Lingua Cosmica

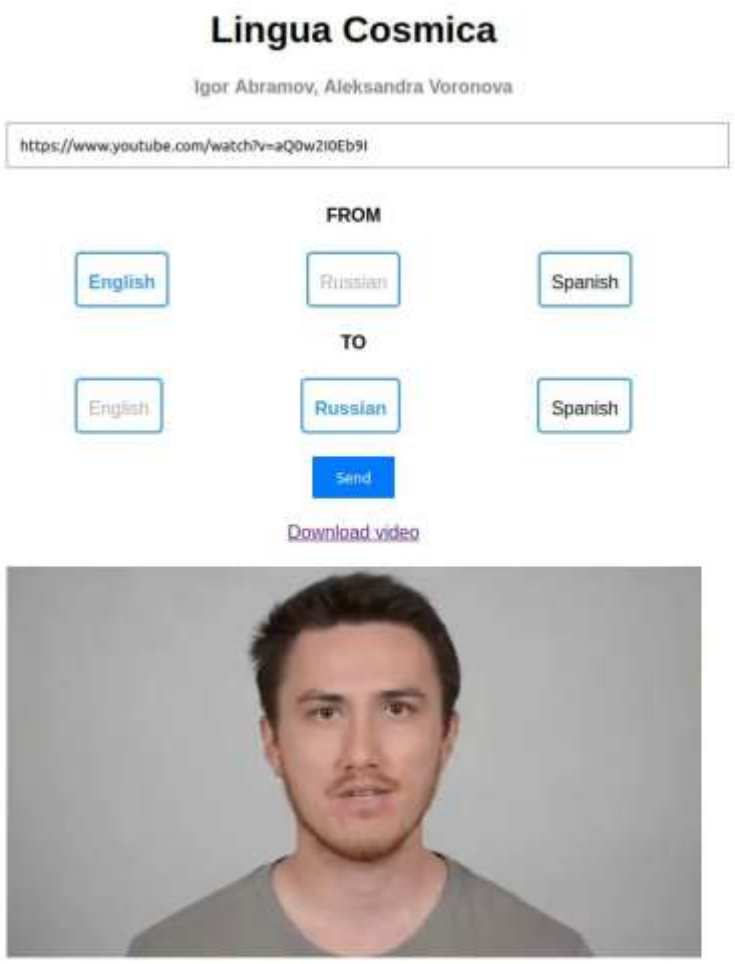
Speech-To-Speech translation service

Igor Abramov

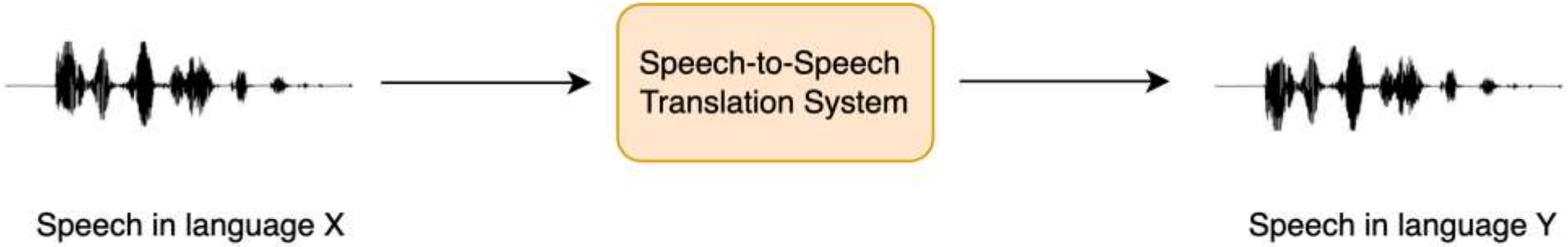
Aleksandra Voronova

I. Result

End-to-end video translation web application that incorporates speech-to-text and text-to-speech ML models.



II. Methodology



II. Methodology

Speech-to-Text Translation

- Download video and audio streams.
- Extract audio signals and divide them into chunks.
- Apply the Whisper model for translation

Text-to-Speech Synthesis

- Apply the MMSTTS model for audio synthesis on each chunk.
- Concatenate chunks and adjust resulting audio to match the original video length.
- Return the video with the translated audio track to the user.

