POLEVAL 2024

Task 3: Polish Automatic Speech Recognition Challenge

Augmenting Polish Automatic Speech Recognition System With Synthetic Data

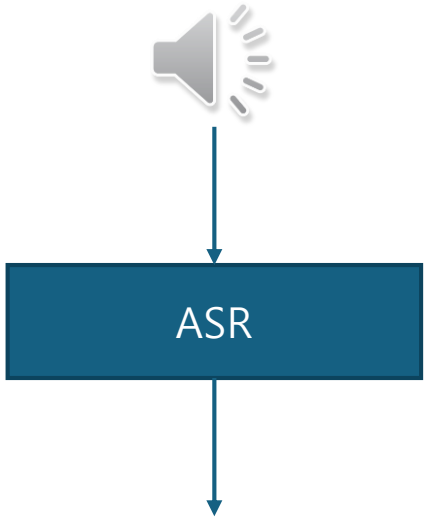
Łukasz Bondaruk, l.bondaruk@samsung.com

Jakub Kubiak, j.kubiak@samsung.com

Mateusz Czyżnikiewicz, m.czyznikiew@samsung.com

2 December 2024

Task & Data



Nieznajomy ocknął się z zamyślenia
i spojrzał nań przytomnie.

Dataset summary [1]

	Number of samples			Duration [h]		
Split	**BIGOS**	**PELCRA**	**Total**	**BIGOS**	**PELCRA**	**Total**
train	82025	229150	311175	236.70	432.26	668.96
dev-0	14254	28532	42786	27.51	49.60	77.11
test-A	1002	1167	2169	2.53	2.14	4.67
test-B	991	1178	2169	2.48	2.15	4.63

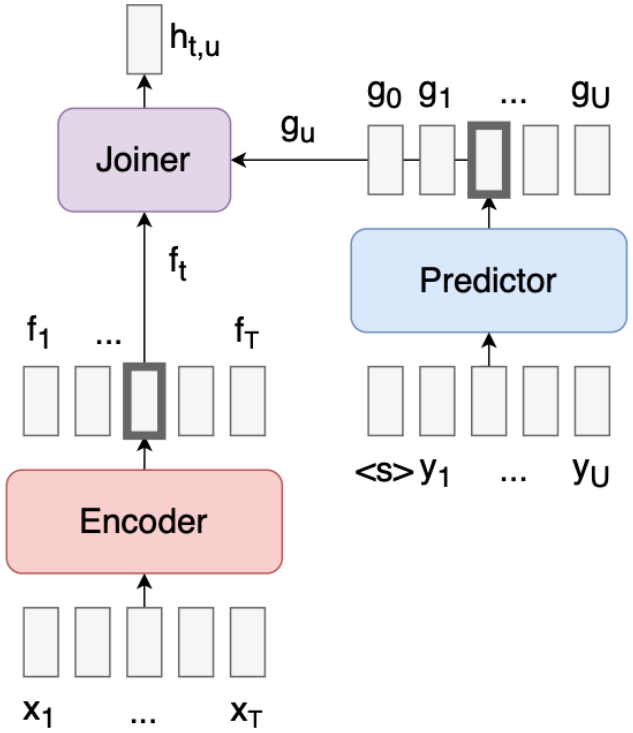
Difficult - many sources:

- PELCRA - spontaneous and conversational
- BIGOS - audiobooks, read speech, many devices, multiple acoustic conditions

