Inter-University Research Institute Corporation /
Research Organization of Information and Systems
National Institute of Informatics

Interspeech 2026
Area 4 - Oral 2 - 3

Deepfake Word Detection by Next-token Prediction using Fine-tuned Whisper

Hoan My Tran[1*], **Xin Wang**[2*], Wanying Ge[2], Xuechen Liu[2], Junichi Yamagishi[2]

[1]Université de Rennes, France

[2]National Institute of Informatics, Japan

wangxin@nii.ac.jp

Slides by Xin Wang
National Institute of Informatics

© 2026, Xin Wang. All rights reserved.

This work is licensed under the Creative Commons Attribution 3.0 license.
See http://creativecommons.org/ for details.



Note: Speech data is from ASVspoof 2019 database
https://datashare.ed.ac.uk/handle/10283/3336

One-minute summary

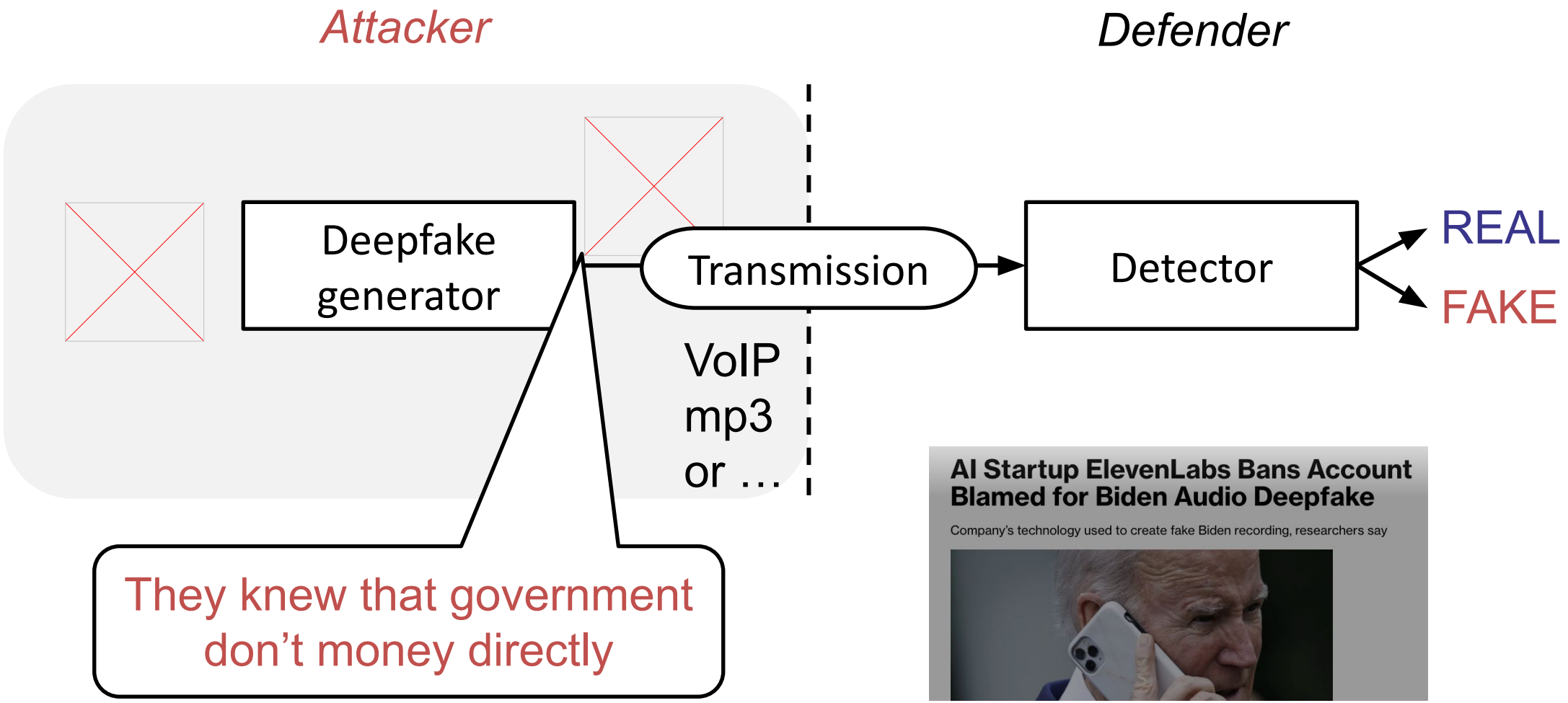
- *Finetune a Whisper to detect fake words via next-token prediction*
  - *ASR + fake words detection **at the same time***
  - *Method*
    1. *Prepare the finetuning data*
       - *Vocoding randomly selected words*
       - *Add <special token> around vocoded words in target token sequences*
    2. *Finetune Whisper*

- *Good results on in-domain test sets*
- *Improvement is needed on out-of-domain data*