III. Related Work

- Whisper: Unsupervised Speech Recognition
- F*ceb**k MMSTTS: Massively Multilingual Speech Technology

III. Related Work

Speech-To-Text Translation with Whisper:

- Model is not trained on translation directly;
- Trick the model into translating voice.
  (it did not go well)

Multitask training data (680k hours)

English transcription

🗣 "Ask not what your country can do for ···"

📝 Ask not what your country can do for ···

Any-to-English speech translation

🗣 "El rápido zorro marrón salta sobre ···"

📝 The quick brown fox jumps over ···

Non-English transcription

🗣 "언덕 위에 올라 내려다보면 너무나 넓고 넓은 ···"

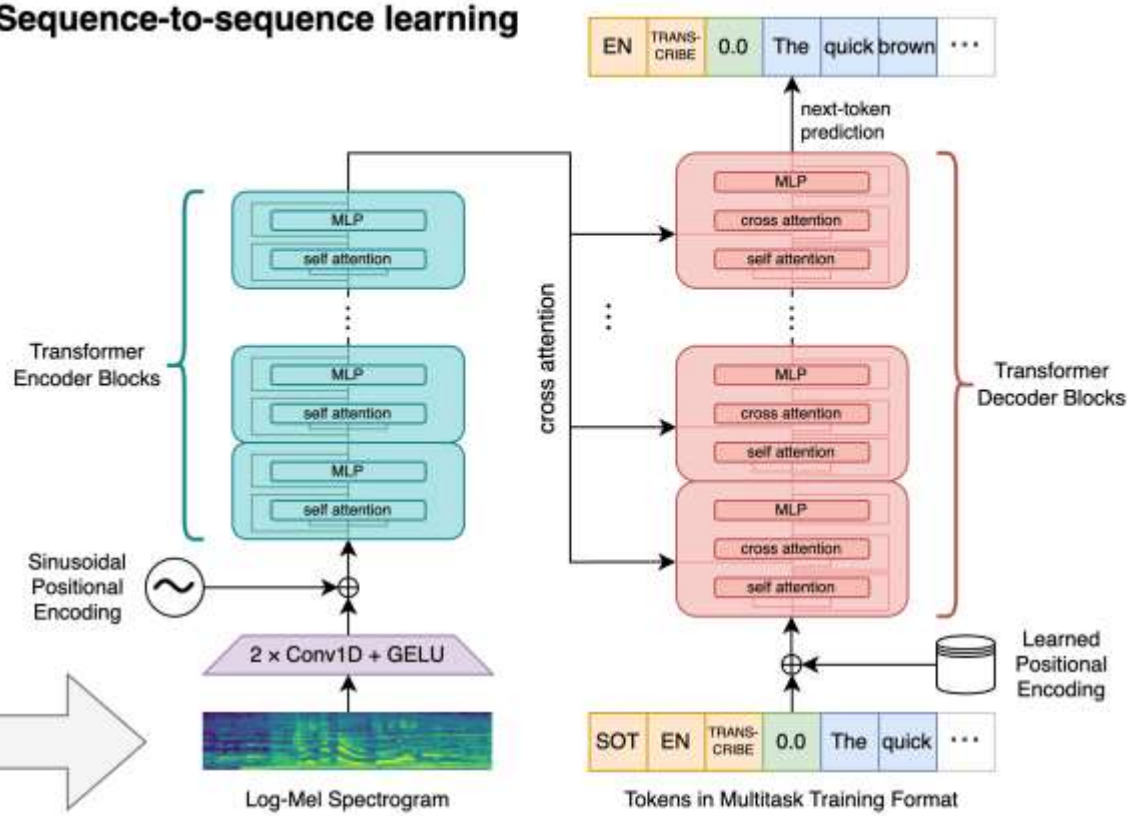
📝 언덕 위에 올라 내려다보면 너무나 넓고 넓은 ···

No speech

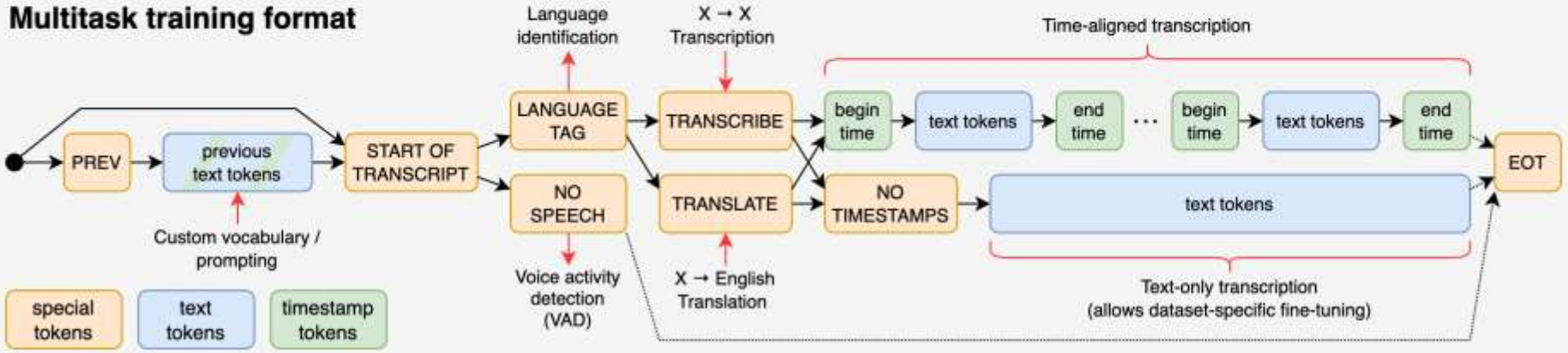
🔊 (background music playing)