Small - only ~700h

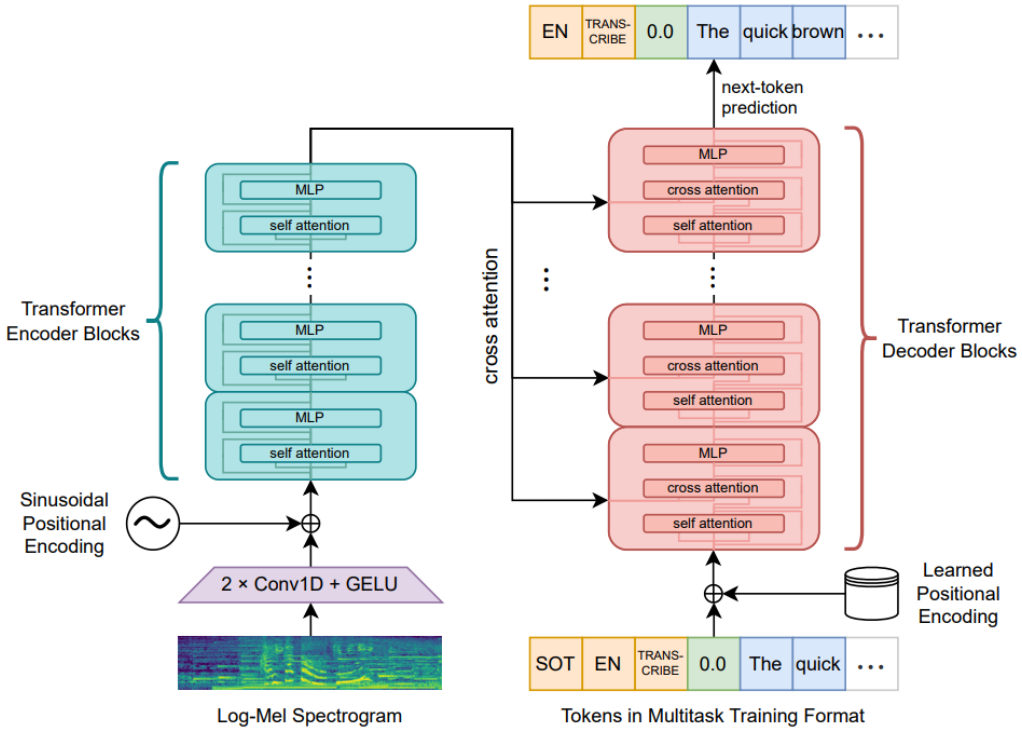
Speech Recognition - Conformer & Whisper

Conformer-based [4] RNN-Transducer:
- lighweight - 60M parameters
- trained from scratch



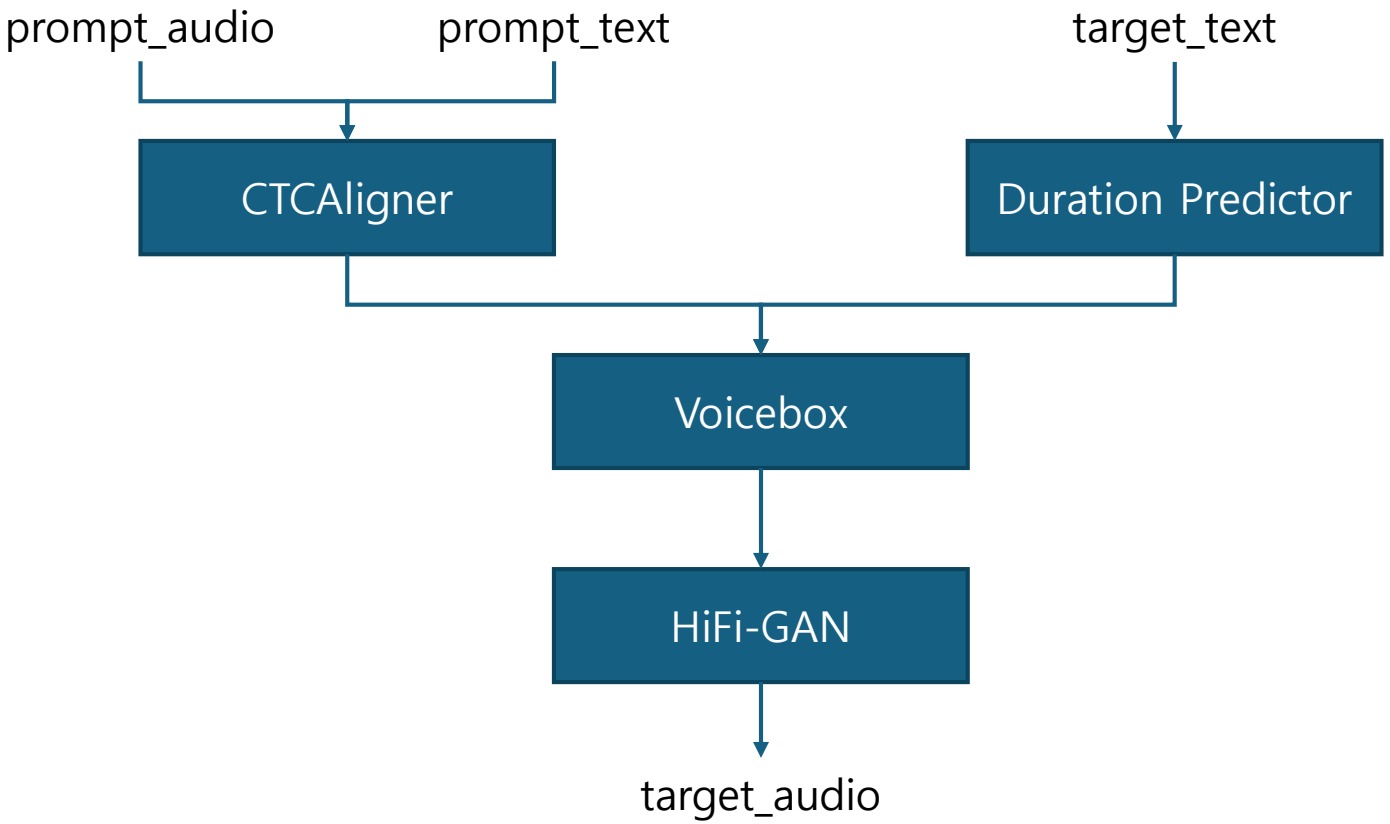
RNN-T model architecture [2]

