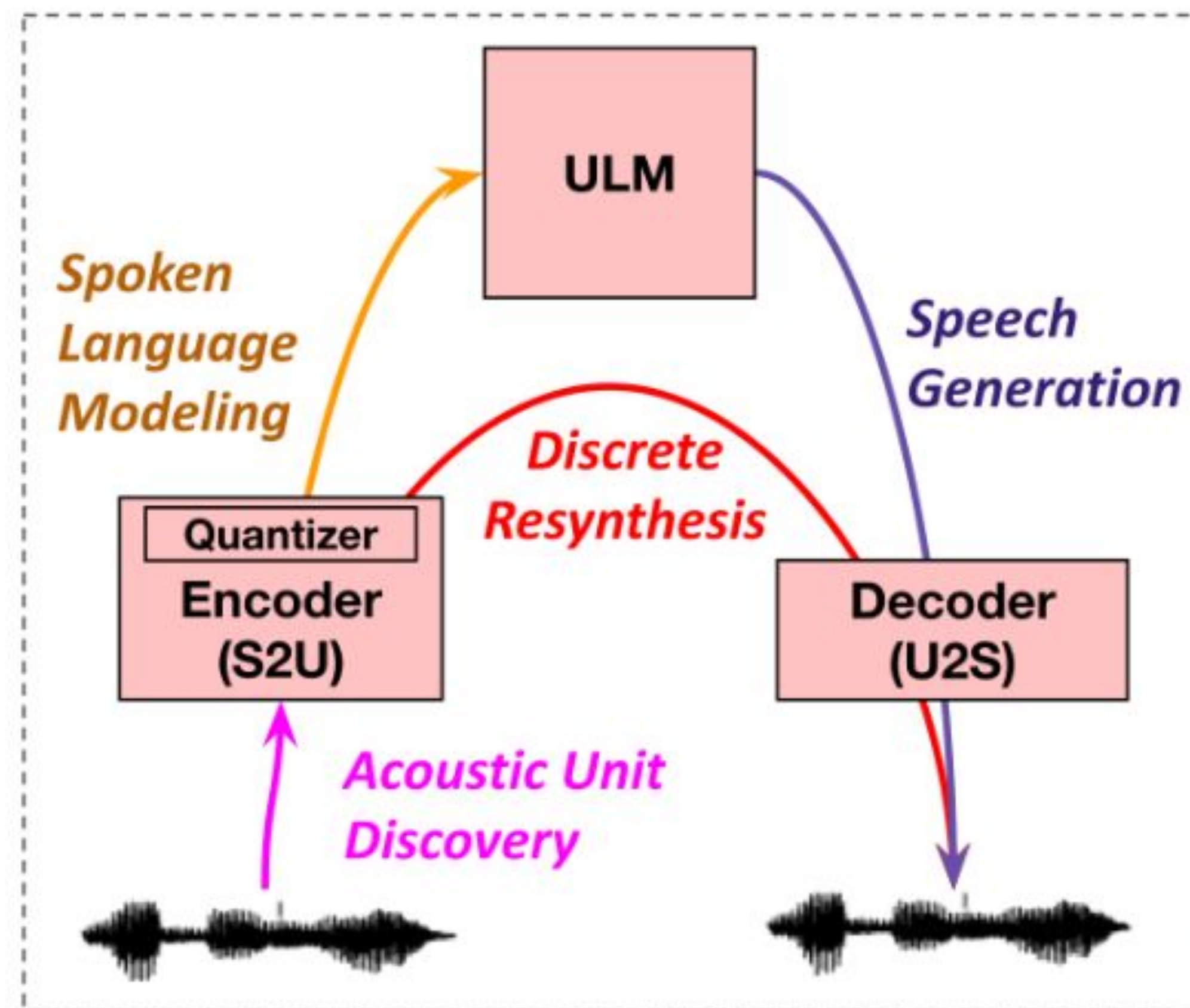
Авторы: Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, Emmanuel Dupoux