📝 ∅

Sequence-to-sequence learning



Multitask training format

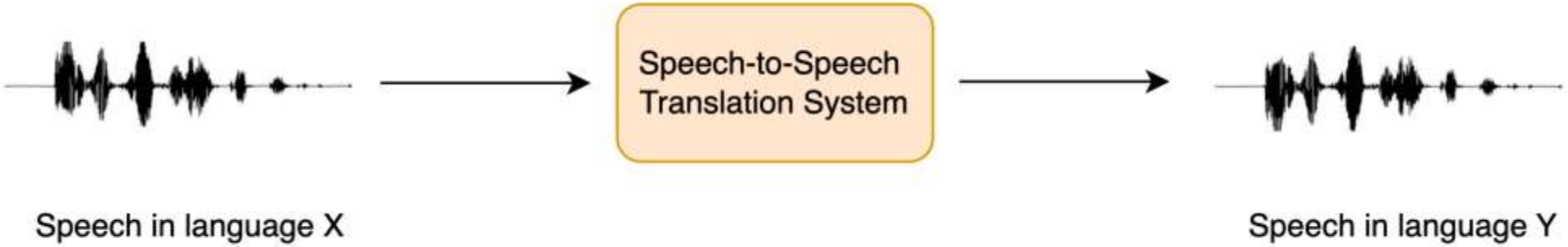


III. Related Work

Text-To-Speech Synthesis
MMSTTS:

- F*ceb**k research create huge model zoo trained for 1144 languages;
- Take specific TTS model for output language.

How to evaluate this thing?



Speech-to-Speech Translation

24 papers with code • 1 benchmarks • 5 datasets

Speech-to-speech translation (S2ST) consists on translating speech from one language to speech in another language. This can be done with a cascade of automatic speech recognition (ASR), text-to-text machine translation (MT), and text-to-speech (TTS) synthesis sub-systems, which is text-centric. Recently, works on S2ST without relying on intermediate text representation is emerging.