Whisper [3]:
- large - 1550M parameters
- pretrained on massive corpus and finetuned

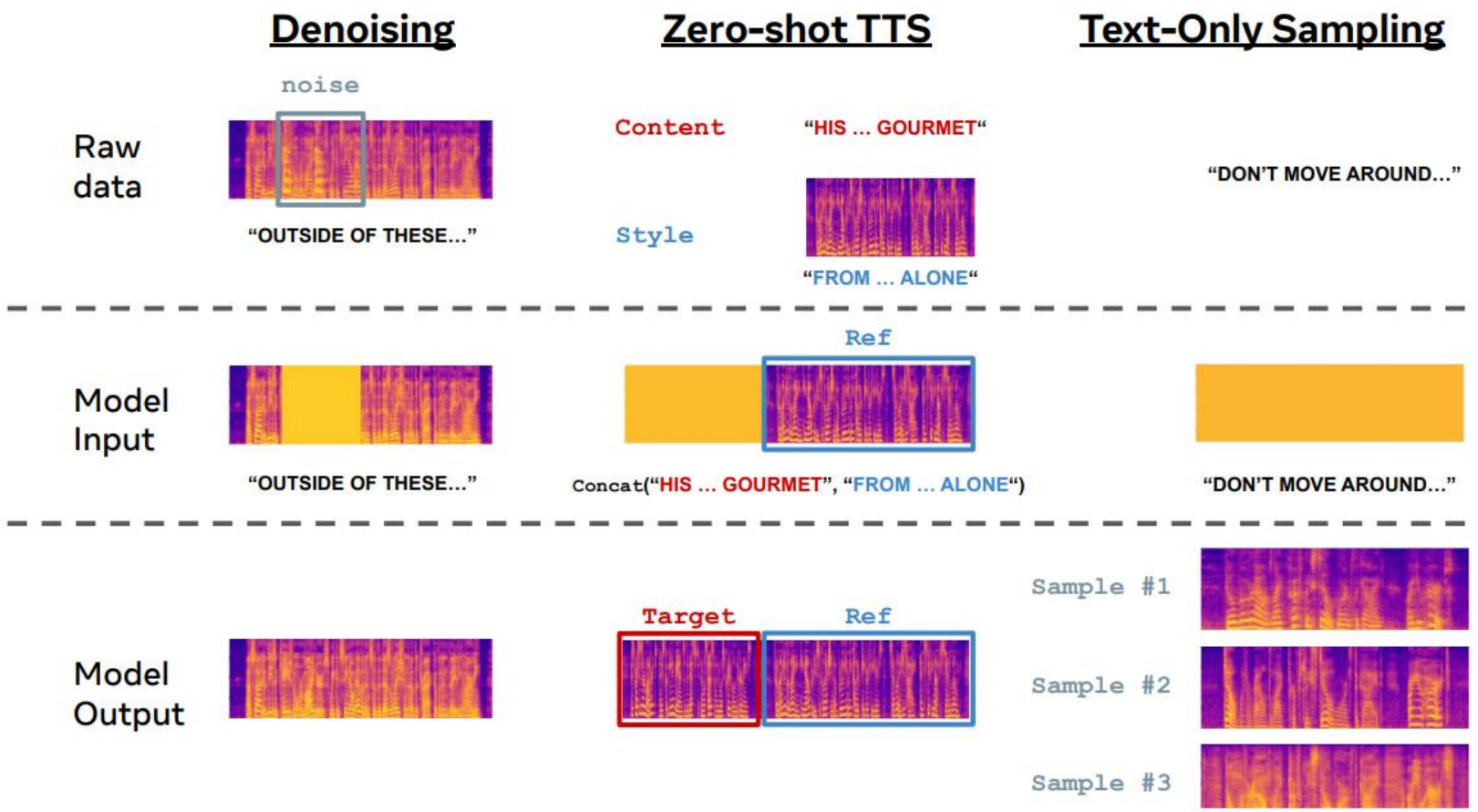


Whisper architecture [3]