Introduction

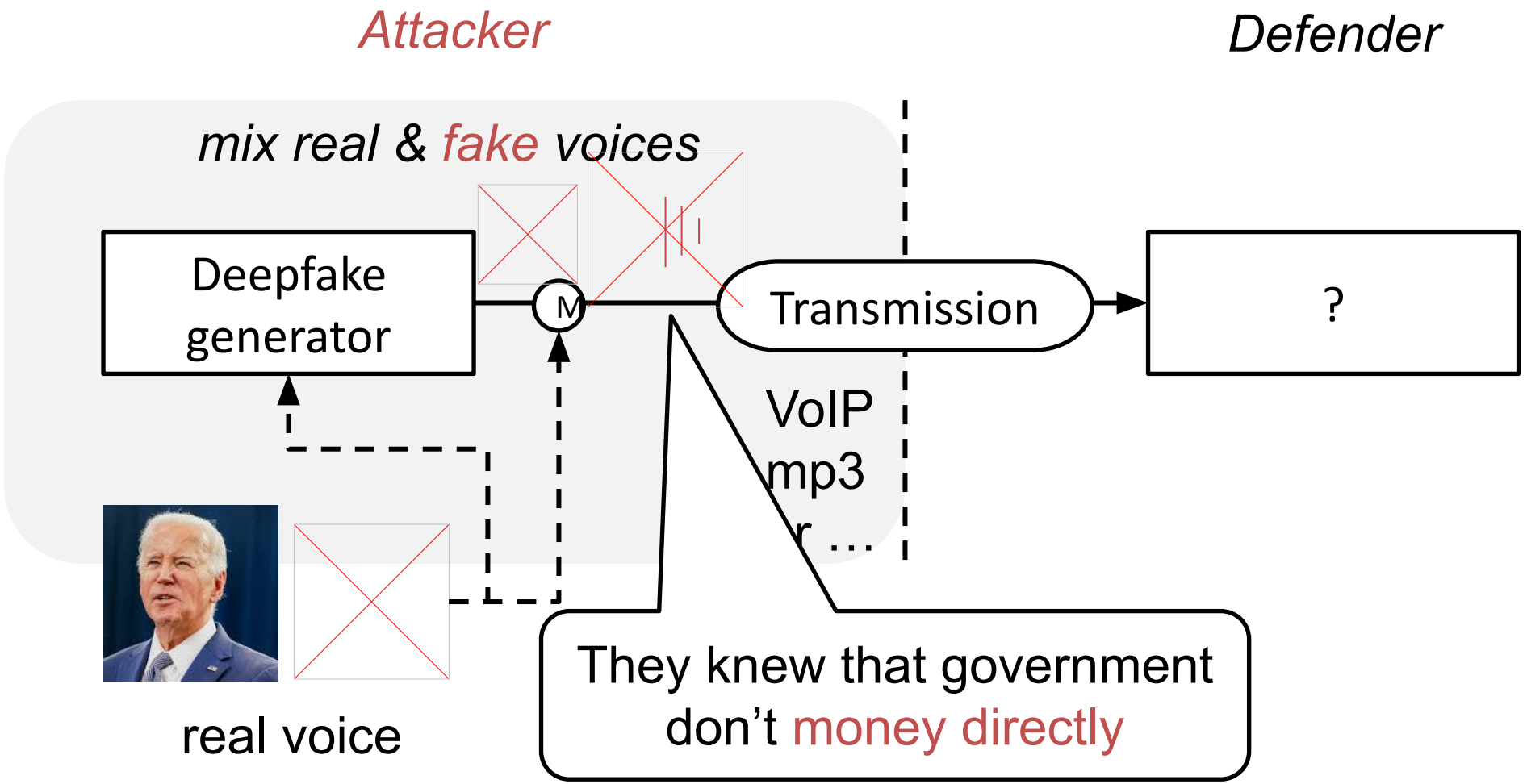
Introduction

- **Popular task: whether an utterance is fake**



Introduction

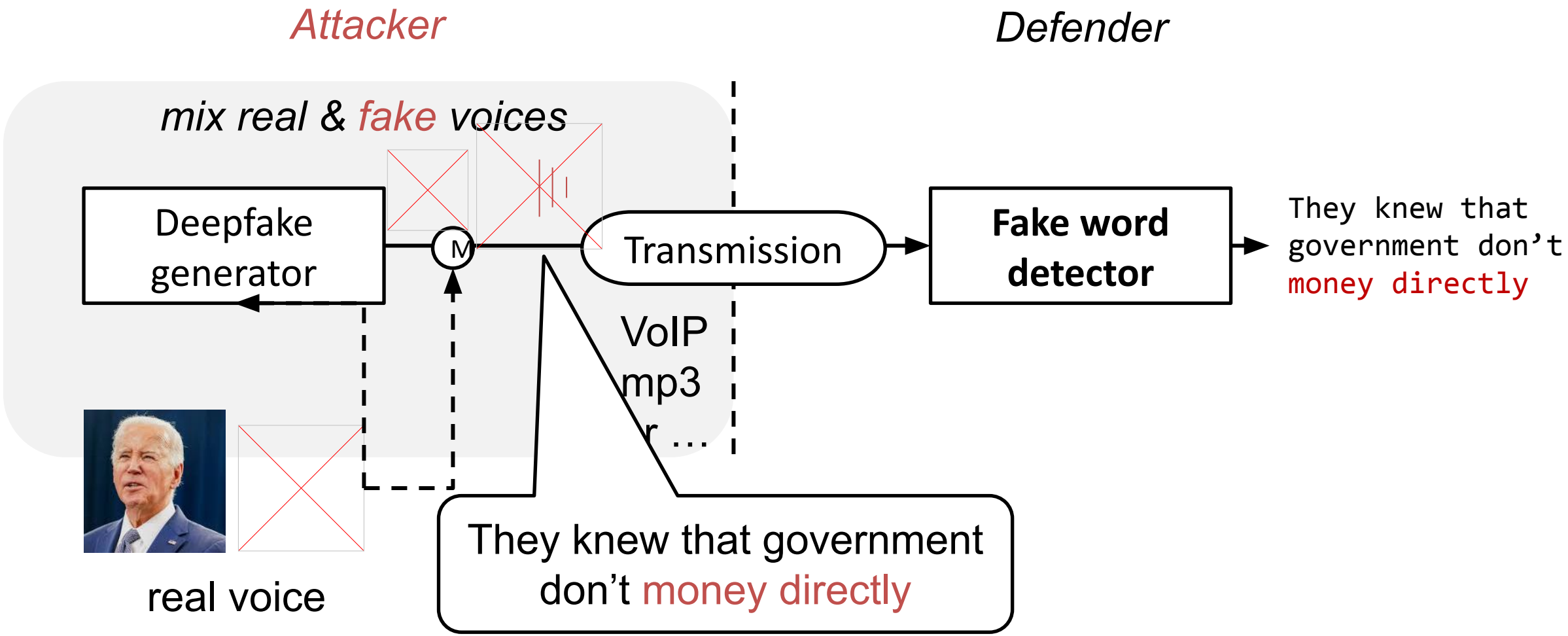
- **Task in this paper: which part is fake?**



Example from Meta VoiceCraft: https://jasonppy.github.io/VoiceCraft_web/
https://www.bloomberg.com/news/articles/2024-01-26/ai-startup-elevenlabs-bans-account-blamed-for-biden-audio-deepfake?embedded-checkout=true

