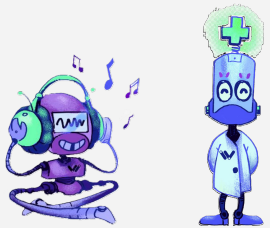
Stenograf

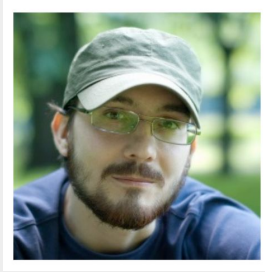
Neural self-supervised audio representation for SpeechLLM

Benchmarking neural audio codecs for Polish language



MLinPL 2025, Warszawa - 16.10. 2025

Who I am



Paweł Cyrta

👨‍🔬 Research Scientist
Speech 💥 & Music🎶 Technology
stenograf.io

Ph.d candidate at Cyfronet AGH

prev @SamsungPolska R&D, @IRCAM, @Rev.ai, @BabbleLabs, @Tooploox,
prev also at F.Chopin Warsaw Music University
weekend lecturer at Warsaw University of Technology,
postgraduate studies *Deep learning in digital media*

#ASR, #TTS, #AI #ML 🎙️🎮🎚️🎛️📣💬〰️📚📒📊📺📻

Pytania badawcze

W ramach badań chcę odpowiedzieć na pytania:

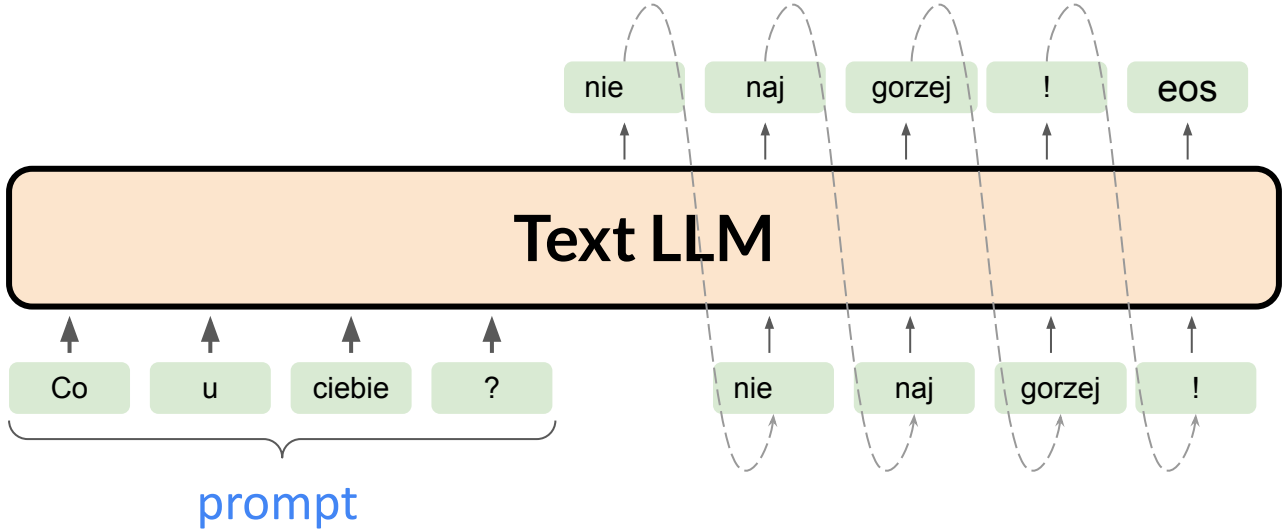
- Czy dostępne neuralne kodeki mowy
  dobrze reprezentują jęz. Polski

- Czy neuralne kodeki mowy są dopasowane
  do istniejących modeli LLM dla jęz. Polskiego np. Bielika

Two main research questions:

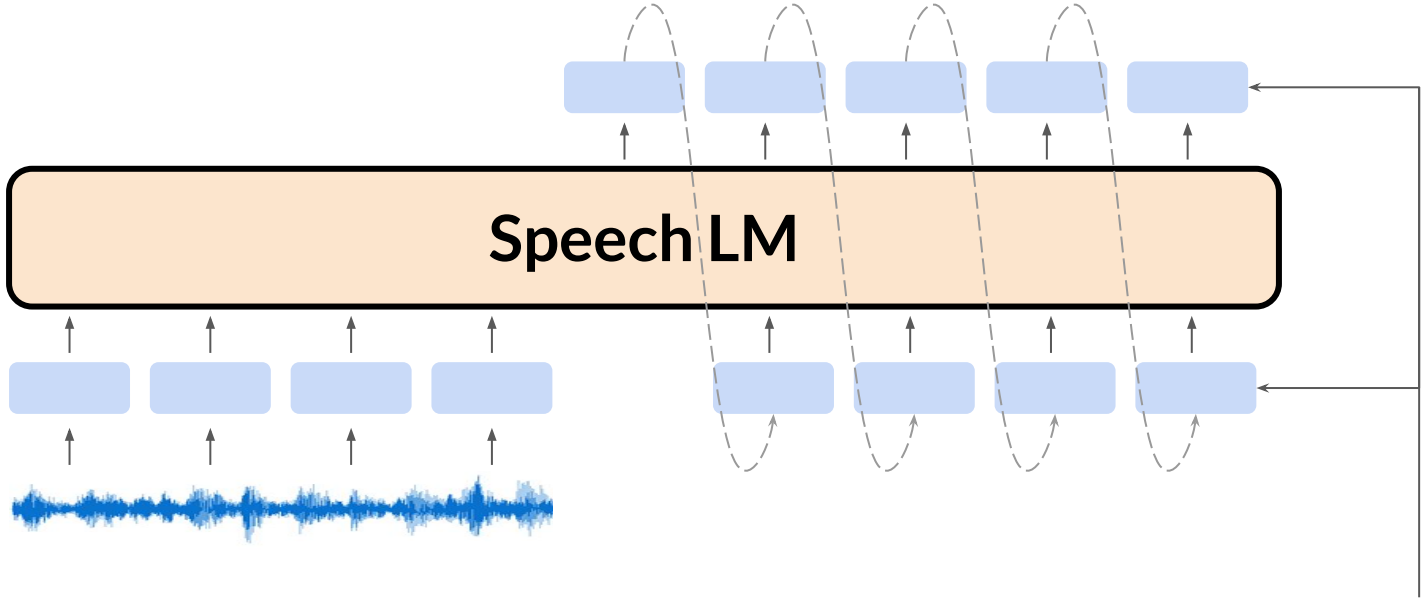
- Do available neural speech codecs
  effectively represent the Polish language?