Москва, 2022

# ASR systems

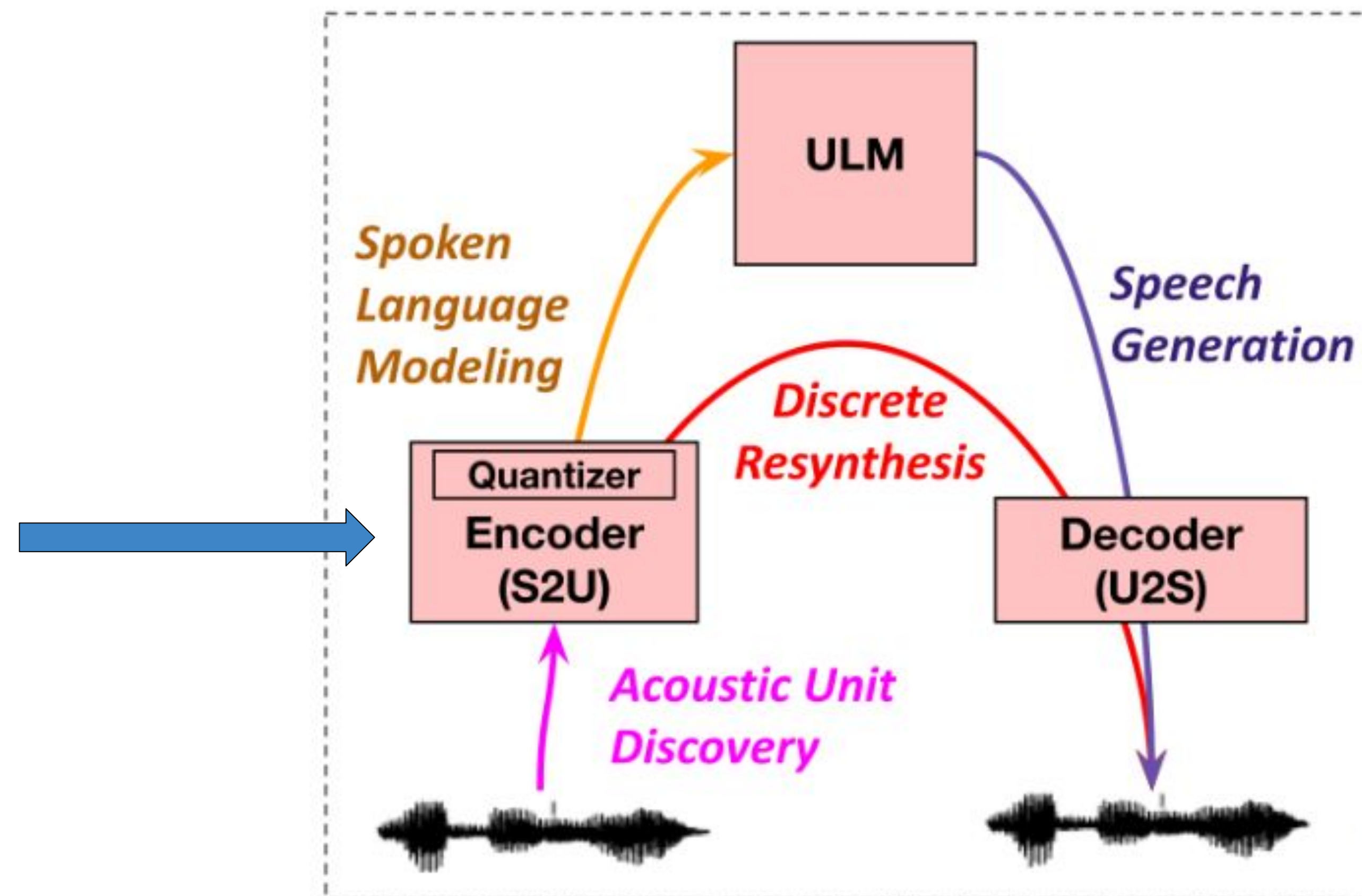


# Модель

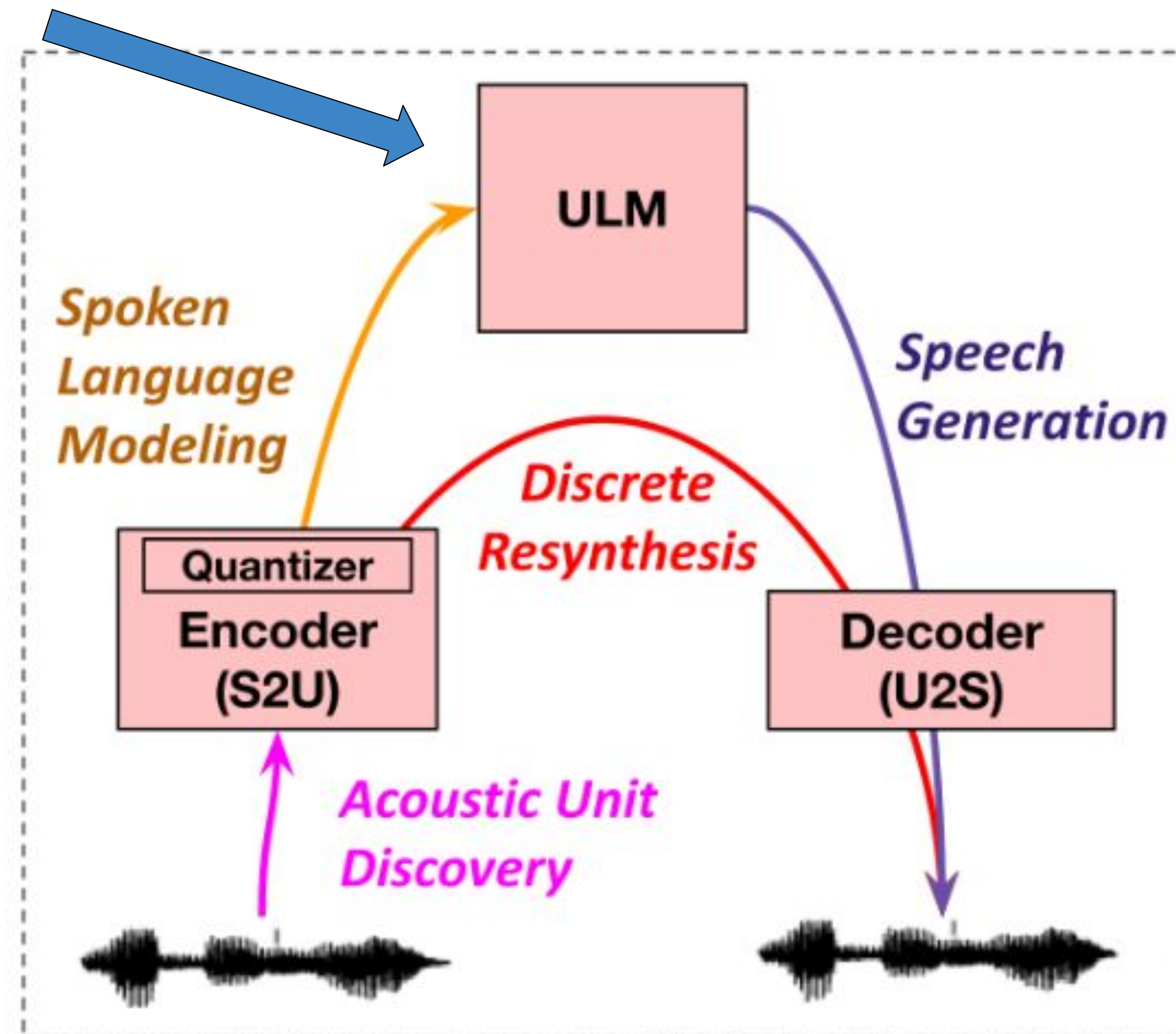




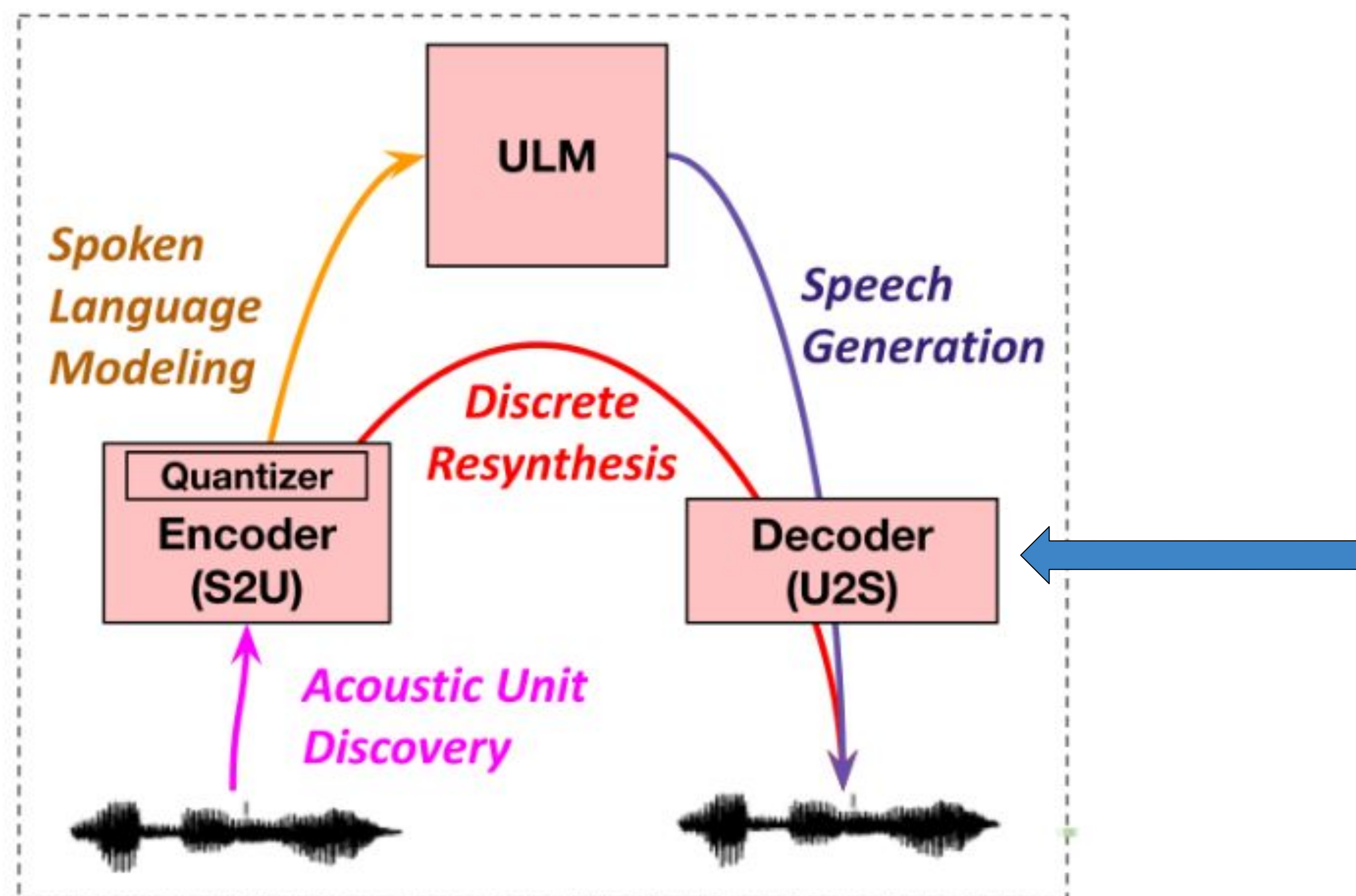
# Модель



# Модель

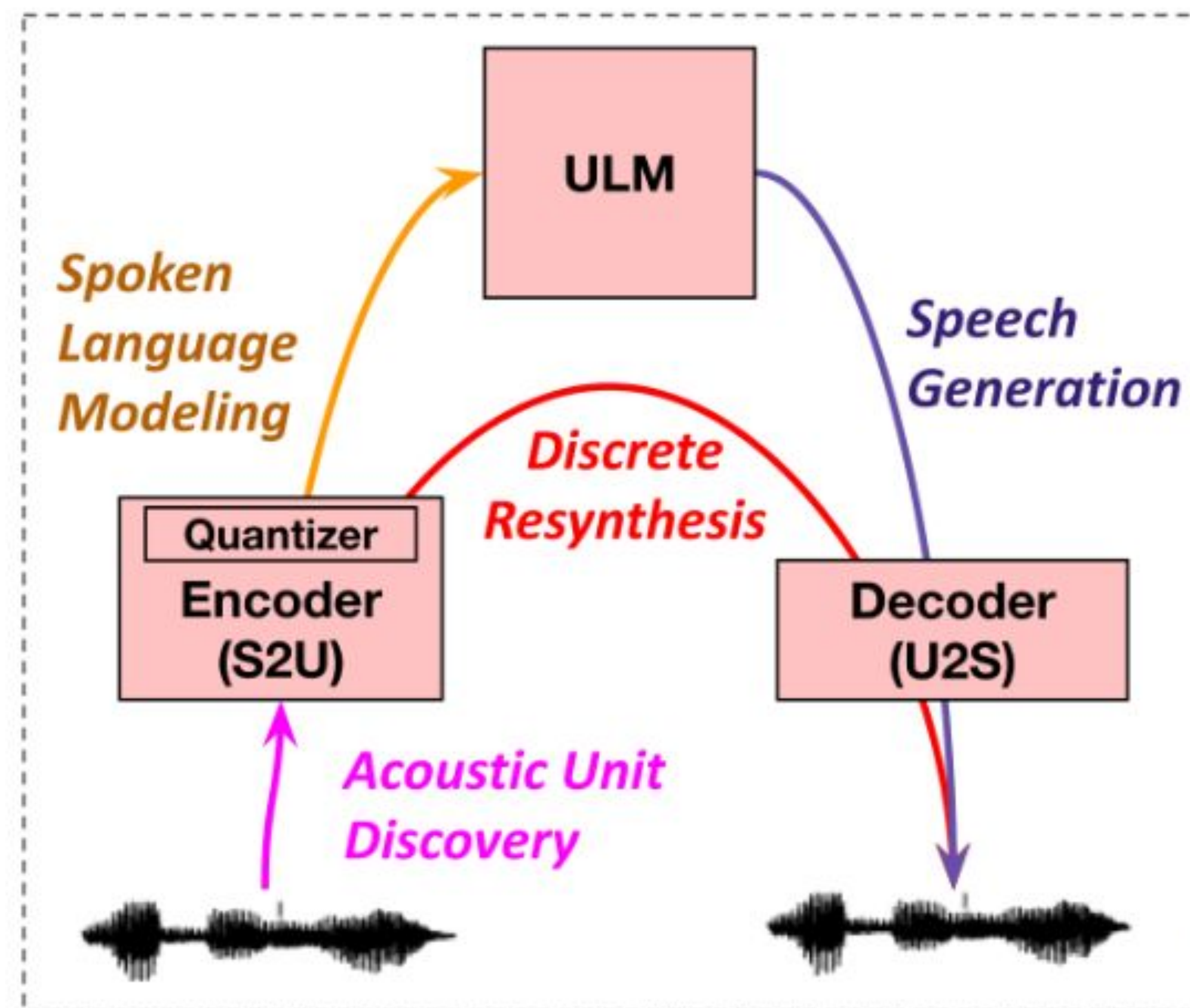


# Модель





# Модель

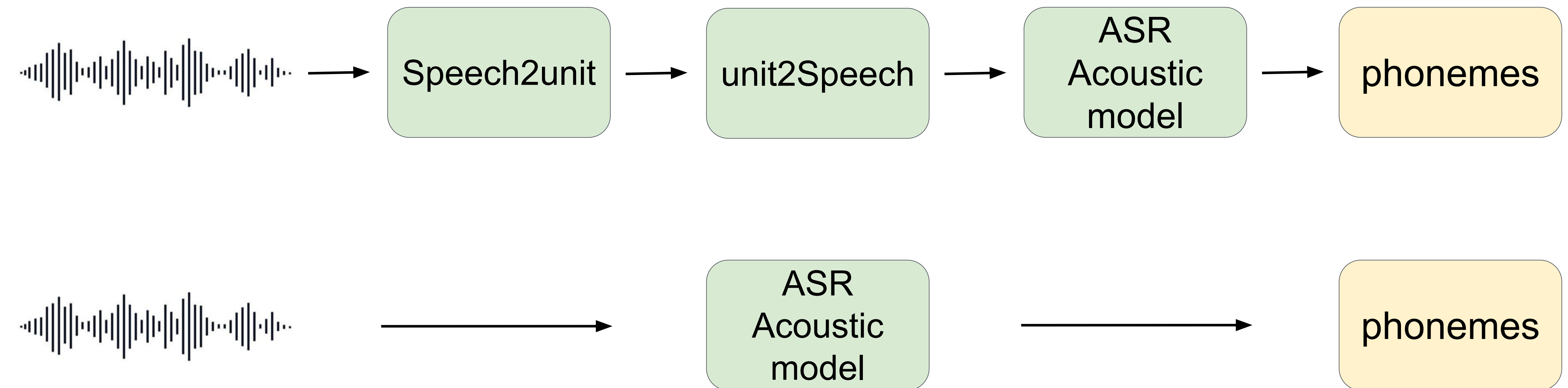


# Задачи и метрики

Level	Encoding		Generation		
	Task	Automatic metric	Task	Automatic metric	Human
Language	Spoken LM	<b>Spot-the-word</b> , Syntax-Acc	Speech Gen.	<b>AUC-of-VERT/PPX</b> , cont- BLEU, PPX@o-VERT	<b>MMOS</b>
Acoustic	Acoustic Unit Disc.	<b>ABX-across</b> , ABX-within	Resynthesis	<b>PER-from-ASR</b> , from-ASR	CER- <b>MOS</b>



# Ресинтез речи



## Phone Error Rate (PER)

$$PER = \frac{S + D + I}{N}$$