Introduction

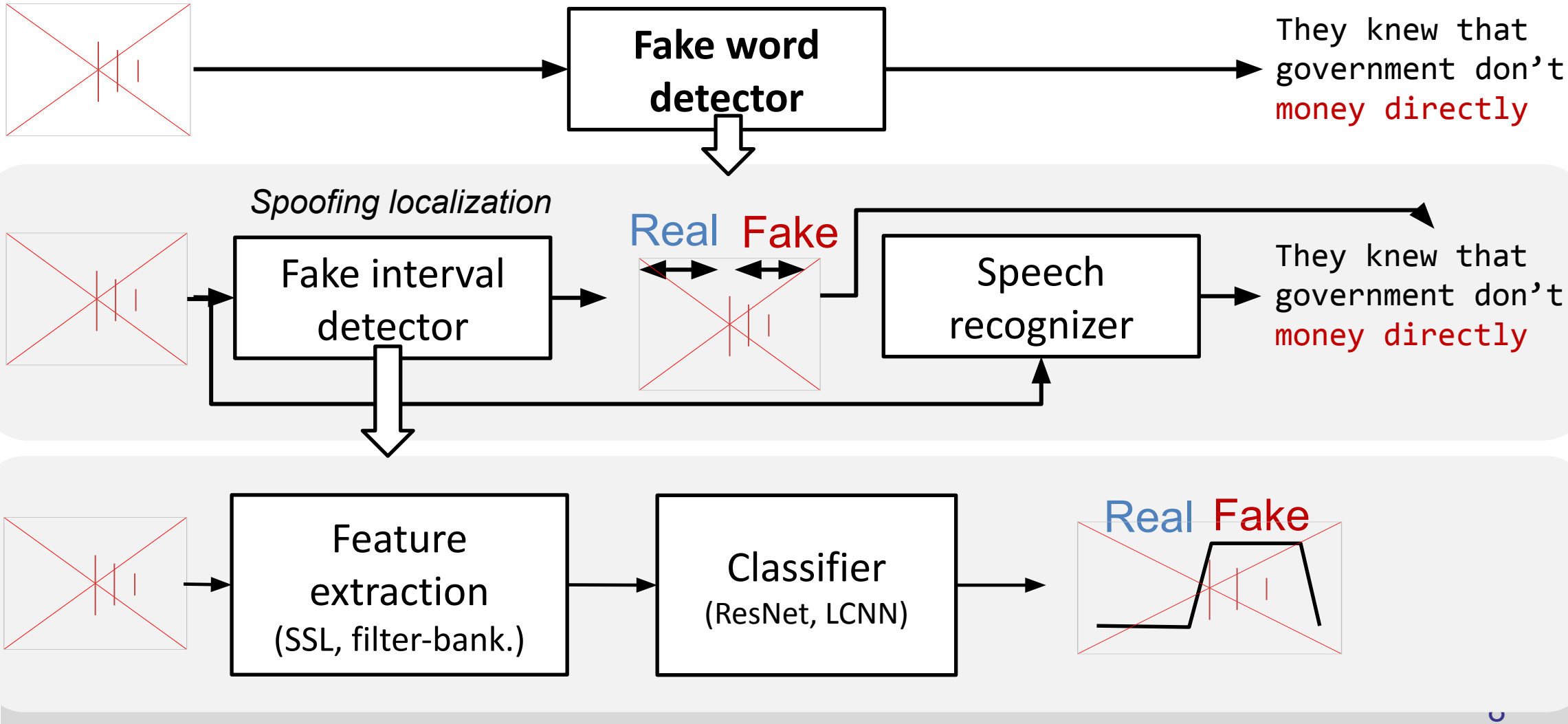
- **Task in this paper: which part is fake?**



Introduction

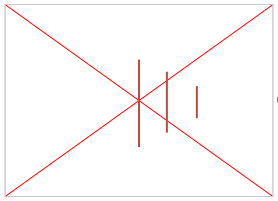
Simpler solution to transcribe the speech & locate fake words?