Speech Synthesis

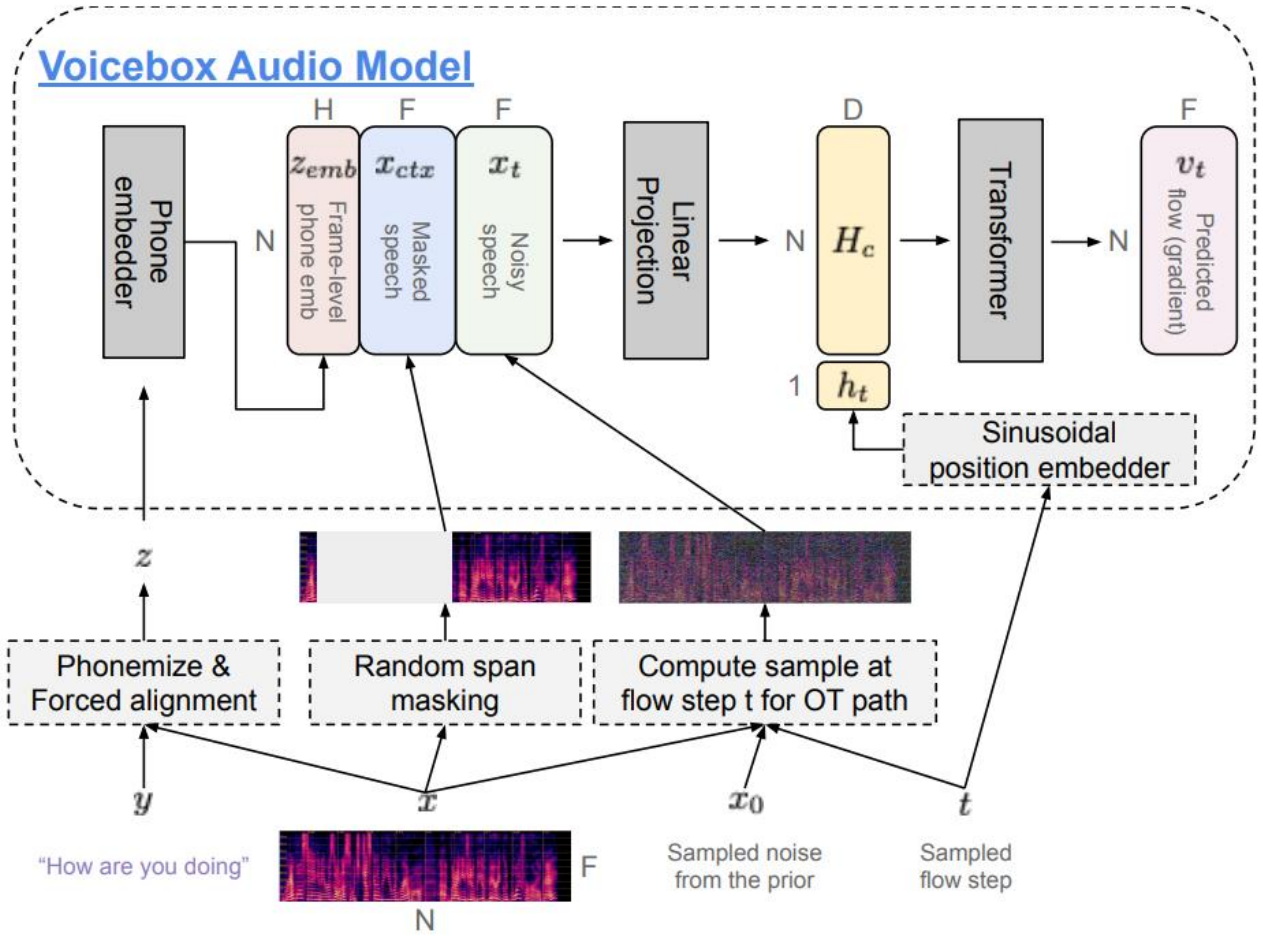


Speech Synthesis - Voicebox



Voicebox task generalization [5]

Speech Synthesis - Voicebox (Conditional Flow Matching)



Voicebox architecture [5]

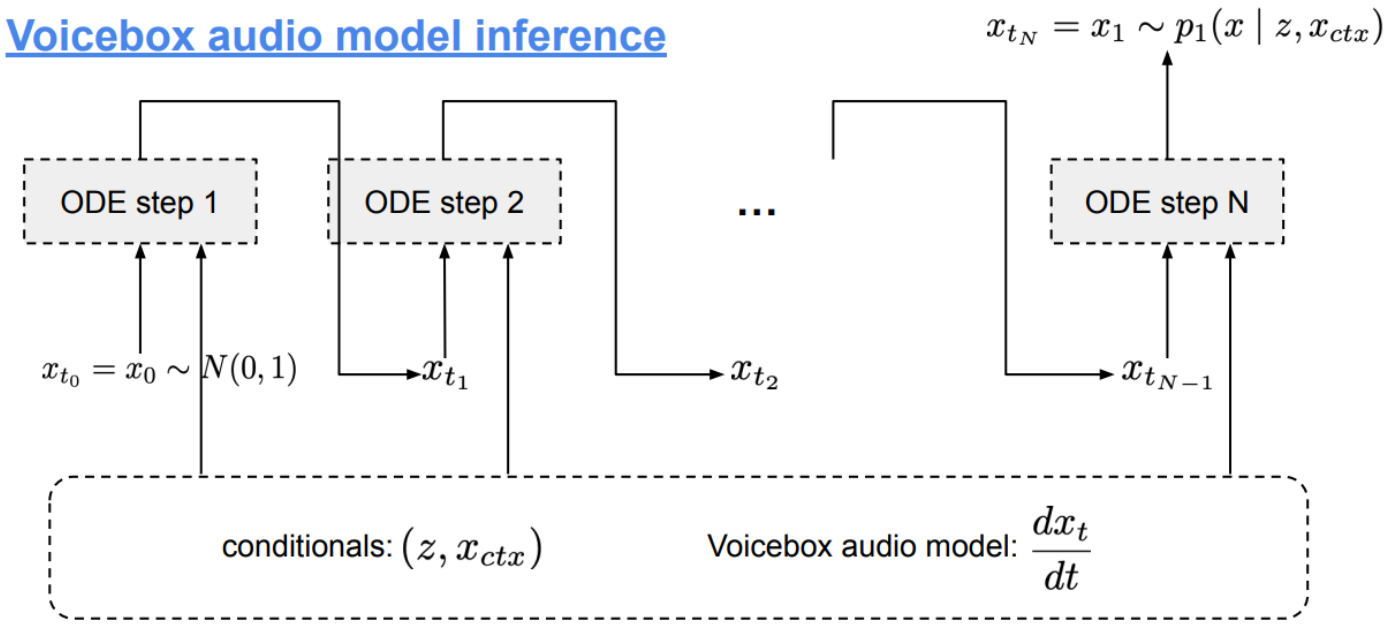
Training procedure for given sample of text and melspec (y, x):