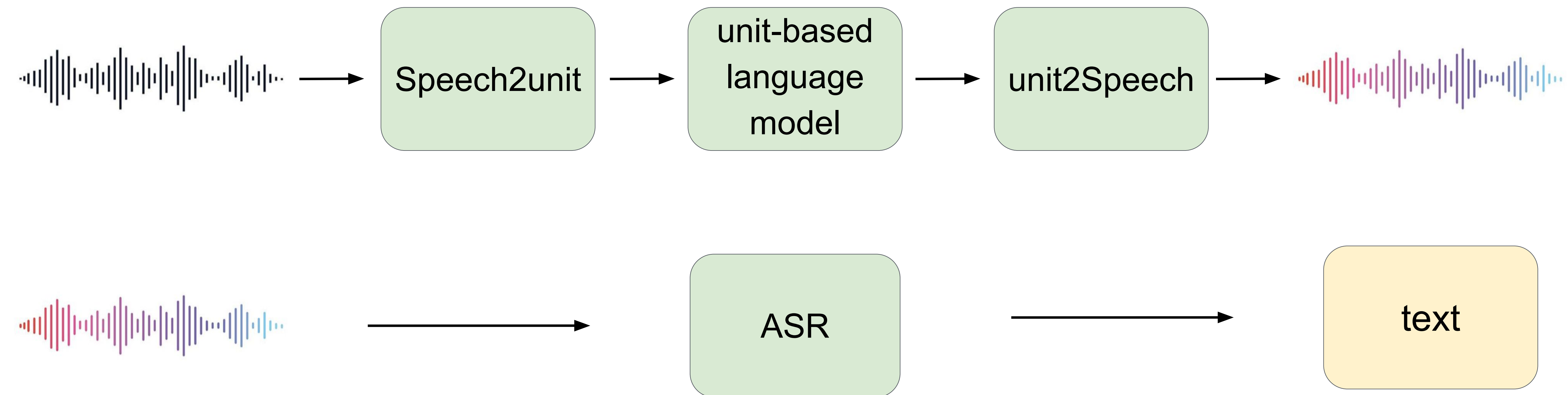
**S** – число замененных фонем

**D** – число пропущенных фонем

**I** – число вставленных фонем

**N** – общее число фонем в исходном тексте

# Генерация текста





# Генерация текста



- ↑  $T \Leftrightarrow$  Сгенерированная речь разнообразная, но качество не очень хорошее
- ↓  $T \Leftrightarrow$  Качество сгенерированной речи хорошее, но она не очень разнообразная

# Разнообразие речи

$$auto - BLEU(u, k) = \frac{\sum_s \left[ s \in (NG_k(u) \setminus s) \right]}{|NG_k(u)|}$$

$NG_k(u)$  – все  $k$ -граммы предложения  $u$

# Разнообразие речи

$$VERT(s) = \sqrt{Avg_u(auto - Bleu(u, 2)) \cdot Avg_u(self - Bleu(u_1, u_2, 2))}$$

(diVERsiTy)