❑ Existing method



Introduction

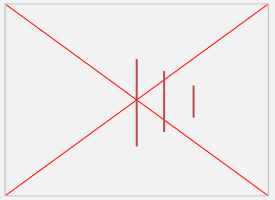
Our method



Fake word detector

They knew that government don’t money directly

Next-token prediction

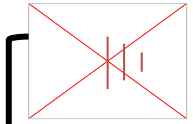


Whisper

They knew that government don’t <TOF>money directly<EOF>

Fine-tuning

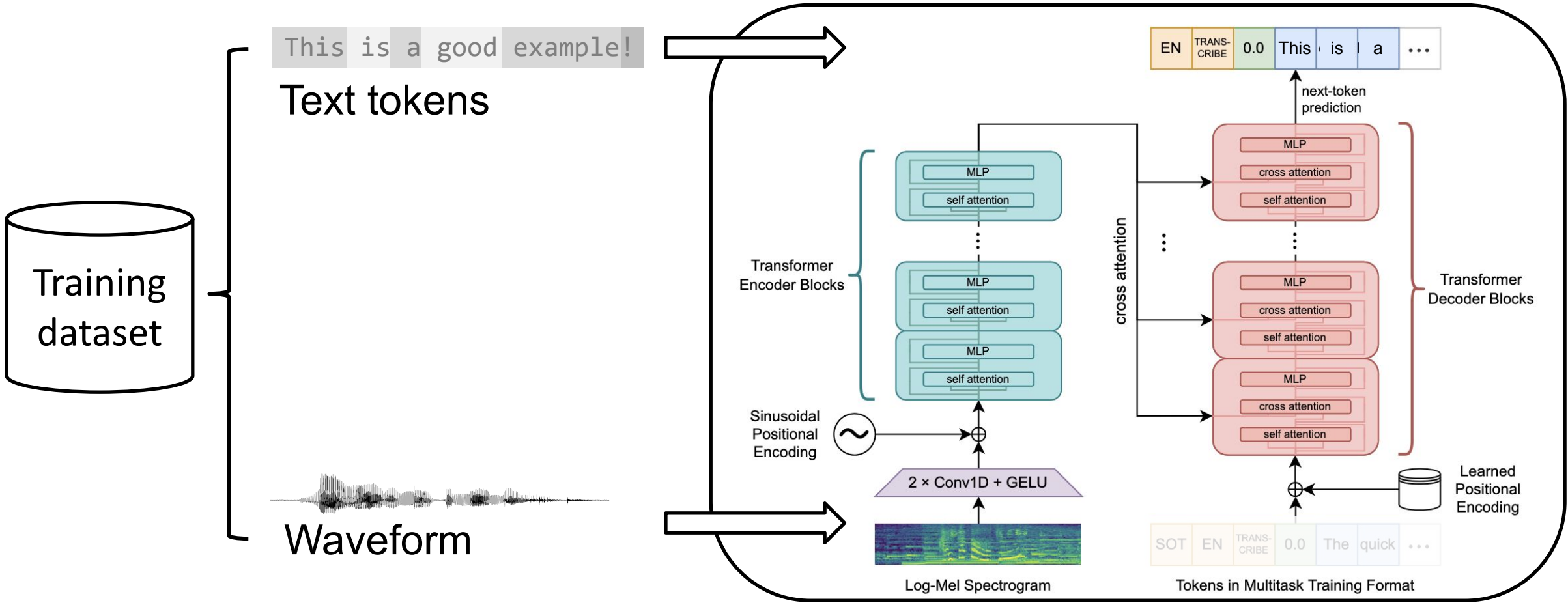
Pseudo fake (vocoded) dataset



This is an <TOF>example<EOF>

Method

Recap: Whisper



Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*, 2023. PMLR, 28492–28518.

Proposed method – at the high level

Prepare vocoded training data

Fine-tune Whisper (using default recipe)

Pseudo fake (vocoded)

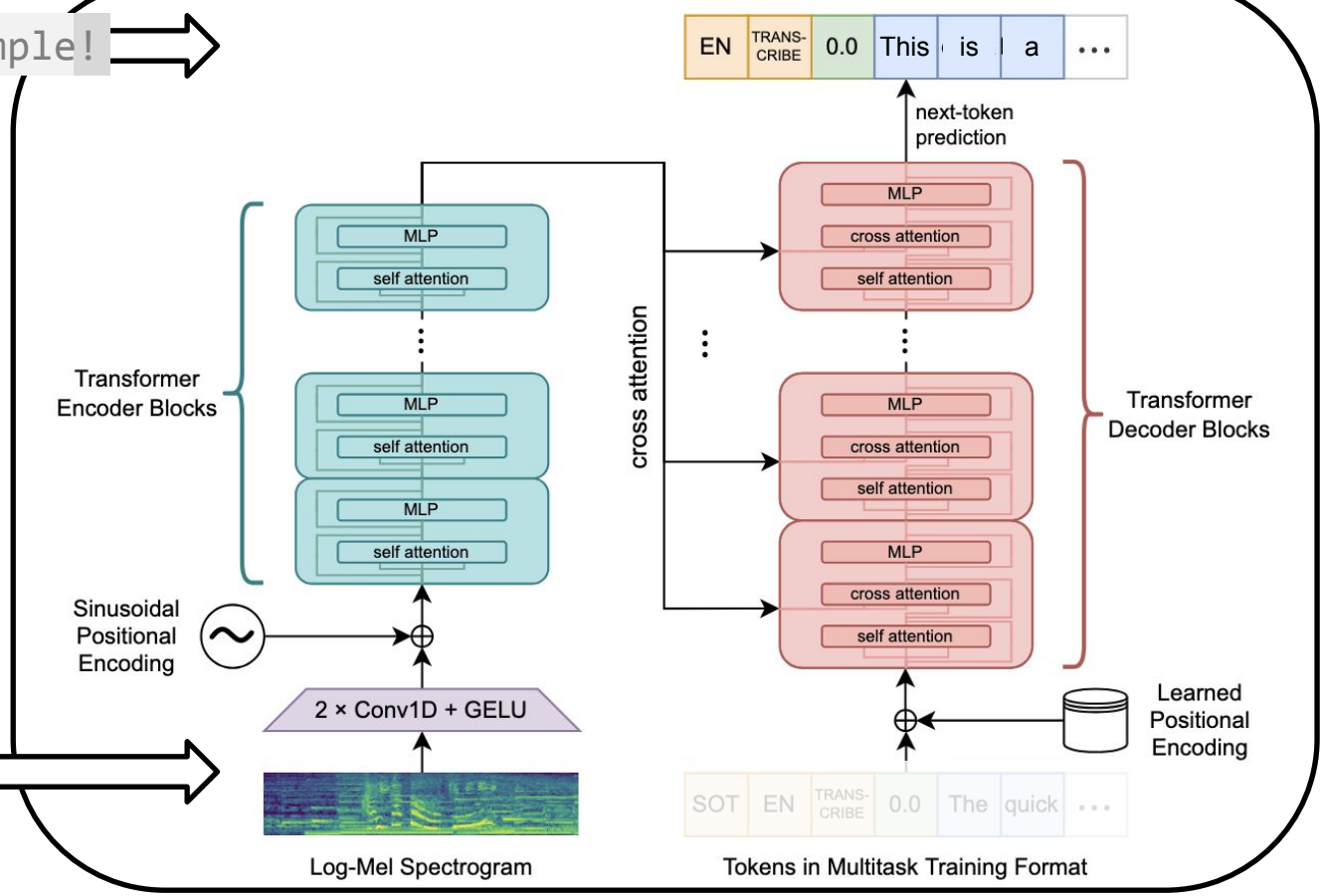
Training dataset

This is a<TOF>good<EOF> example!