- Are neural speech codecs well-suited
  for integration with existing Polish LLMs such as Bielik?



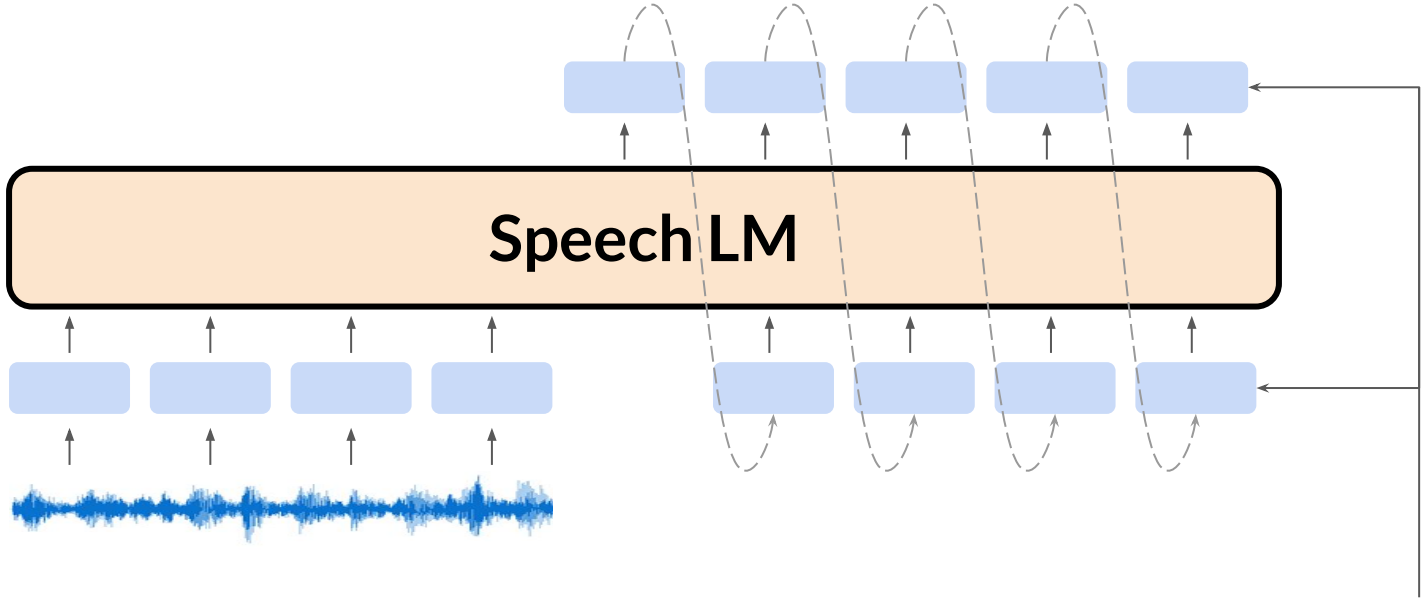
Text LLMs performs next-token-prediction.

What about in speech processing?