$s$  – сгенерированный текст

$u_1, u_2, u$  – сгенерированные предложения



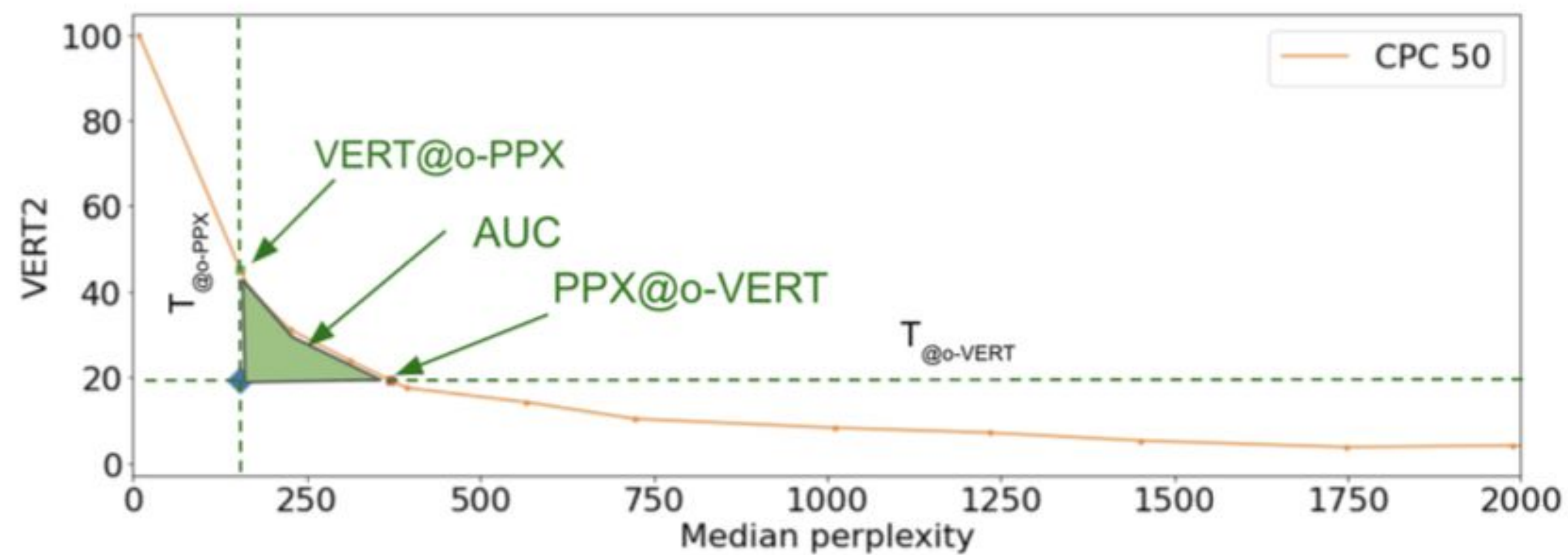
## Качество речи

$$PPX(w) = P(s_1, \dots, s_m)^{-\frac{1}{m}}$$

$s_1, \dots, s_m$  — части сгенерированной речи

$m$  — ЧИСЛО СЛОВ

# Разнообразие и качество речи



# Кодирование на акустическом уровне

$$ABX - score = \frac{\sum_{a, x \in A, x \neq a} \sum_{b \in B} \left( [d(b, x) < d(a, x)] + \frac{1}{2} [d(b, x) = d(a, x)] \right)}{n_A (n_A - 1) n_B}$$

$a, b, x$  — токены

$A, B$  — фонетические категории

$n_A, n_B$  — число элементов категории  $A$  и  $B$

$d(\cdot, \cdot)$  — мера расстояния



# Кодирование на языковом уровне

spot-the-word accuracy  
(лексика)

acceptability judgement  
(синтаксис)

brick

blick

She likes it

She like it

↑  
Такое слово есть

↑  
Такого слова нет

↑  
Так верно

↑  
Так не верно

# Человеческие оценки

Mean opinion score  
(MOS)

meaningfulness-MOS  
(MMOS)