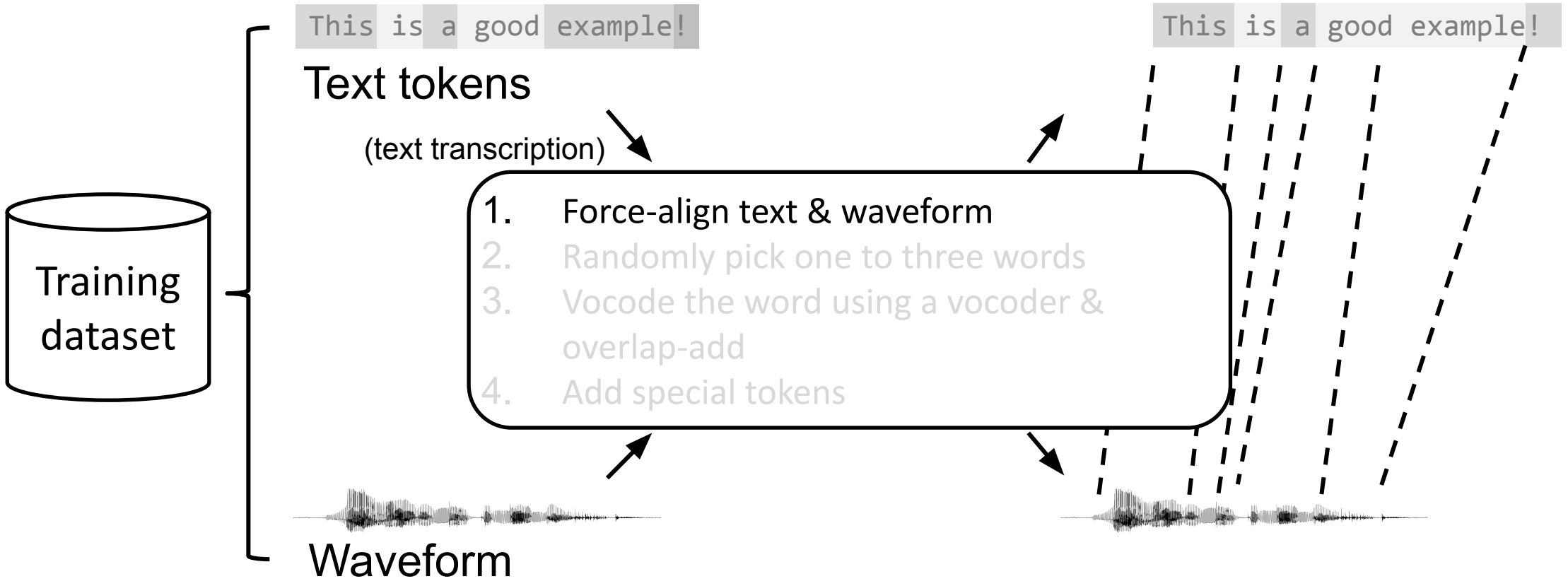
Text tokens



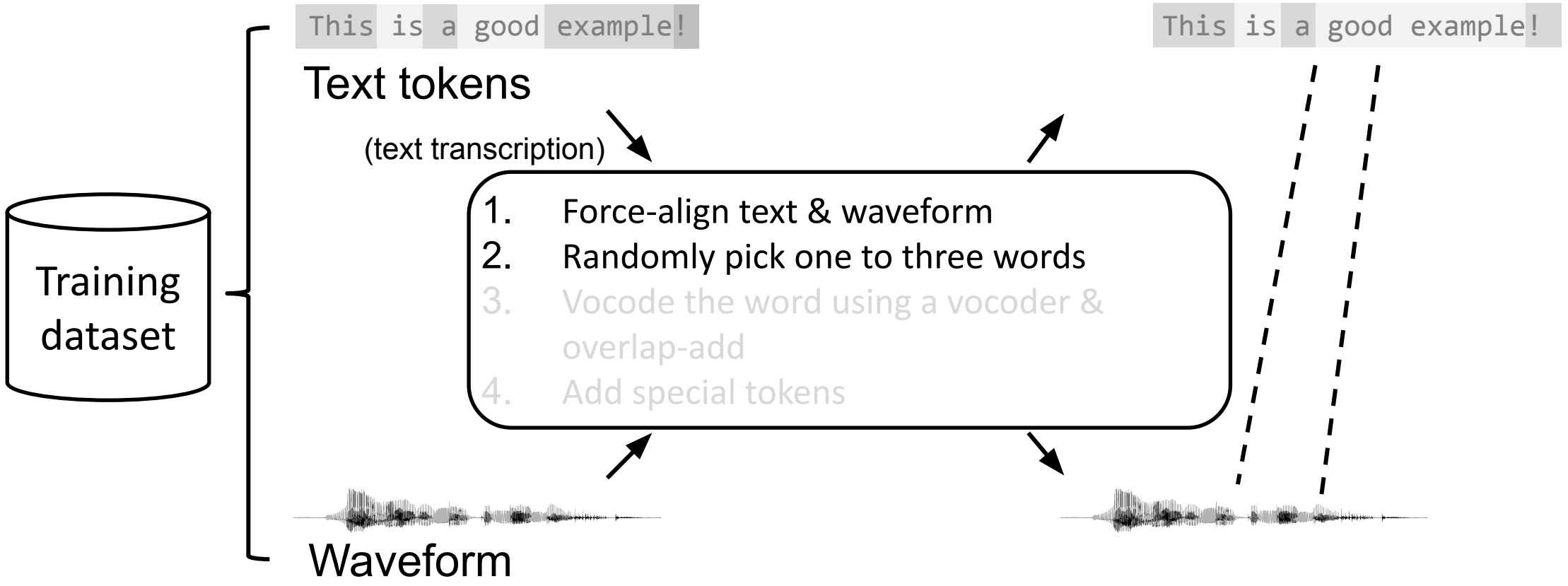
Waveform partially vocoded



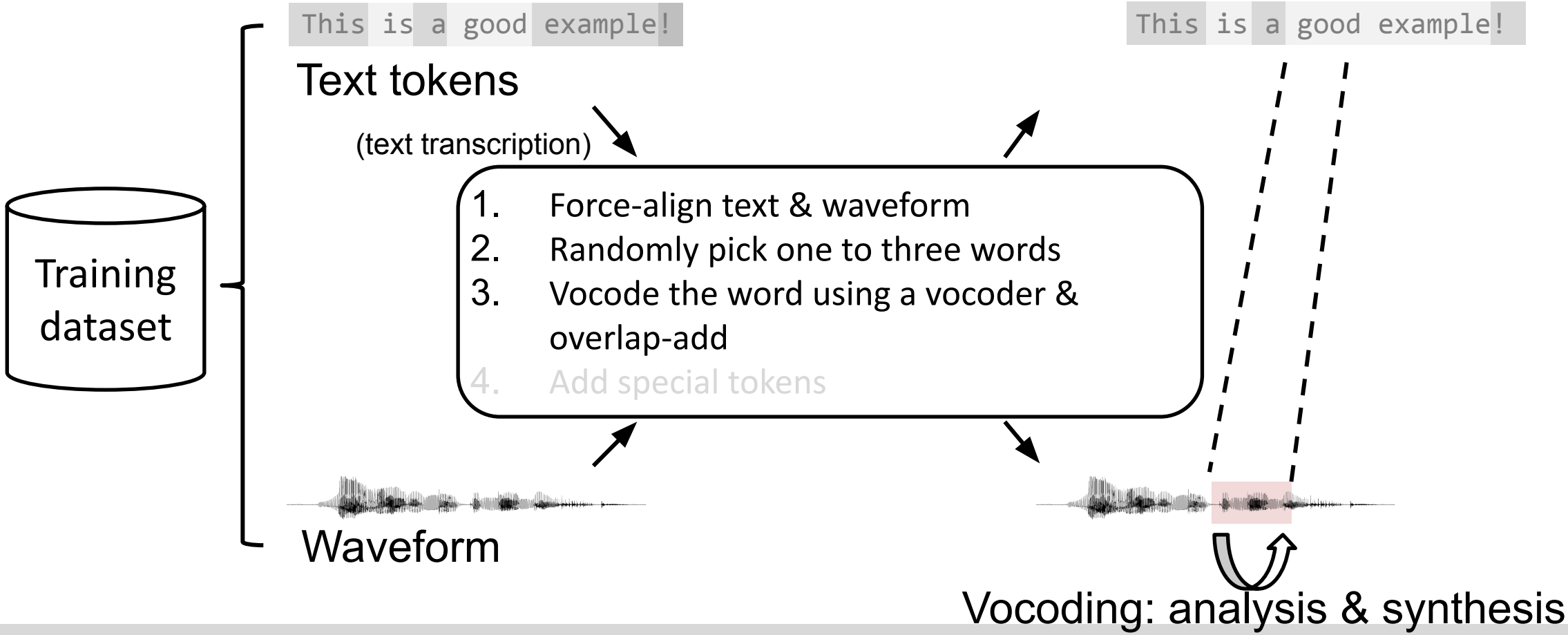
Proposed method – prepare data