1. Preprocess text and forced alignment to melspec
2. Mask span of melspec (context modelling)
3. Sample at flow step $t \in [0,1]$:
   $x_t = (1-t)x_0 + tx$,
   where $x_0 \sim \mathcal{N}(0,1)$
4. Calculate target for the model:

$$v_t = \frac{dx_t}{dt} = -x_0 + x$$

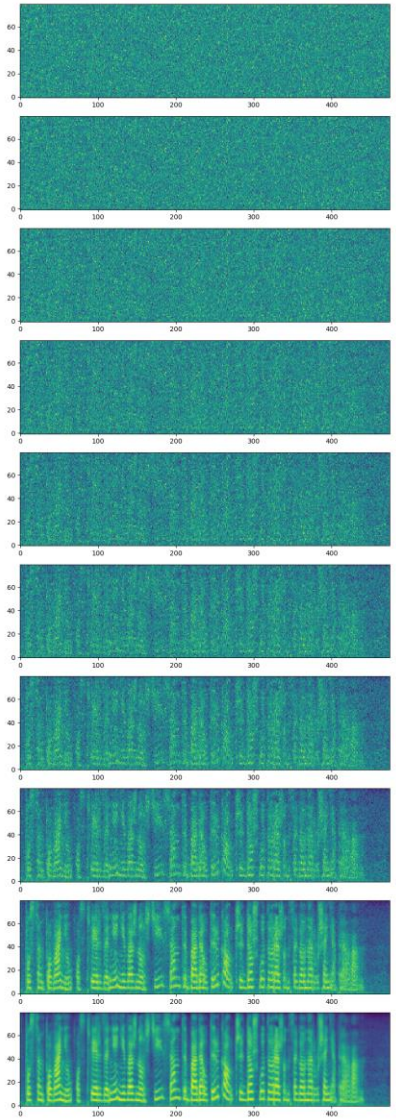
5. Calculate loss only for masked span of x

Speech Synthesis - Voicebox (Inference)