Similar to text LM, there are also some models trained on speech tokens

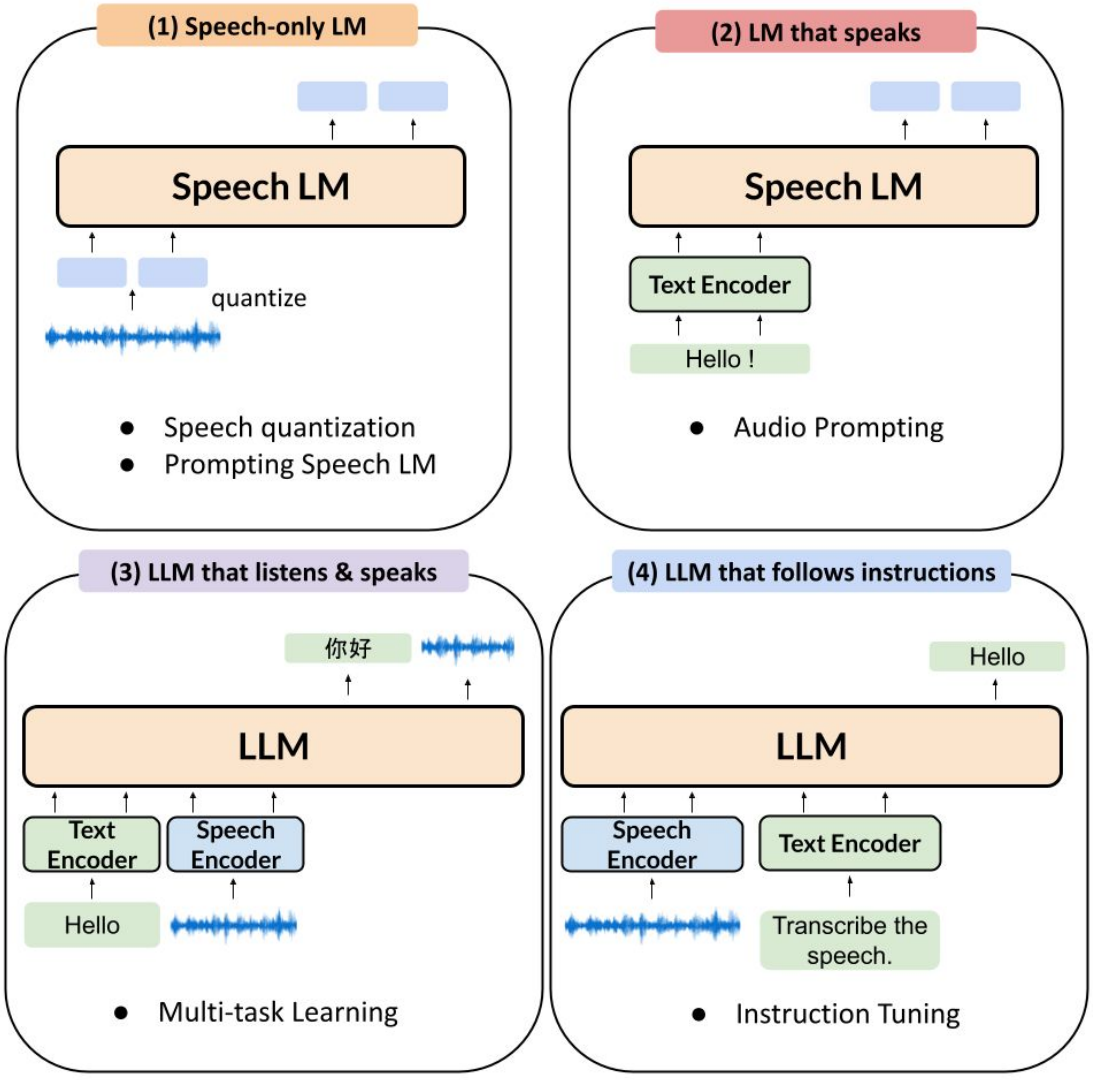
Discrete speech tokens



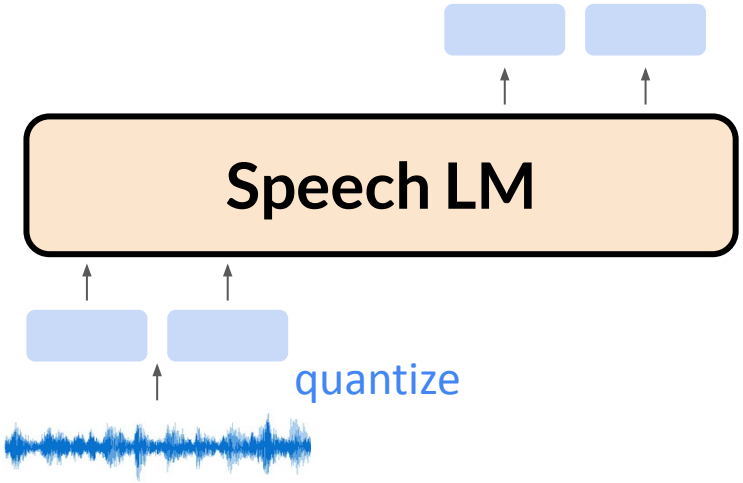
What are these tokens?

What do speech LM look like?

How to do speech LM for Polish lang.?


(1) Speech-only LM
Speech LM
quantize
Speech quantization
Prompting Speech LM
(2) LM that speaks
Speech LM
Text Encoder
Hello !
Audio Prompting
(3) LLM that listens & speaks
你好
LLM
Text Encoder
Speech Encoder
Hello
Multi-task Learning
(4) LLM that follows instructions
Hello
LLM
Speech Encoder
Text Encoder
Transcribe the speech.
Instruction Tuning