Proposed method – prepare data

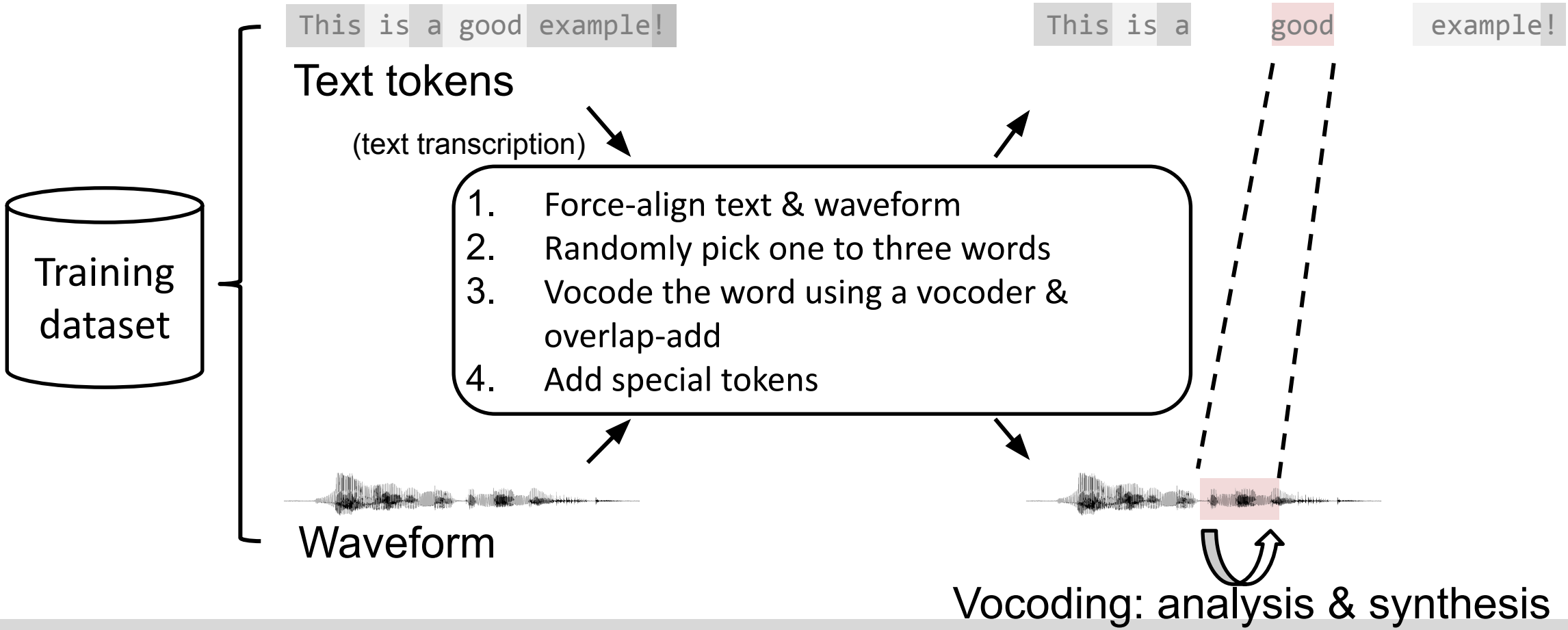


Proposed method – prepare data



See code pipeline: https://github.com/nii-yamagishilab/project-NN-Pytorch-scripts/tree/master/package