Benchmarks

Add a Result

These leaderboards are used to track progress in Speech-to-Speech Translation

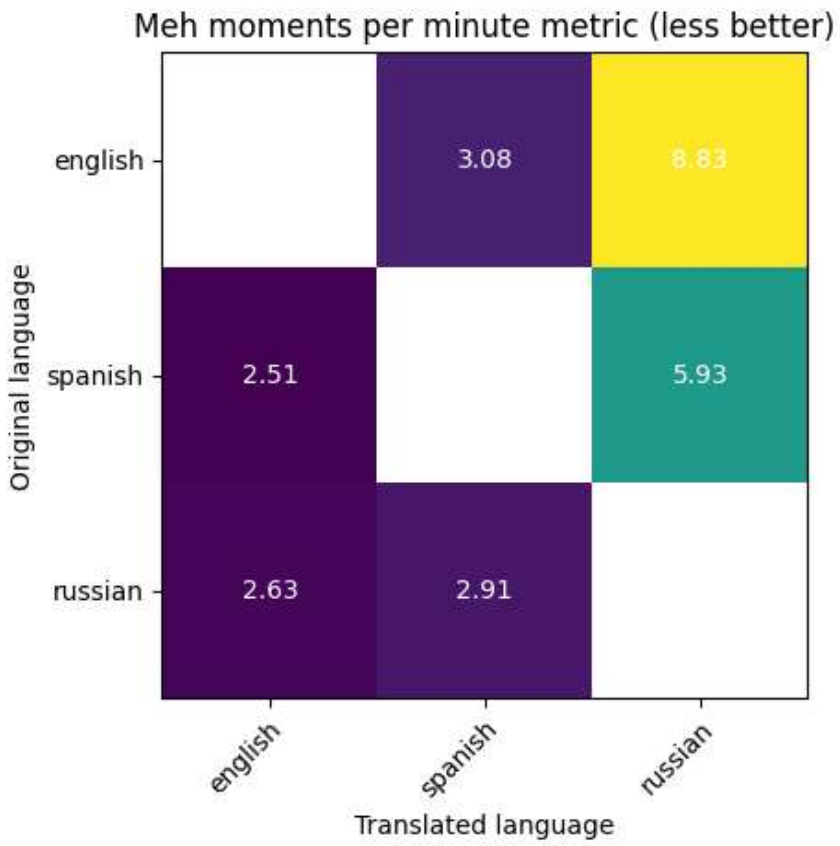
Trend	Dataset	Best Model	Paper	Code	Compare
	TAT	Hokkien→En (Two-pass decoding)			See all

Human evaluation!