(1) Speech-only LM

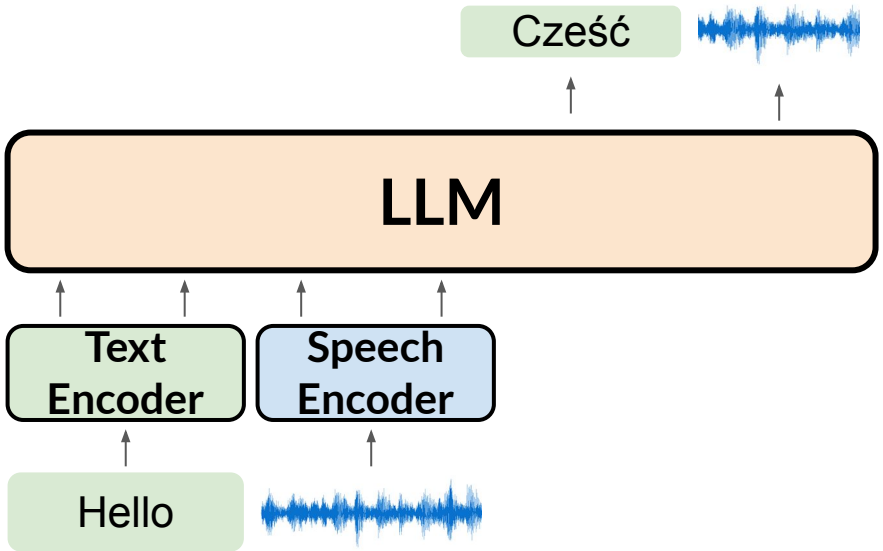


- Speech quantization
- Prompting Speech LM

(2) LM that speaks

- Audio Prompting

(3) LLM that listens & speaks



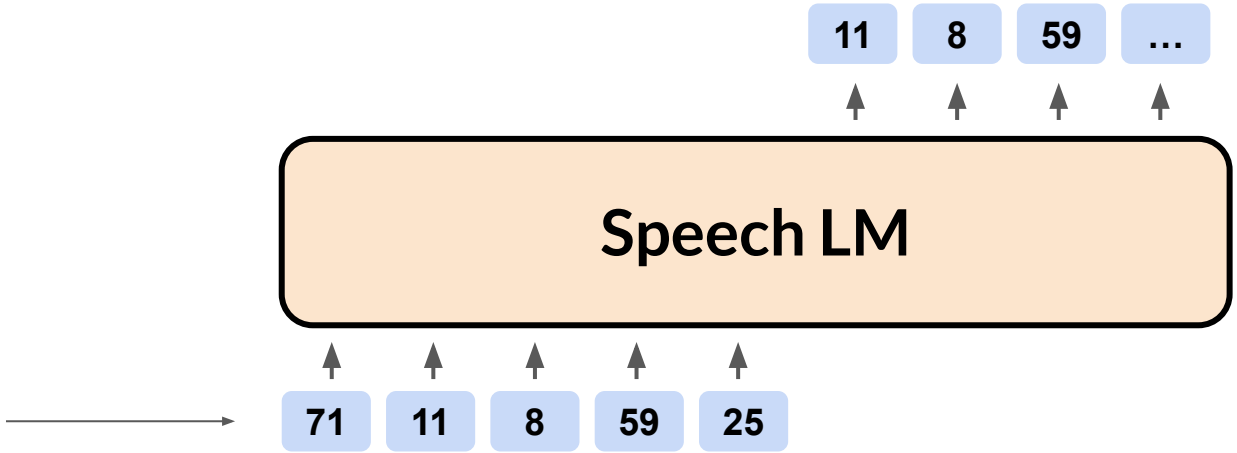
- Multi-task Learning

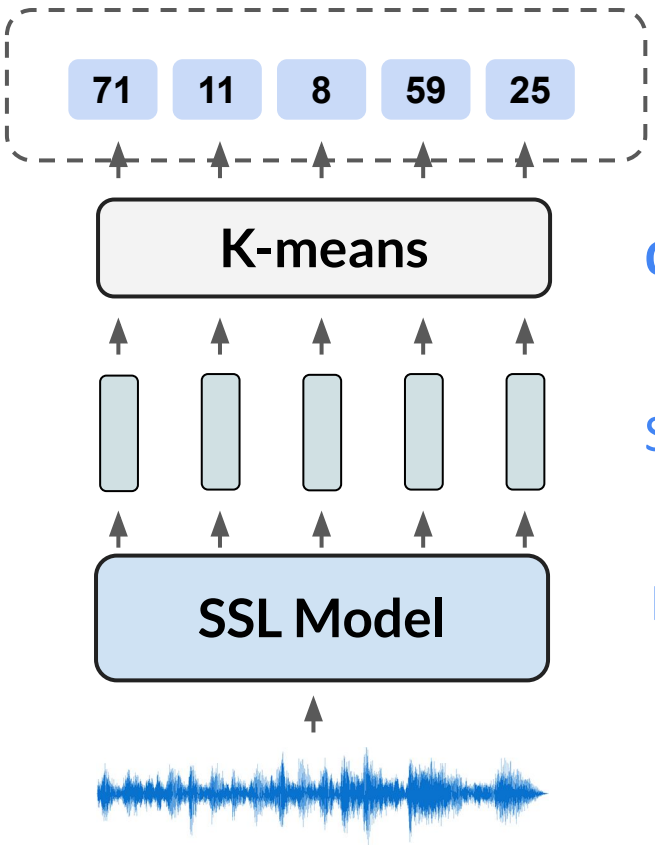
(4) LLM that follows instructions

- Instruction Tuning

What are these speech tokens?

Speech LM trained on speech tokens





Discrete speech tokens:

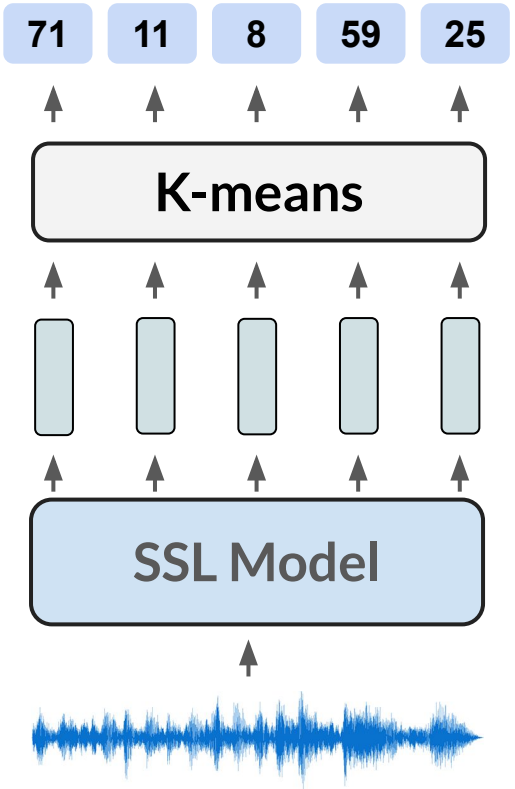
- pseudo text
- semantic tokens

Quantization

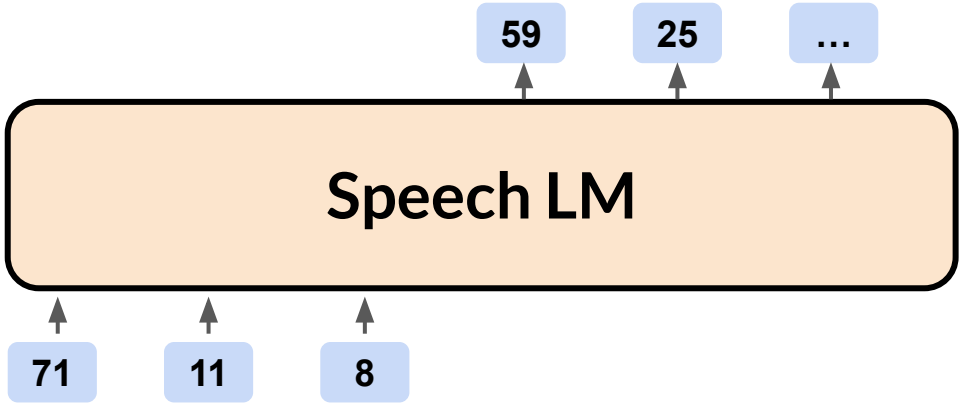
Speech representation

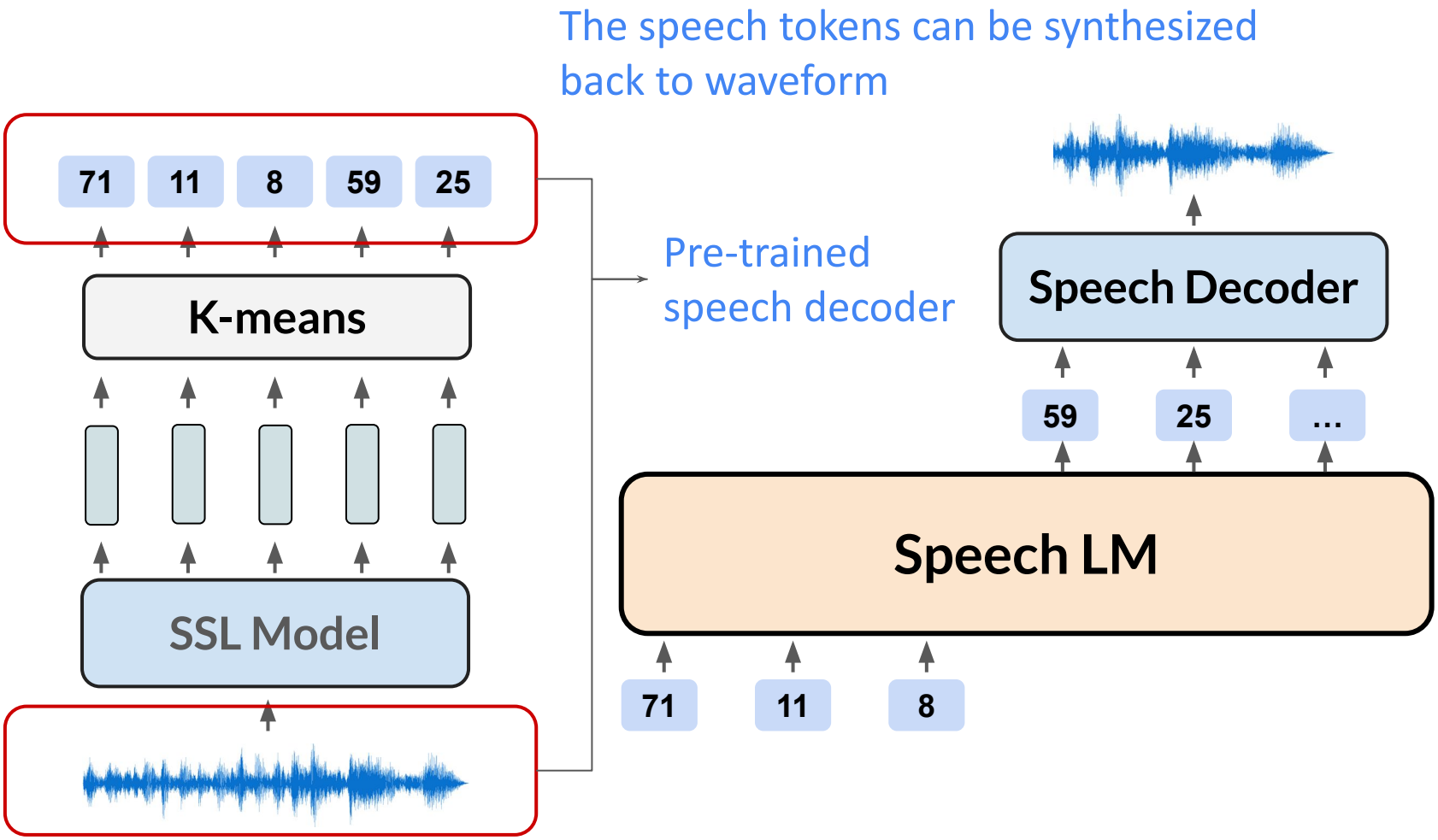
pre-trained SSL representation models