**5** Excellent

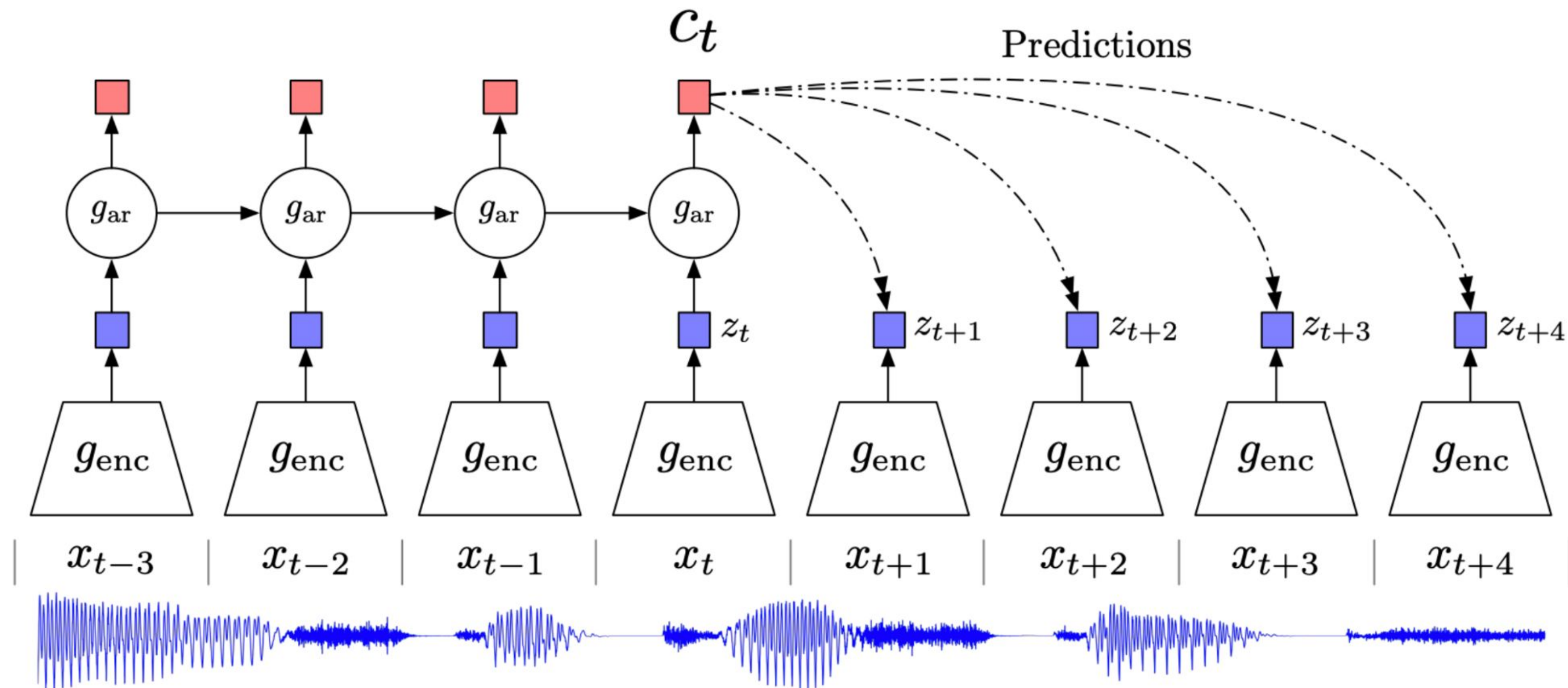
**4** Good

**3** Fair

**2** Poor

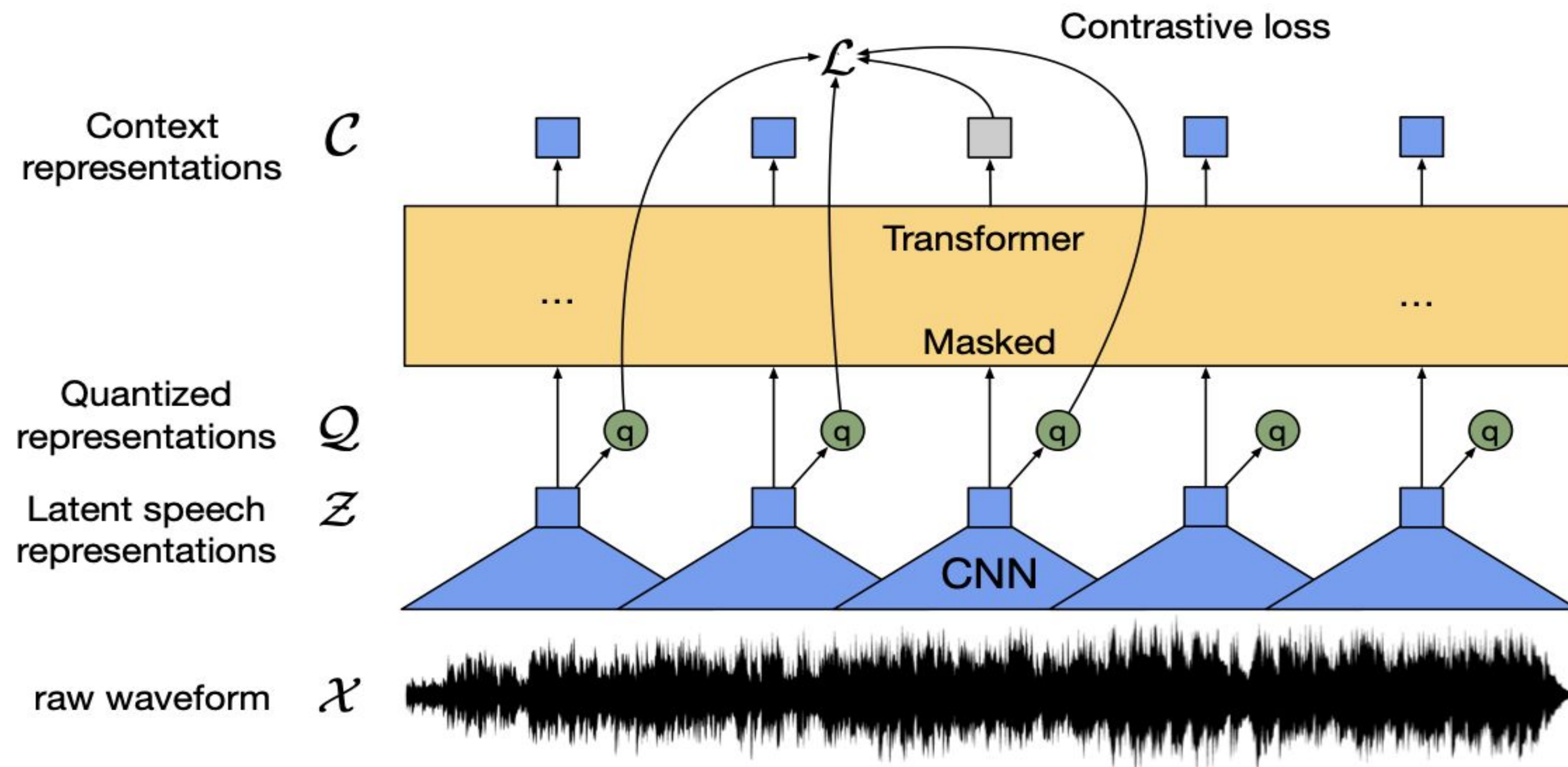
**1** Bad

# CPC модель (Speech2unit)

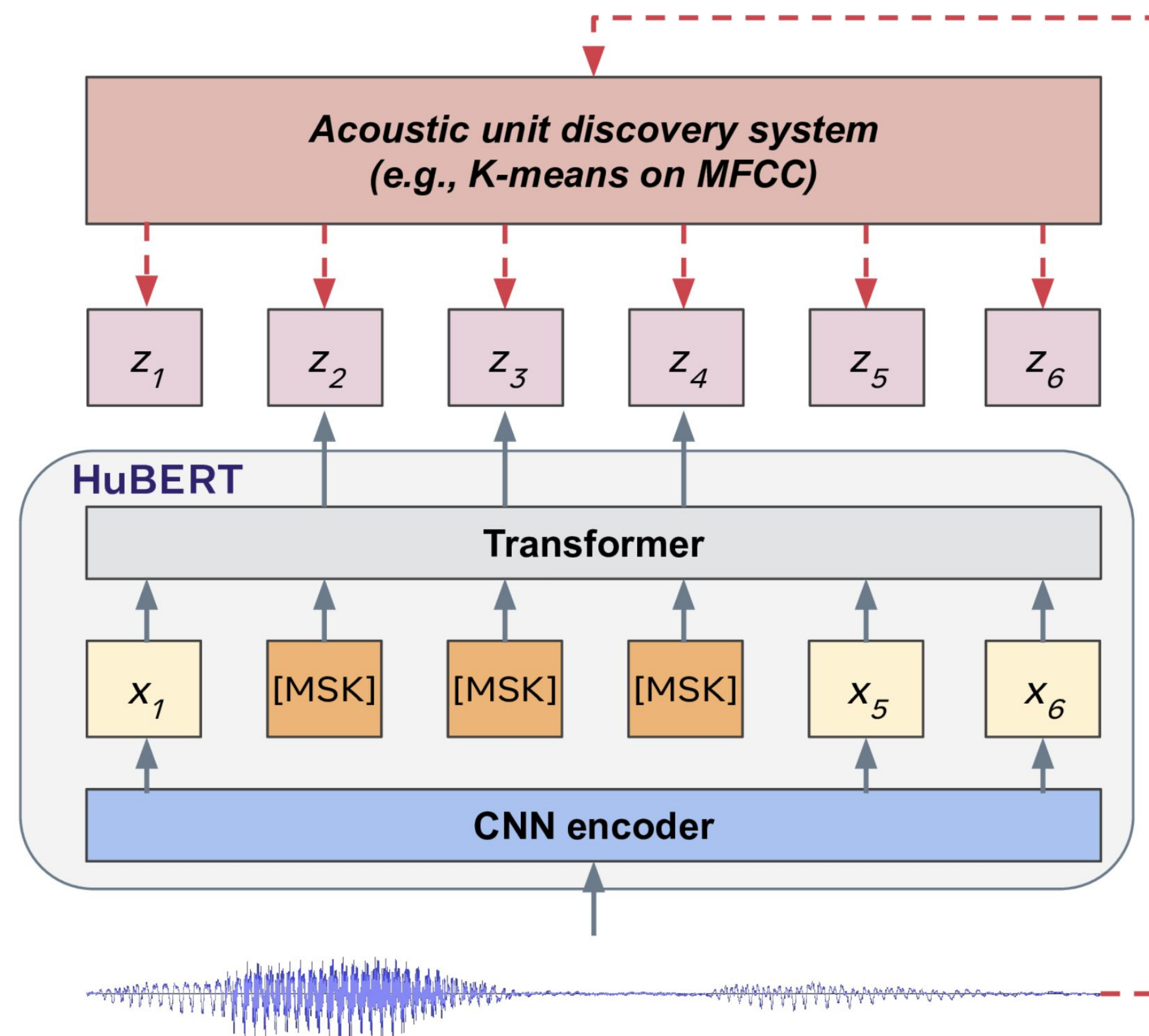




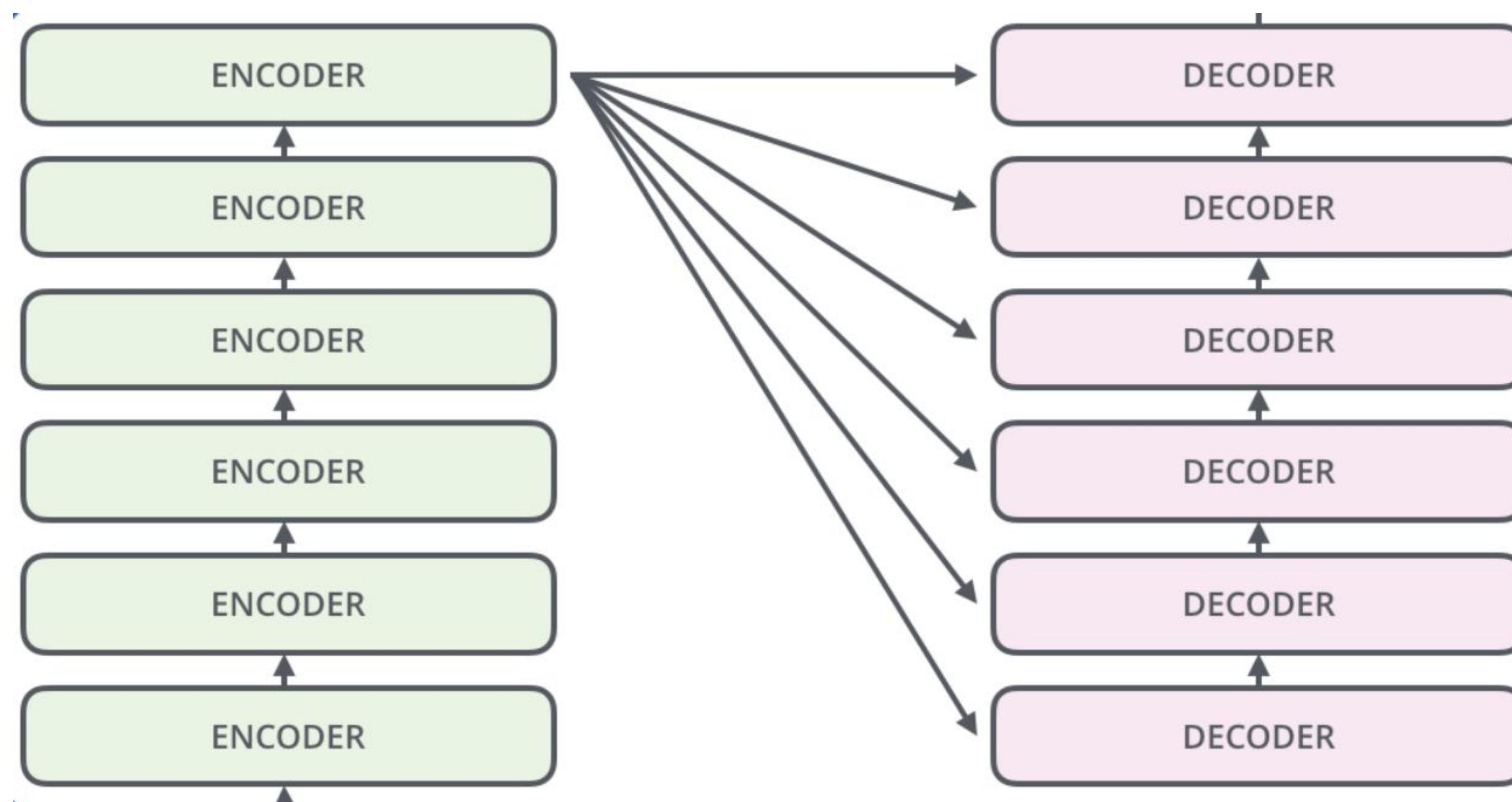
# Wave2vec модель (Speech2unit)