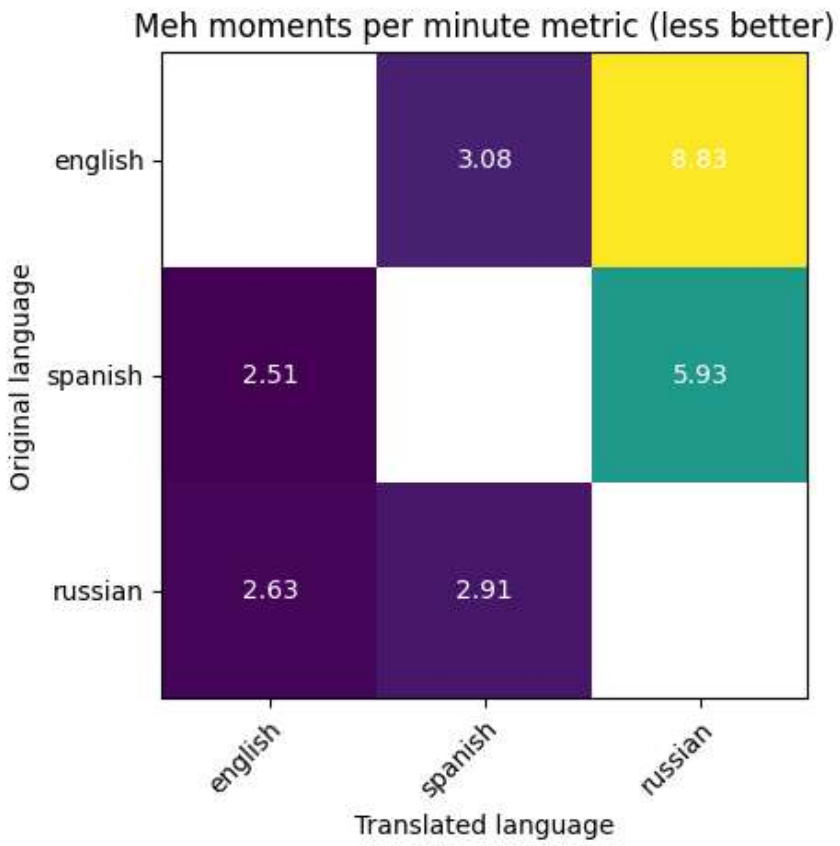
Introducing: meh moments per minute

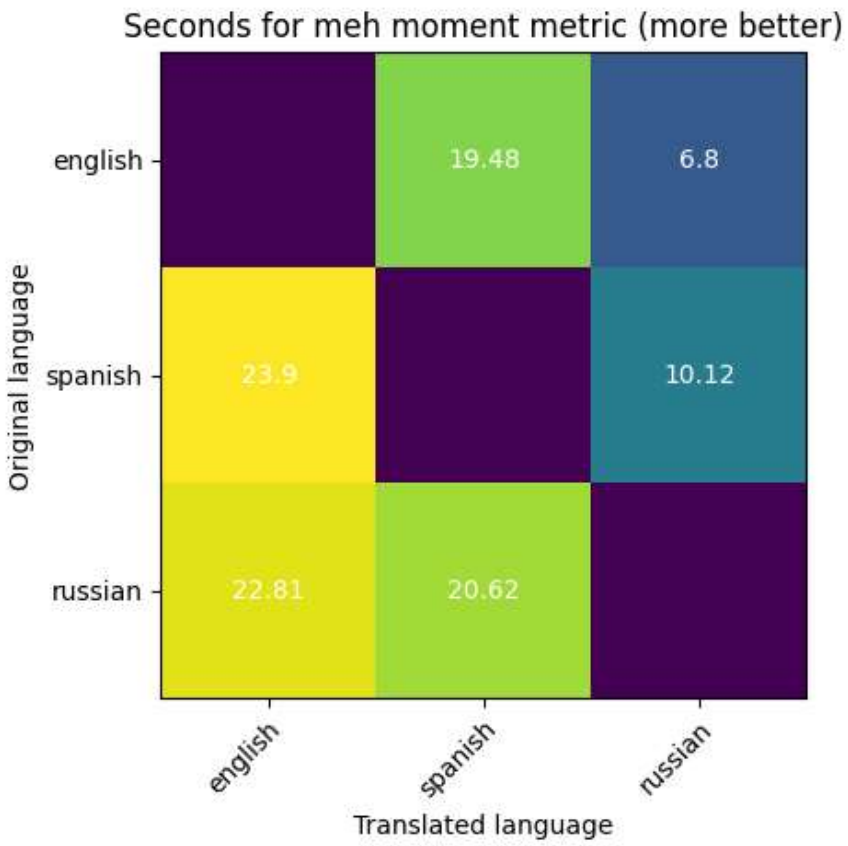
A	B	C	D	E	F	G	H	I
video_name	original_lang	translated_to	length	meh_moments	whisper_size	split_sentences	meh_moments_per_minute	
MinecraftSpanis	english	spanish	50	2	small	0	2.40	
MinecraftSpanis	english	russian	120	16	medium	1	8.00	
ListeningEnglish	english	spanish	120	8	medium	1	4.00	
MinutePhysicsR	russian	english	121	5	medium	1	2.48	
PlanningRussiar	russian	spanish	270	18	medium	1	4.00	
PlanningRussiar	russian	english	270	12	medium	1	2.67	
ParkinsonRussia	russian	english	197	9	medium	1	2.74	
ParkinsonRussia	russian	spanish	197	6	medium	1	1.83	
MinutePhysicsEı	english	spanish	247	11	medium	1	2.67	
MinutePhysicsEı	english	russian	80	15	medium	1	11.25	
MinutePhysicsEı	english	spanish	176	8	medium	1	2.73	
MinutePhysicsEı	english	spanish	149	9	medium	1	3.62	
MinutePhysicsEı	english	russian	174	21	medium	1	7.24	
StoriesSpanish	spanish	russian	103	13	large	1	7.57	
StoriesSpanish	spanish	english	135	5	medium	1	2.22	
SpeechTipsSpar	spanish	english	120	7	medium	1	3.50	
SpeechTipsSpar	spanish	russian	120	12	medium	1	6.00	
MarquezSpanisl	spanish	russian	157	11	large	1	4.20	
MarquezSpanisl	spanish	english	166	5	medium	1	1.81	