e.g. HuBERT, w2v-BERT, WavLM

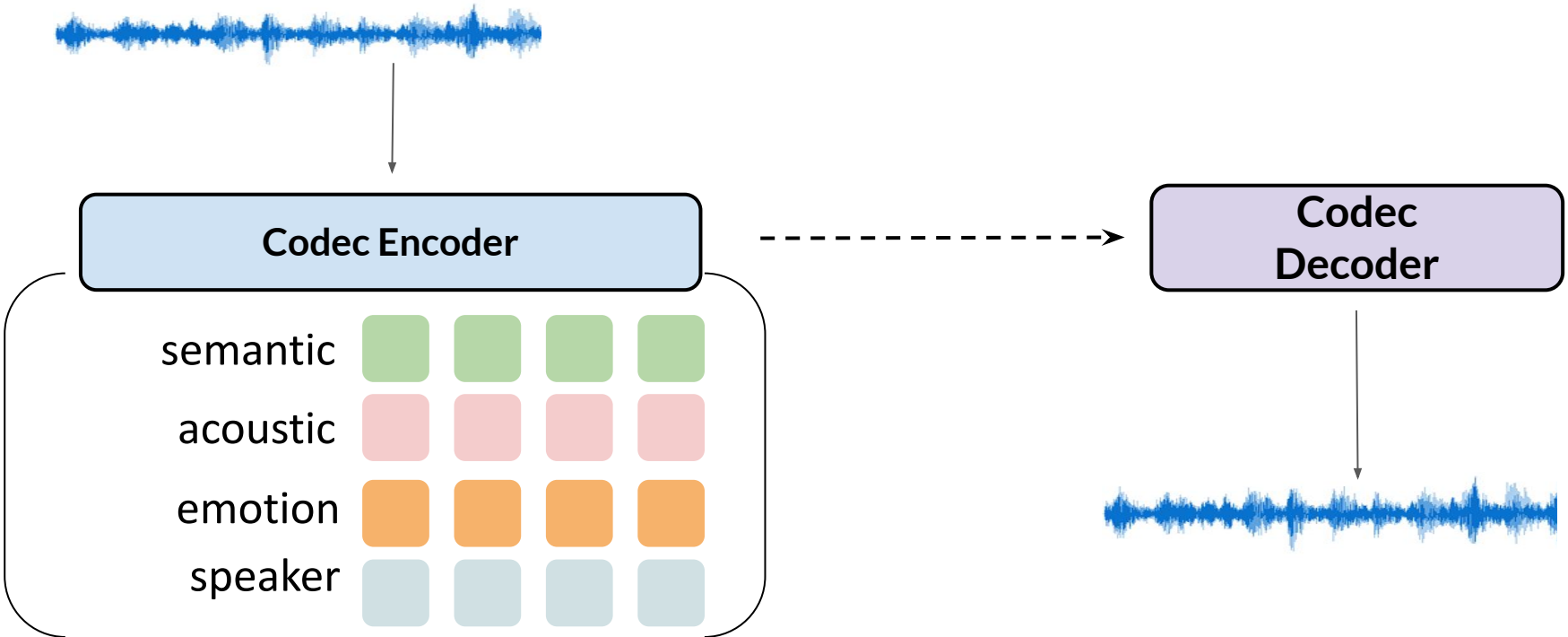


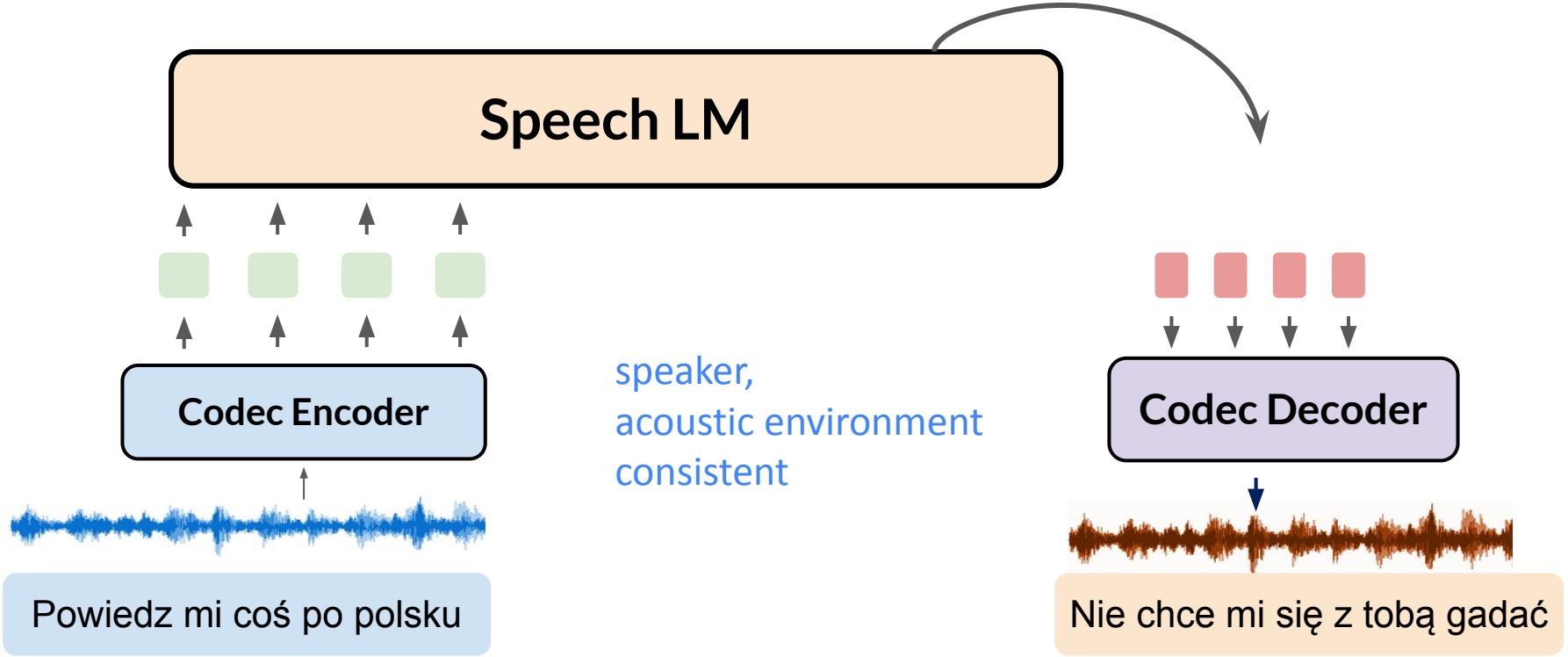
Speech LM performs next-token prediction on the speech tokens autoregressively