# HuBERT модель (Speech2unit)

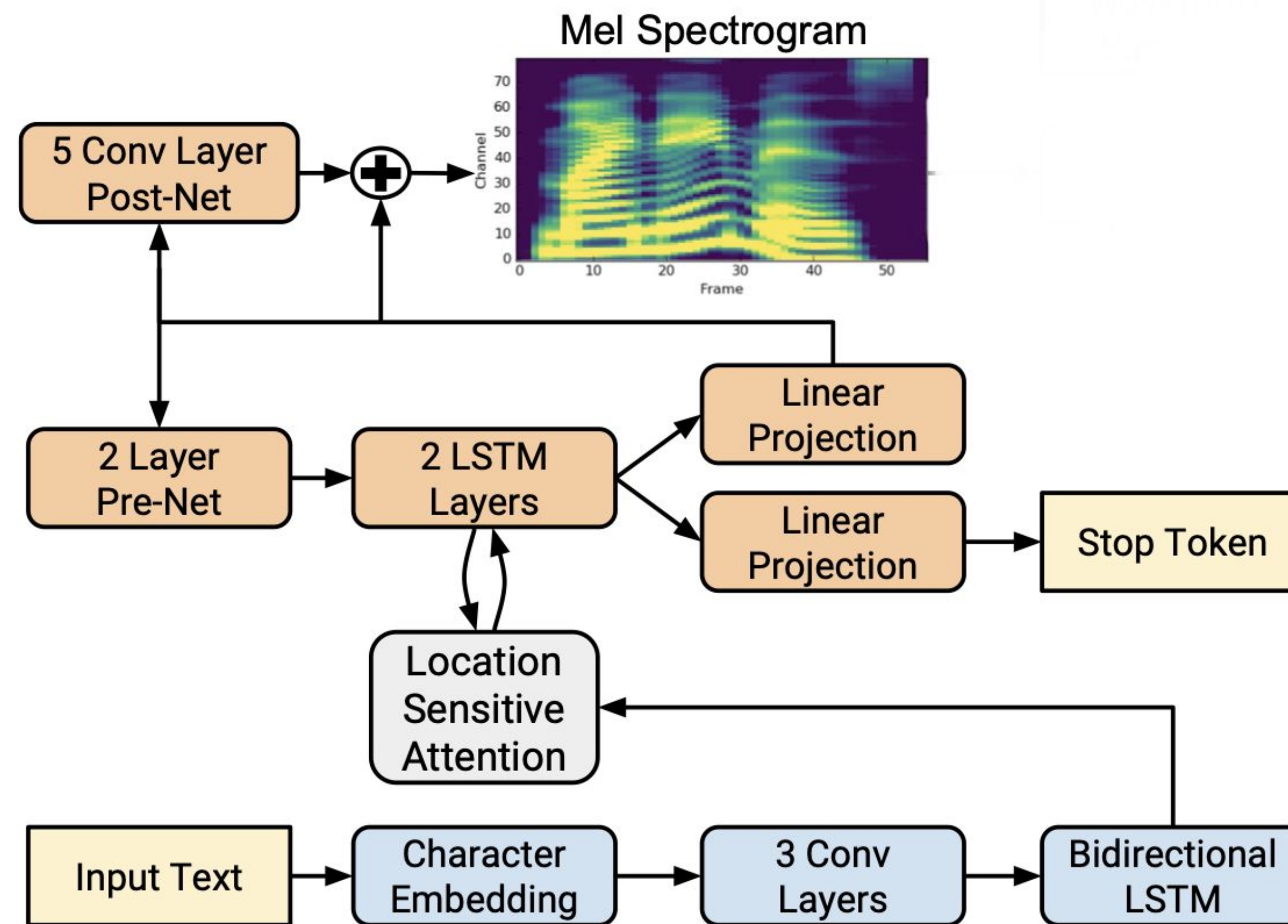


# Transformer (uLM)



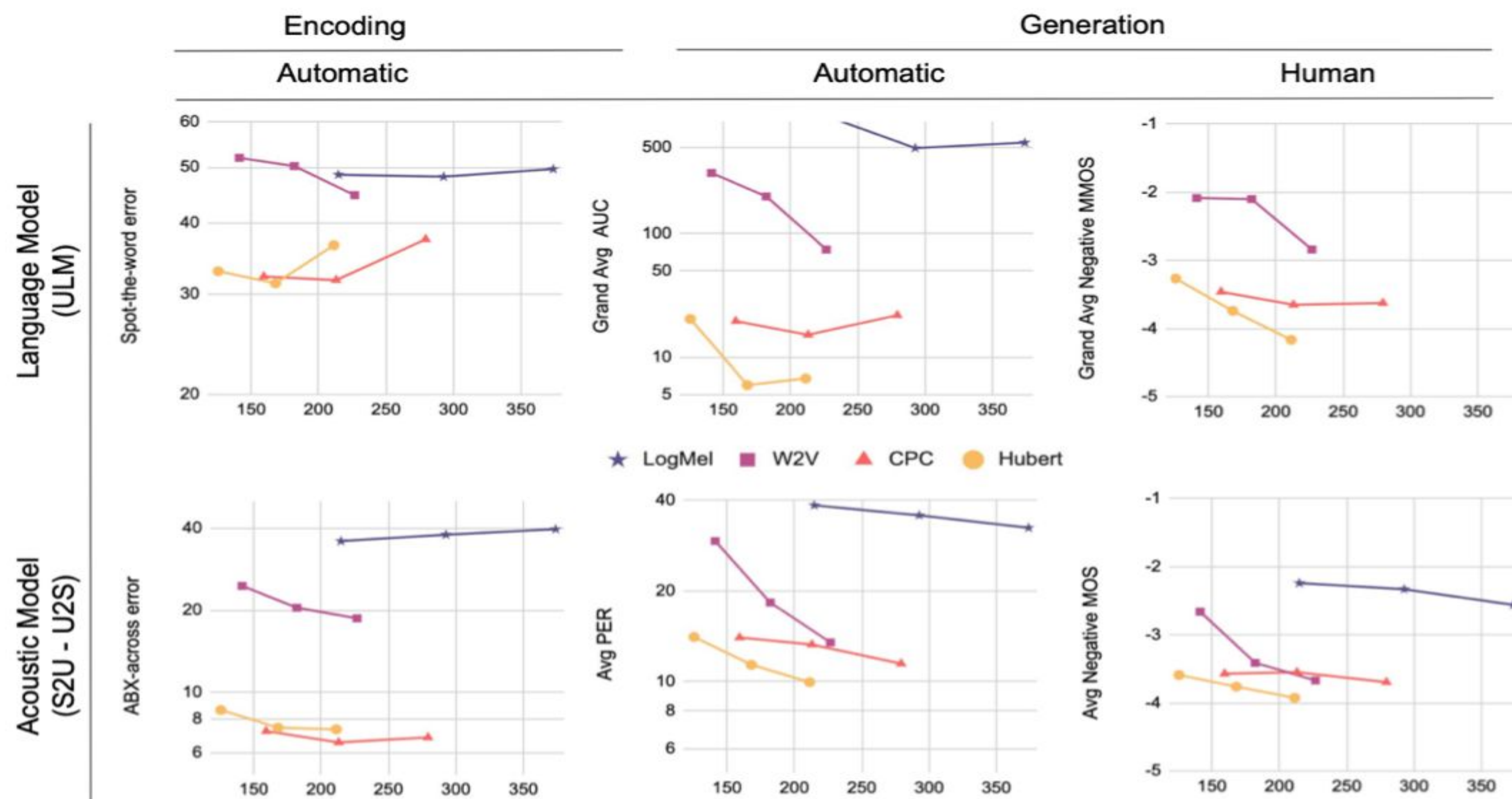


# Tacotron-2 модель (unit2Speech)

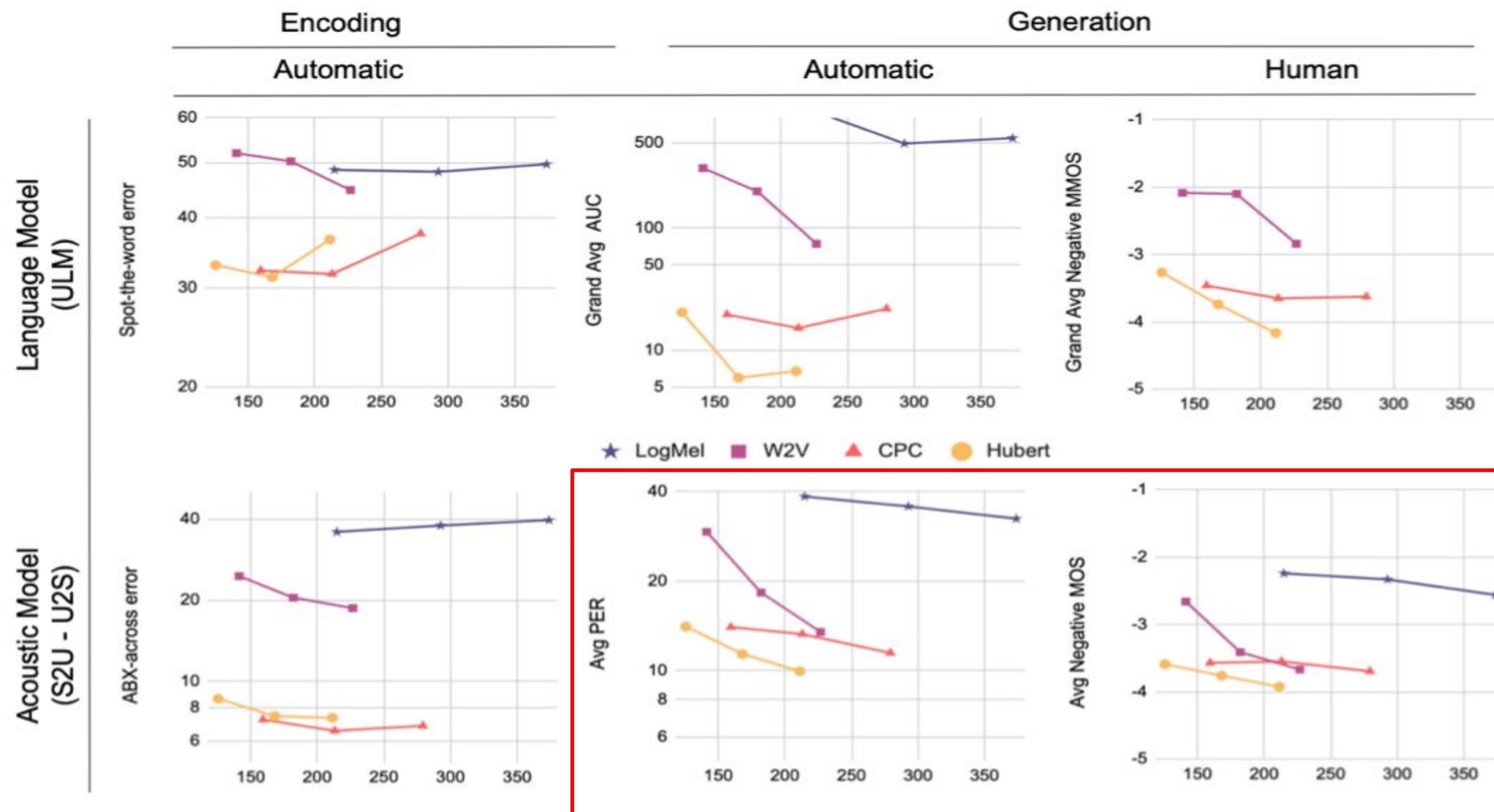




# Результаты



# Задача ресинтеза

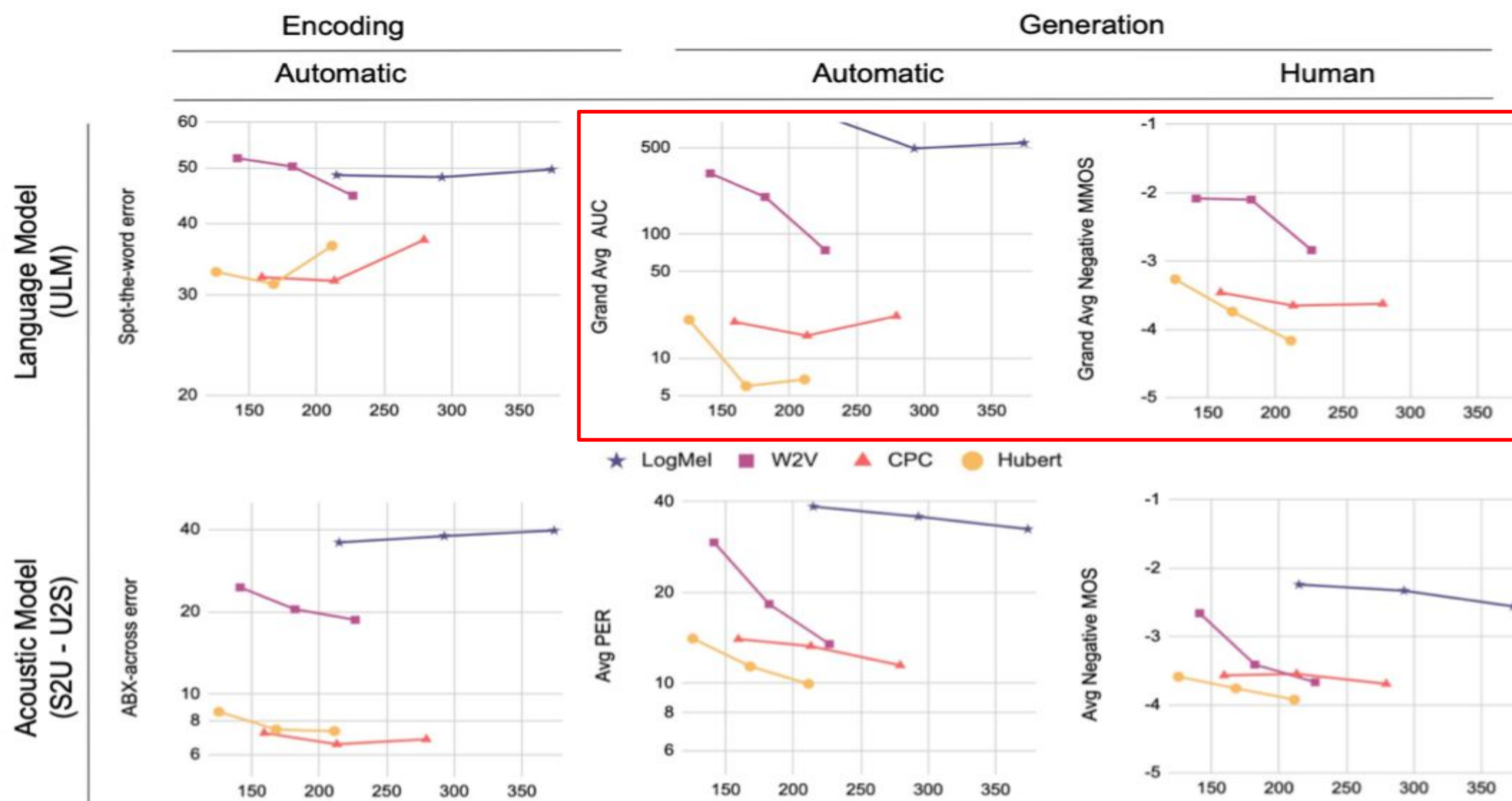




# Задача ресинтеза

Systems			End-to-end ASR-based metrics				Human Opinion			
S2u architect.	Nb units	Bit- rate	PER↓ (LJ)	PER↓ (LS)	CER↓ (LJ)	CER↓ (LS)	MOS↑ (LJ)	MOS↑ (LS)	CER↓ (LJ)	CER↓ (LS)
<i>Toplines</i>										
original wav			-	-	-	-	4.83	4.30	8.88	6.73
orig text+TTS			7.78	7.92	8.87	5.14	4.02	4.03	13.25	10.73
ASR + TTS	27		9.45	8.18	9.48	5.30	4.04	4.06	15.98	11.56
<i>Baselines</i>										
LogMel	50	214.8	27.72	49.38	27.73	52.05	2.41	2.07	43.78	66.75
LogMel	100	292.7	25.83	45.58	24.88	48.71	2.65	2.01	37.39	62.72
LogMel	200	373.8	19.78	45.16	17.86	46.12	2.96	2.16	23.33	62.6
<i>Unsupervised</i>										
CPC	50	159.4	10.87	17.16	10.68	12.06	3.63	3.51	13.97	19.92
CPC	100	213.1	10.75	15.82	9.84	9.46	3.42	3.68	13.53	14.73
CPC	200	279.4	<b>8.74</b>	14.23	9.20	8.29	3.85	3.54	<b>9.36</b>	14.33
HuBERT-L6	50	125.7	11.45	16.68	11.02	11.85	3.69	3.49	14.54	13.14