Pretty plots

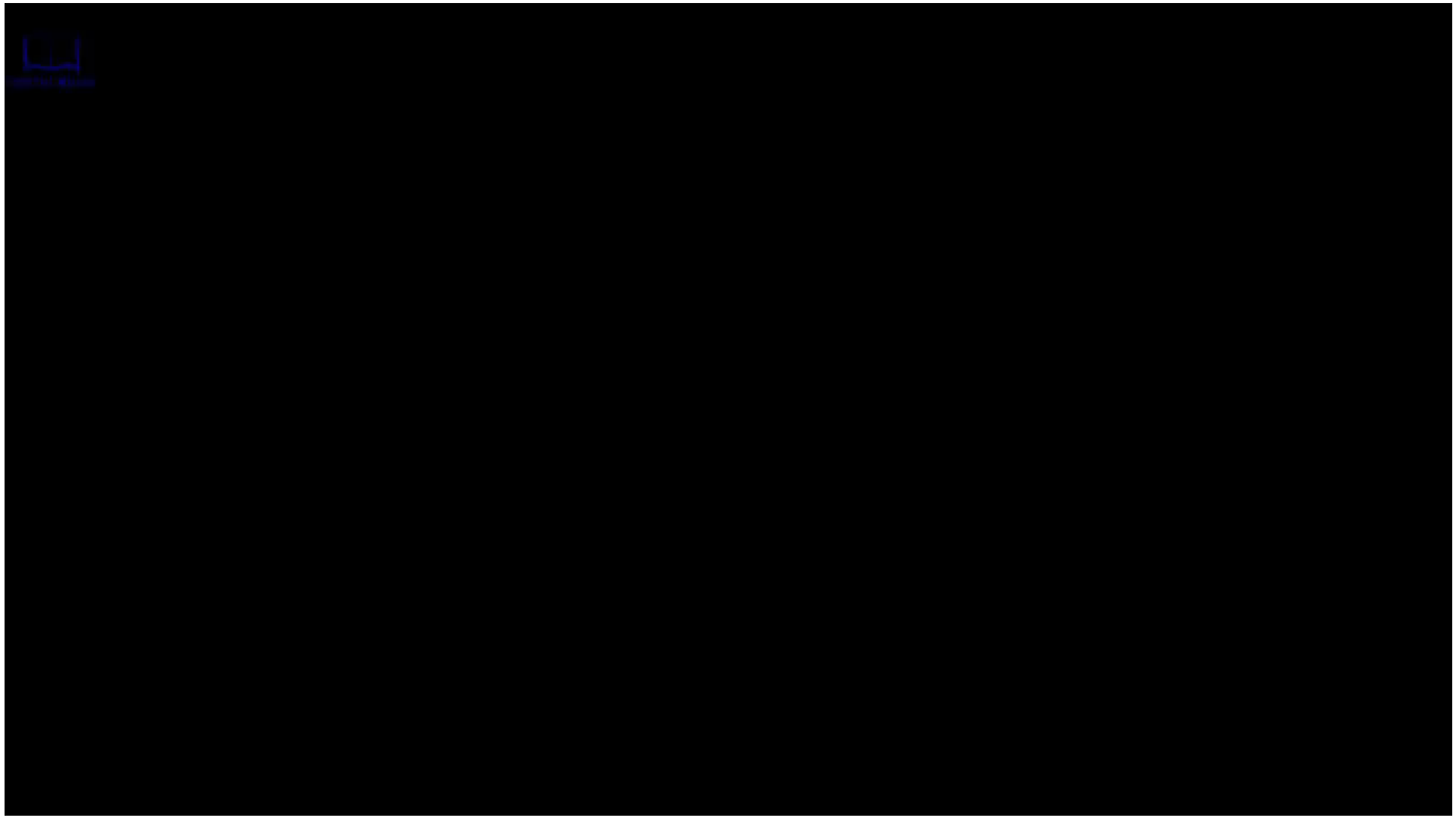


Pretty plots





Small demo (link if no video)