The speech tokens can be synthesized back to waveform
71
11
8
59
25
K-means
SSL Model
Pre-trained speech decoder
Speech Decoder
59
25
...
Speech LM
71
11
8

neural audio codec -
as self-supervisedly learned representation




Speech LM
Codec Encoder
Codec Decoder
speaker,
acoustic environment
consistent
Powiedz mi coś po polsku
Nie chce mi się z tobą gadać

Here, are list of representation to evaluate

self-supervisedly learned representation

wav2vec2.0, HuBERT, WavLM, w2v-bert-2.0

neural audio codecs:

EnCodec, DAC, SoundStream,

neural speech codecs for speechLLM

SpeechTokenizer, WavTokenizer, XCodec, FunCodec, Mimi

Dataset:

Bigos-v2 test set https://huggingface.co/datasets/amu-cai/pl-asr-bigos-v2

Neural Speech Representation Comparison

Ground Truth (GT) vs. Reconstructed (REC) • ↑ higher is better • ↓ lower is better

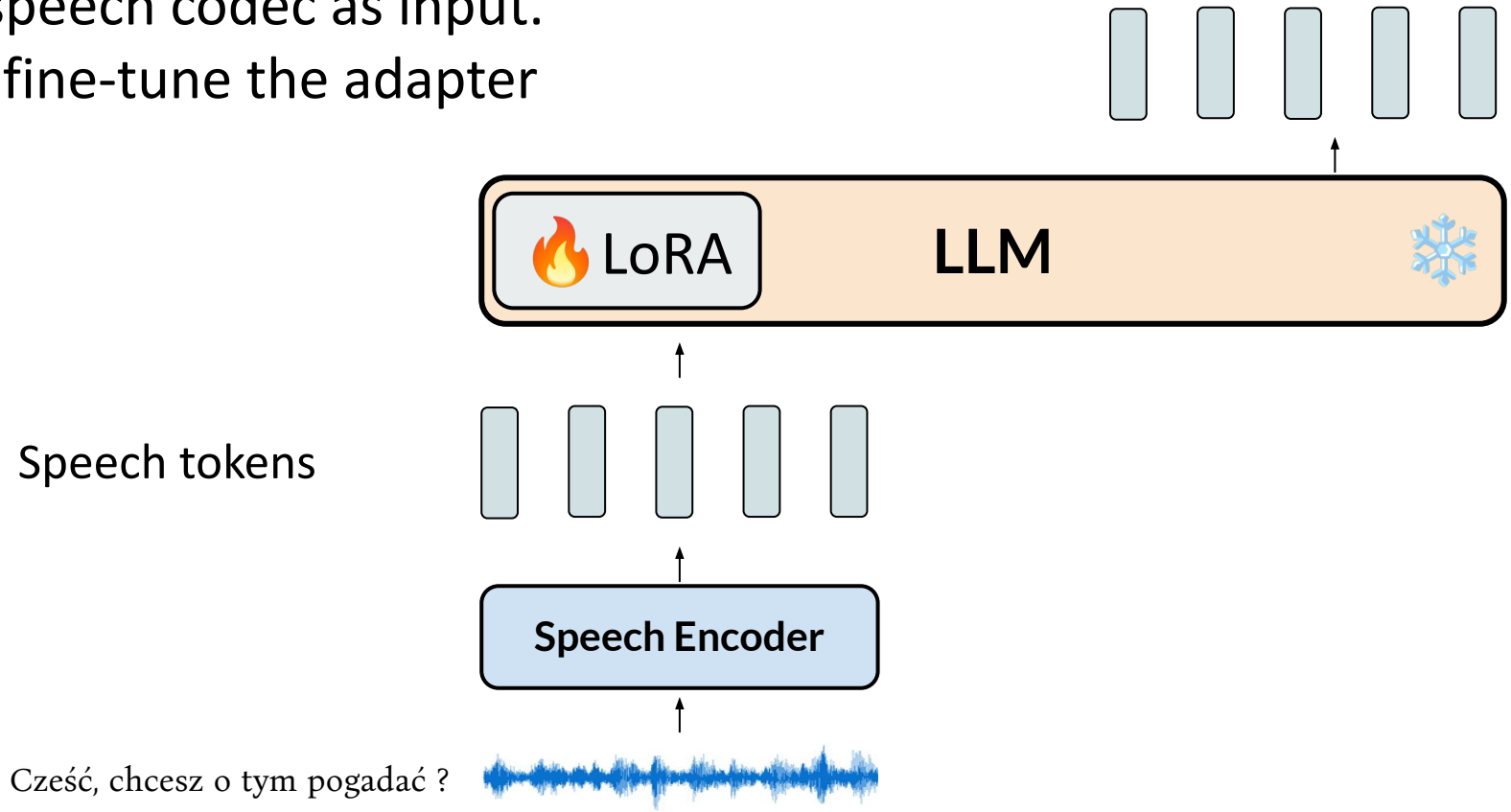
Codec	Category	PESQ ↑	Spk-Sim ↑	WER ↓	CER ↓	STOI ↑	Params	Bitrate
wav2vec 2.0	*SSL Model*	- / -	0.850 / -	**0.155 / 0.245**	**0.09 / 0.158**	- / -	95M	-
HuBERT	*SSL Model*	- / -	0.875 / -	**0.155 / 0.238**	**0.09 / 0.152**	- / -	95M	-
WavLM	*SSL Model*	- / -	0.892 / -	**0.155 / 0.225**	**0.09 / 0.145**	- / -	95M	-
w2v-BERT 2.0	*SSL Model*	- / -	0.865 / -	**0.155 / 0.242**	**0.09 / 0.155**	- / -	600M	-
DAC	*Neural Codec*	**3.69 / 2.66**	**0.965 / -**	**0.155 / 0.202**	**0.09 / 0.125**	**0.94 / 0.86**	73M	8 kbps
EnCodec	*Neural Codec*	3.21 / 2.27	0.919 / -	**0.155 / 0.198**	**0.09 / 0.114**	0.93 / 0.85	16M	6 kbps
SoundStream	*Neural Codec*	3.15 / 2.18	0.908 / -	**0.155 / 0.215**	**0.09 / 0.128**	0.92 / 0.83	30M	6 kbps
Mimi	*Neural Codec*	2.77 / -	0.928 / -	**0.155 / 0.287**	**0.09 / 0.173**	0.88 / -	95M	12.5 kbps
SpeechTokenizer	*Neural Codec*	2.97 / -	0.924 / -	**0.155 / 0.216**	**0.09 / 0.120**	0.89 / -	35M	1 kbps