HuBERT-L6	100	168.1	9.53	13.24	9.31	7.19	3.84	3.68	13.02	11.43
HuBERT-L6	200	211.3	8.87	<b>11.06</b>	<b>8.88</b>	<b>5.35</b>	<b>4.00</b>	<b>3.85</b>	11.67	<b>10.84</b>
wav2vec-L14	50	141.3	24.95	33.69	25.42	32.91	2.45	2.87	46.82	54.9
wav2vec-L14	100	182.1	14.58	22.07	13.72	17.22	3.50	3.32	23.76	28.1
wav2vec-L14	200	226.8	10.65	16.34	10.21	10.50	3.83	3.51	13.14	15.27

# Задача генерации



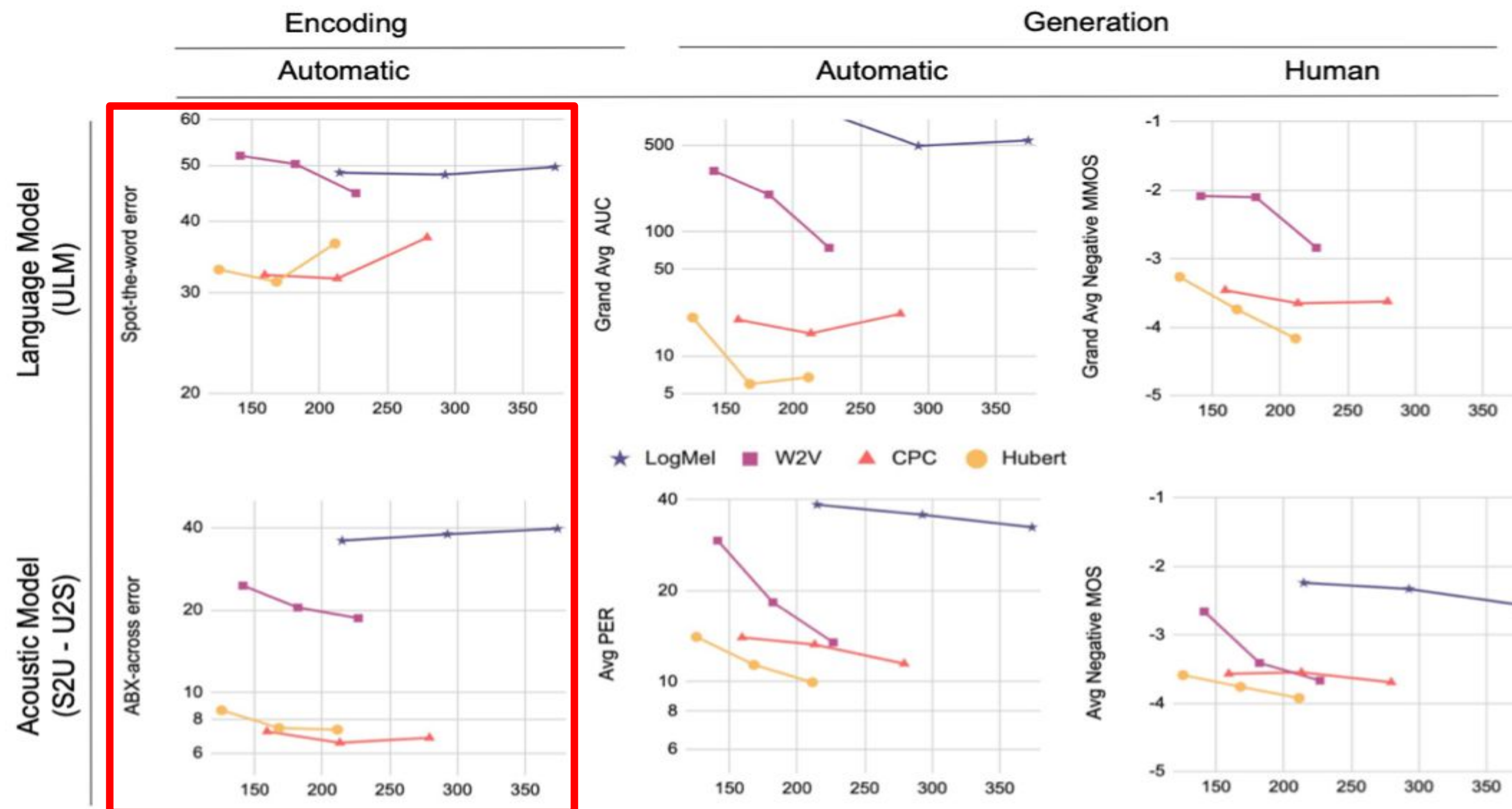


# Задача генерации

Systems		Generation based metrics						Human Opinion	
Encoder architect.	Nb units	<u>unconditional</u>			<u>prompt</u>			<u>uncond.</u>	<u>prompt</u>
		PPX↓	VERT↓	AUC↓	PPX↓	VERT↓	AUC↓	MMOS↑	MMOS↑
<i>Controls</i>									
oracle text		154.5	19.43	-	154.5	19.43	-	4.02	4.26
ASR + LM		178.4	21.31	0.18	162.8	20.49	0.04	3.91	4.38
<i>Baseline</i>									
LogMel	50	1588.97	-	1083.76	-	-	-	-	-
LogMel	100	1500.11	95.50	510.26	-	-	-	-	-
LogMel	200	1539.00	-	584.16	-	-	-	-	-
<i>Unsupervised</i>									
CPC	50	374.26	46.26	19.68	323.9	39.92	18.44	3.31	3.61
CPC	100	349.56	41.797	15.74	294.7	42.93	14.06	3.65	3.65
CPC	200	362.84	40.28	16.46	303.5	43.42	26.67	3.58	3.67