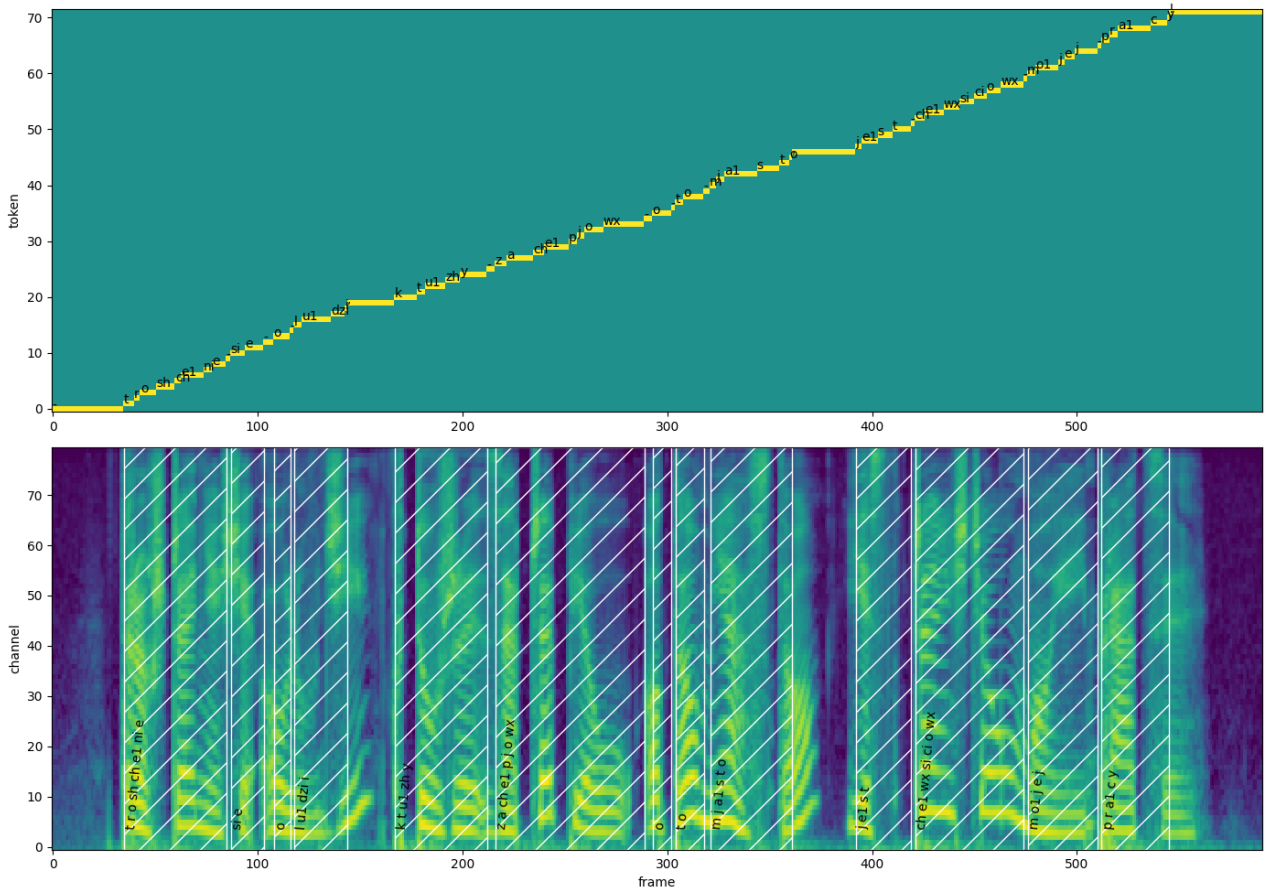


Voicebox inference as solving an ODE with initial condition x_o sampled from prior, derivative $\frac{dx_t}{dt}$ specified by the model, and conditional inputs (z, x_{ctx}). [4]

Selecting number of steps allows for trade-off between speed and quality. Usually even after 15 steps quality is very good.



Speech Synthesis - Forced Alignment



Audio:

Troszczenie się o ludzi, którzy zaczepią o to miasto jest częścią mojej pracy. 6.29[s]

Segments:

troszczenie 0.37[s] - 0.91[s]

ludzi 1.26[s] - 1.53[s]

miasto 3.42[s] - 3.85[s]

Synthetic Data

Prompt	Synthesis
	
	
	

Utilized (recorded+synthethic) datasets summary [1]

Dataset	Composition	Number of samples	Duration [h]
baseline	*train*	311175	669
mix-00	*train + synth-00*	604671	1109
mix-01	*train + synth-00 + synth-01*	1191663	1999

Prompts for synthesis were selected randomly from audio files that:

- achieved CER of at most 25%
- had a speech rate variation of up to 2.5 standard deviations from mean

Results

Results on dev split of data [1]

Model	BIGOS	PELCRA	Total
whisper-large-v3	6.08	29.04	21.39
whisper-large-v3-baseline	6.16	23.35	17.62
whisper-large-v3-mix-00	5.04	22.58	16.74
whisper-large-v3-mix-01	3.93	20.98	15.30
conformer-baseline	11.22	30.55	24.11
conformer-mix-00	7.85	27.32	20.84
conformer-mix-01	7.26	25.38	19.34

Results on test split of data [1]

	test-A		*test-B*	
Model	**CER**	**WER**	**CER**	**WER**
whisper-large-v3-baseline	7.15	11.52	7.10	11.23
whisper-large-v3-mix-00	6.85	11.07	6.91	11.15
whisper-large-v3-mix-01	6.90	11.27	6.85	11.07
conformer-baseline	8.77	17.48	8.37	16.82
conformer-mix-00	7.60	15.25	7.16	14.33
conformer-mix-01	7.08	13.99	6.90	13.40

Conclusions & Next Steps

- Addition of synthetic data improves results for both tested models
- No clear saturation even with tripling the amount of data

- Language model could be used to introduce even more variability in synthetic data
- More careful procedure for choosing audio prompts for synthesis could be beneficial

- Decent voice-cloning speech synthesis system can be trained with as little as 700h of labelled speech data

Thank You!

References

[1] Ł. Bondaruk, J. Kubiak, and M. Czyżnikiewicz, Augmenting Polish Automatic Speech Recognition System With Synthetic Data. 2024. [Online]. Available: https://arxiv.org/abs/2410.22903
[2] https://lorenlugosch.github.io/posts/2020/11/transducer/?ref=assemblyai.com
[3] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, Robust Speech Recognition via Large-Scale We ak Supervision. 2022. [Online]. Available: https://arxiv.org/abs/2212.04356
[4] A. Gulati et al., Conformer: Convolution-augmented Transformer for Speech Recognition. 2020. [Online]. Available: https ://arxiv.org/abs/2005.08100
[5] M. Le et al., Voicebox: Text-Guided Multilingual Universal Speech Generation at Scale. 2023. [Online]. Available: https://a rxiv.org/abs/2306.15687