WavTokenizer	*Neural Codec*	2.17 / 1.14	0.743 / -	**0.155 / 0.494**	**0.09 / 0.325**	0.83 / 0.49	48M	0.7 kbps
XCodec	*Neural Codec*	3.23 / 1.85	0.942 / -	**0.155 / 0.185**	**0.09 / 0.106**	0.91 / 0.76	52M	5 kbps
FunCodec	*Neural Codec*	3.08 / 2.05	0.915 / -	**0.155 / 0.225**	**0.09 / 0.135**	0.90 / 0.82	45M	4 kbps
SemantiCodec	*Semantic Codec*	2.64 / 1.32	0.907 / -	**0.155 / 0.318**	**0.09 / 0.195**	0.86 / 0.60	120M	0.5 kbps

Note: Values shown as GT/REC format. Best values highlighted with **bold underline**.

Metrics:

- **PESQ:** Perceptual Evaluation of Speech Quality (1–5 scale)
- **Spk-Sim:** Speaker Similarity (cosine similarity, 0–1)
- **WER/CER:** Word/Character Error Rate from ASR
- **STOI:** Short-Time Objective Intelligibility (0–1)

Use speech codec as input.
then fine-tune the adapter



SpeechLLM Perplexity on BIGOS Dataset

Evaluated on Bielik-based SpeechLLM after pretraining

Codec	Category	PPL ↓	Params	Bitrate
wav2vec 2.0	*SSL Model*	156 / 128	95M	-
HuBERT	*SSL Model*	142 / 115	95M	-
WavLM	*SSL Model*	138 / 109	95M	-
w2v-BERT 2.0	*SSL Model*	145 / 118	600M	-
DAC	*Neural Codec*	247 / 194	73M	8 kbps
EnCodec	*Neural Codec*	76 / 141	16M	6 kbps
SoundStream	*Neural Codec*	85 / 152	30M	6 kbps
Mimi	*Neural Codec*	127 / -	95M	12.5 kbps
SpeechTokenizer	*Neural Codec*	14 / -	35M	1 kbps
WavTokenizer	*Neural Codec*	105 / 38	48M	0.7 kbps
XCodec	*Neural Codec*	30 / 48	52M	5 kbps
FunCodec	*Neural Codec*	**12 / 18**	45M	4 kbps
SemantiCodec	*Semantic Codec*	8 / 16	120M	0.5 kbps

Note: Values shown as before/after format (initial/fine-tuned). Best value highlighted with **bold underline**.

PPL: Perplexity metric measuring prediction uncertainty (lower is better). Evaluated after pretraining on Polish conversational data.

Pytania ?
questions ?

Contact me:
pawel@cyrta.com