HuBERT-L6	50	376.33	43.06	19.27	339.8	45.85	21.03	3.53	3.00
HuBERT-L6	100	<b>273.86</b>	<b>31.36</b>	<b>5.54</b>	<b>251.2</b>	<b>33.67</b>	<b>5.88</b>	3.95	3.53
HuBERT-L6	200	289.36	33.04	7.49	262.4	34.30	6.13	<b>4.01</b>	<b>4.32</b>
wav2vec-L14	50	936.97	-	307.91	1106.3	-	330.8	2.26	1.91
wav2vec-L14	100	948.96	79.51	208.38	775.1	-	205.7	2.28	1.92
wav2vec-L14	200	538.56	61.06	61.48	585.8	-	91.07	2.64	3.04



# Задача кодирования



# Задача кодирования

System	Metrics	S2u		uLM	
	Nb units	ABX with.↓	ABX acr.↓	spot-the-word↓	accept. judg.↓
<i>Toplines</i>					
ASR+LM		-	-	3.12	29.02
<i>Baselines</i>					
LogMel	50	23.95	35.86	48.52	46.78
LogMel	100	24.33	37.86	48.12	46.83
LogMel	200	25.71	39.65	49.62	47.76
<i>Unsupervised</i>					
CPC	50	5.50	7.20	32.18	45.43
CPC	100	<b>5.09</b>	<b>6.55</b>	31.72	44.35
CPC	200	5.18	6.83	37.40	45.19
HuBERT-L6	50	7.37	8.61	32.88	44.06
HuBERT-L6	100	6.00	7.41	<b>31.30</b>	<b>42.94</b>
HuBERT-L6	200	5.99	7.31	36.52	47.03
wav2vec-L14	50	22.30	24.56	51.92	45.75
wav2vec-L14	100	18.16	20.44	50.24	45.97
wav2vec-L14	200	16.59	18.69	44.68	45.70



# Пример

<https://speechbot.github.io/gslm/index.html>