Proposed method – prepare data



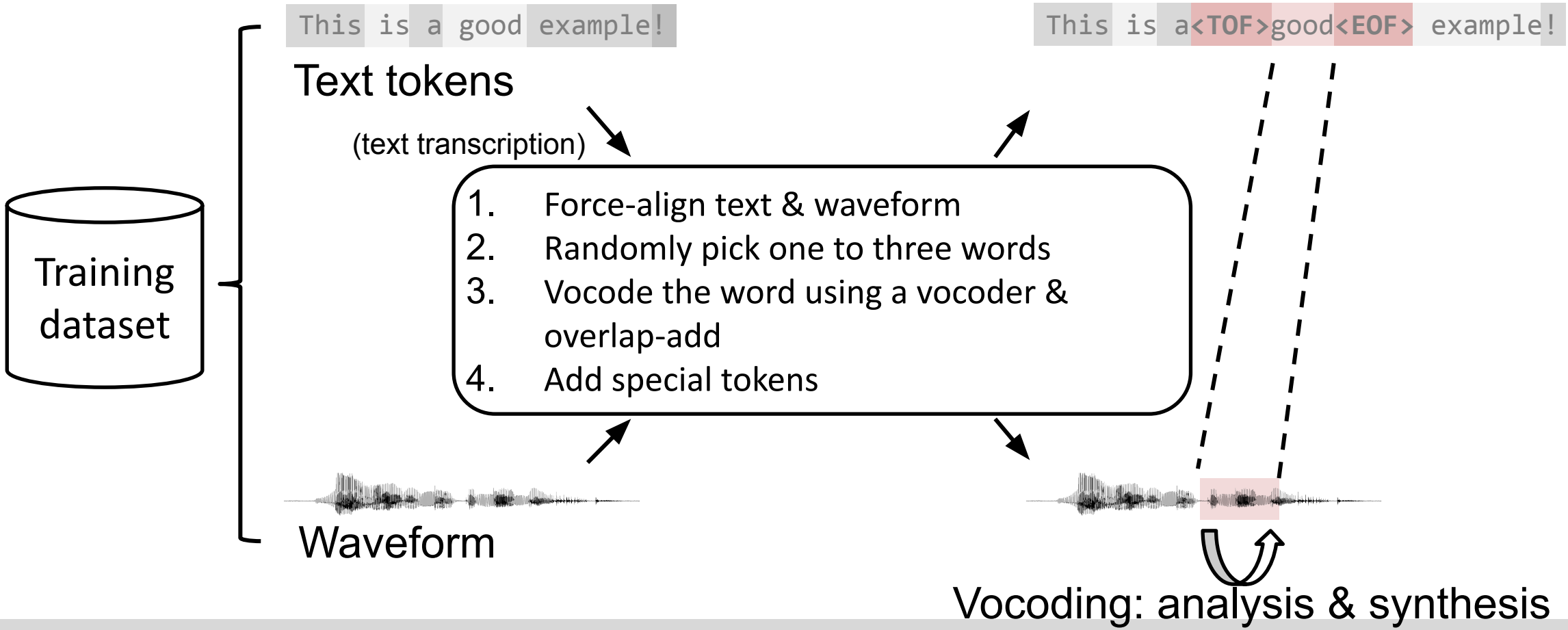
See code pipeline: https://github.com/nii-yamagishilab/project-NN-Pytorch-scripts/tree/master/package

Proposed method – prepare data

We re-use existing tokens:

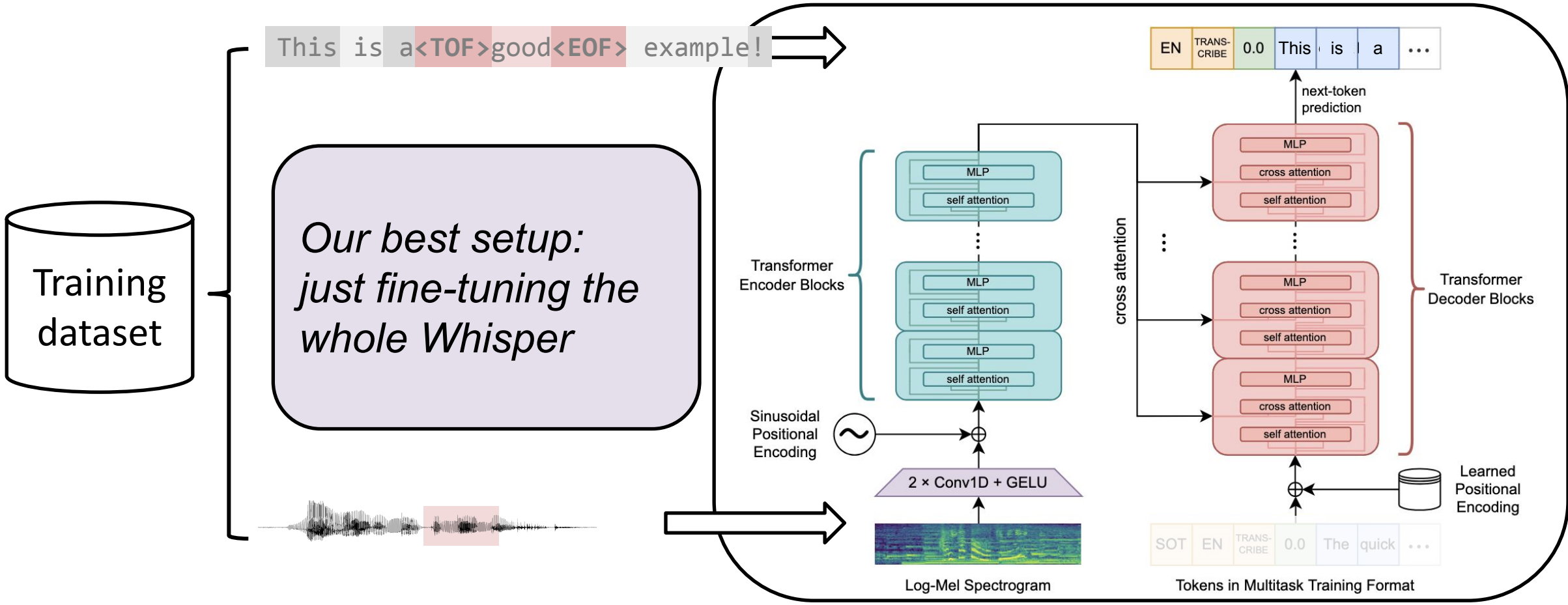
<TOF> !!!!!!

<EOF> ~~~



See code pipeline: https://github.com/nii-yamagishilab/project-NN-Pytorch-scripts/tree/master/package

Proposed method – fine-tuning Whisper



We tried LoRA, or fine-tune encoder / decoder only, but performance was sub-optimal

Examples of Whisper's prediction

Ground-truth: !!!!!!Now~~~ this !!!!!!isn't~~~ absolutely definitive.
Whisper: !!!!!!Now~~~ this !!!!!!isn't~~~ absolutely definitive.