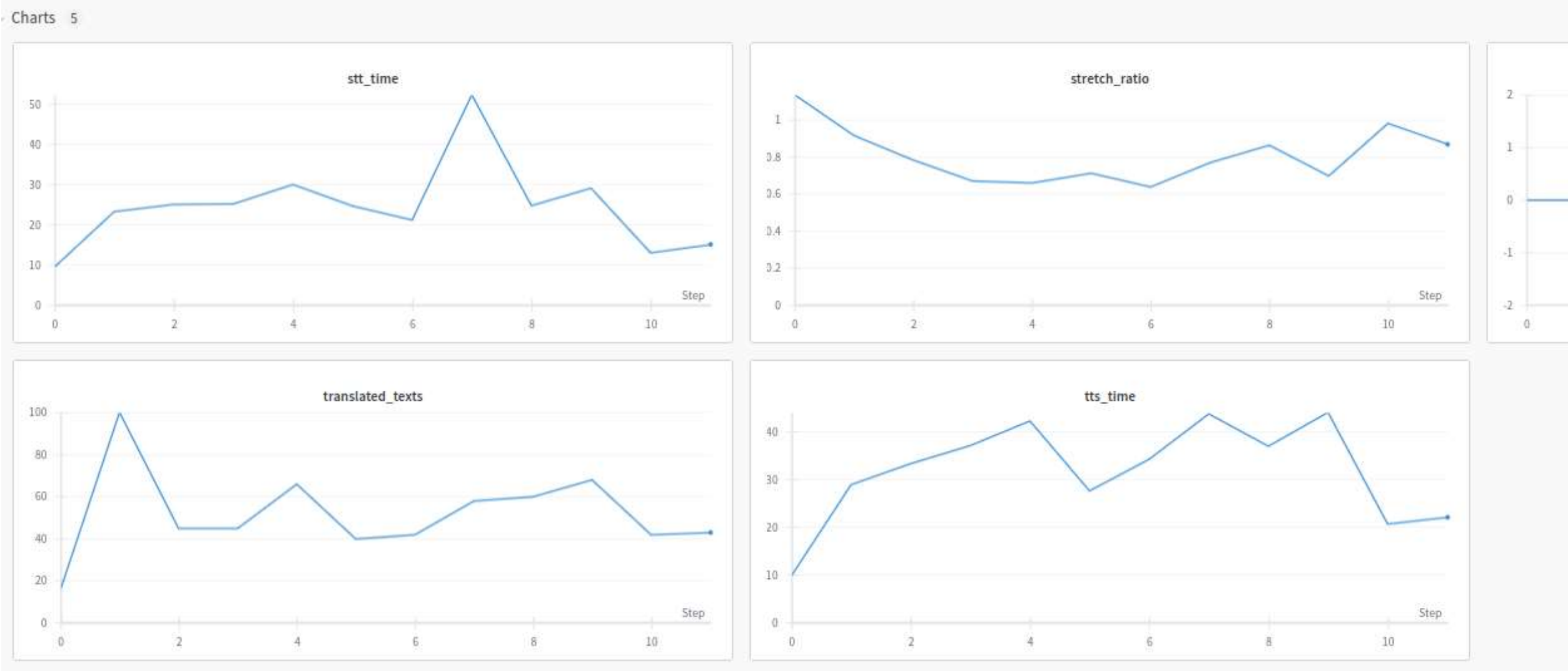


Q&A

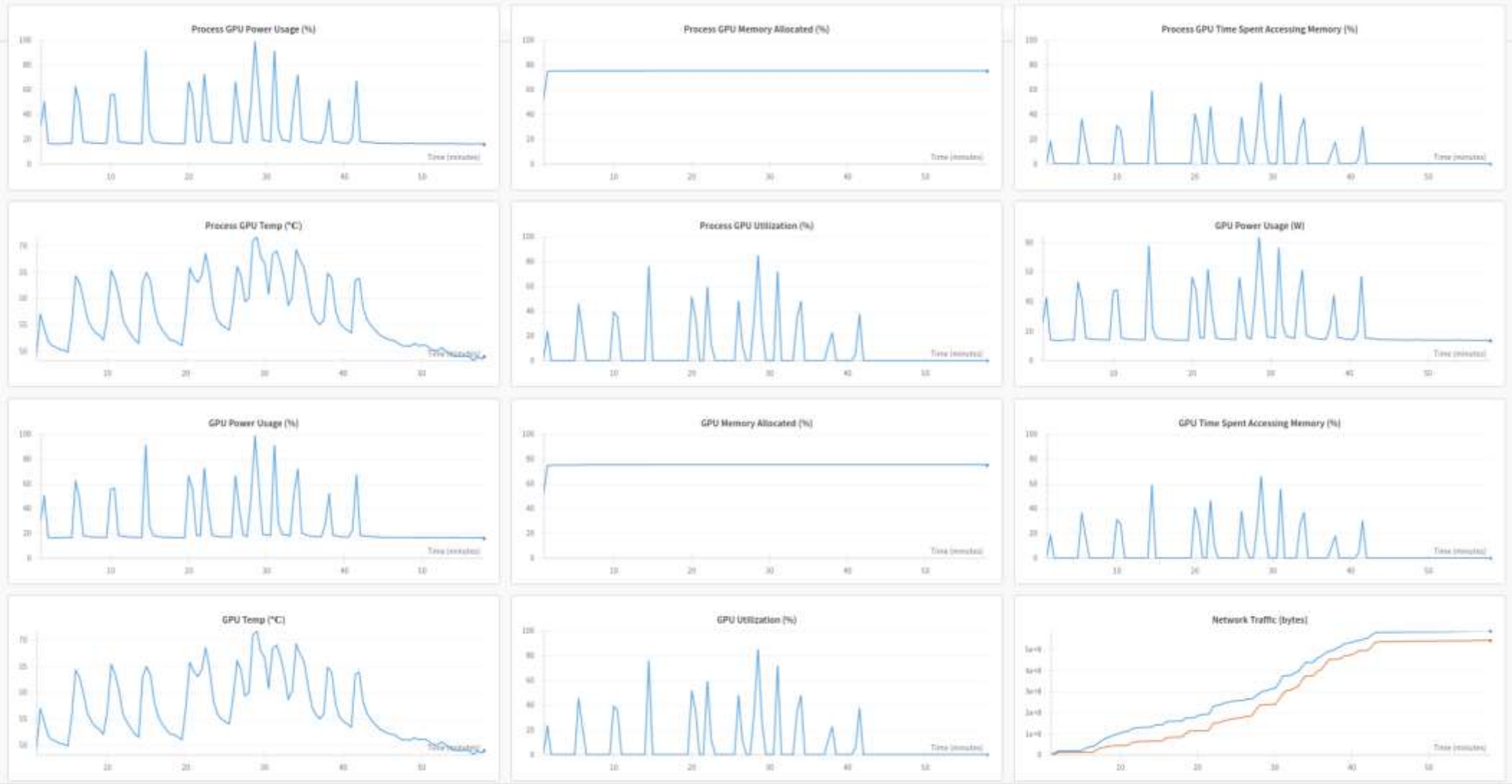
References

- Hugging Face, STS Translation
- A. Radford et al., Robust Speech Recognition via Large-Scale Weak Supervision
- V. Pratap et al., Scaling Speech Technology to 1,000+ Languages

Deployment


Charts 5
stt_time
Step
stretch_ratio
Step
translated_texts
Step
tts_time
Step

Deployment


Process GPU Power Usage (%)
Process GPU Memory Allocated (%)
Process GPU Time Spent Accessing Memory (%)
Process GPU Temp (°C)
Process GPU Utilization (%)
GPU Power Usage (W)
GPU Power Usage (%)
GPU Memory Allocated (%)
GPU Time Spent Accessing Memory (%)
GPU Temp (°C)
GPU Utilization (%)
Network Traffic (bytes)