Dataset: E.VOC

Ground-truth: It seemed a moving and !!!!!!fitting~~~ !!!!!!destruction.~~~
Whisper: It seemed a moving and fitting !!!!!!distraction.~~~

Dataset: PartialEdit

False acceptance (of a fake word)

Ground-truth: I really !!!!!!didn't~~~ enjoy putting ...
Whisper: I really !!!!!!didn't~~~ !!!!!!kind~~~ of putting ...

Dataset: AVDeepfake-1m

False rejection (of a real word)

Experiment

Experiment setup

Research questions

- Can fine-tuned Whisper detect synthetic words while transcribing?
- Is the vocoding-based training data helpful?

Evaluation metrics

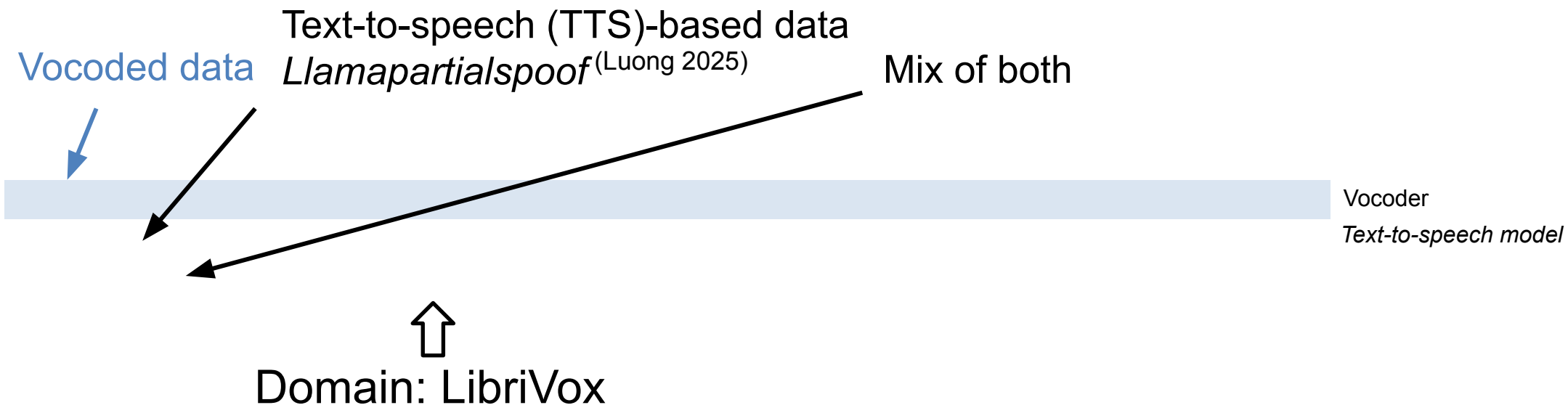
- False rejection rate (FRR, *real* words misclassified as fake)
- False acceptance rate (FAR, *fake* words misclassified as real)
- Word error rates (WER)

Experiment setup

Research questions

- Can fine-tuned Whisper detect synthetic words while transcribing?
- Is the vocoding-based training data helpful?

Data

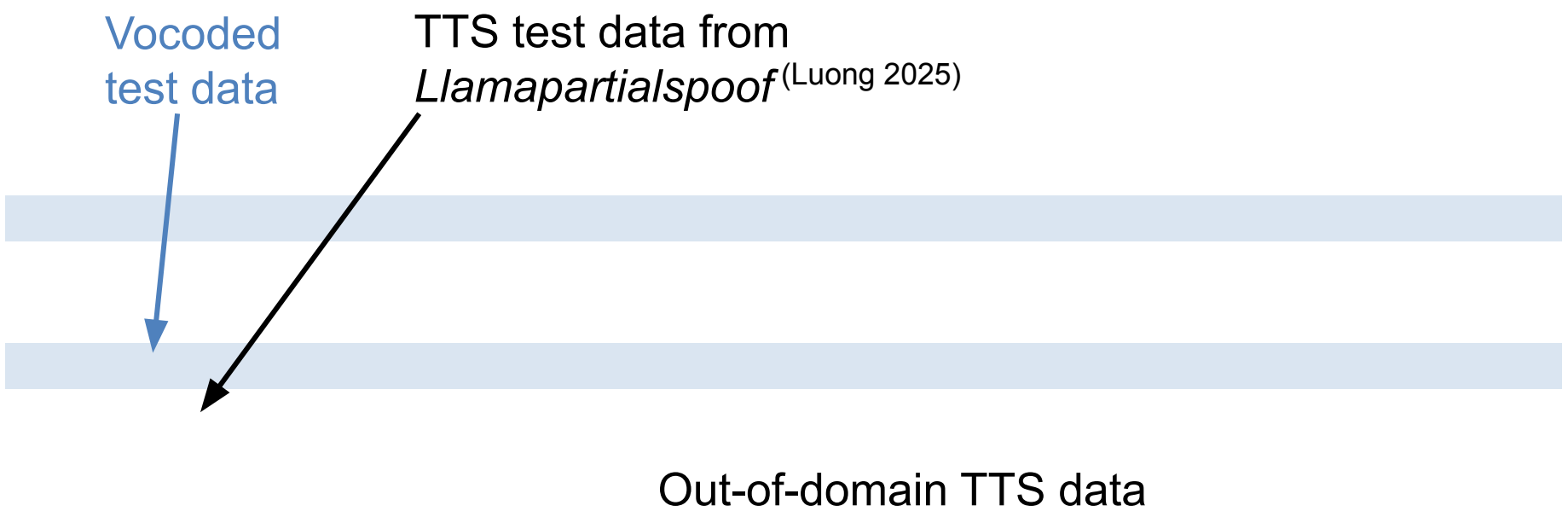


Experiment setup

Research questions

- Can fine-tuned Whisper detect synthetic words while transcribing?
- Is the vocoding-based training data helpful?

Data



Results – Audiobook test data

The lower the better

Evaluate on vocoded data (E.VOC)

		WER (%)	FAR (%)	FRR (%)
Pre-trained	Whisper large v3	23.89	-	-
Fine-tuned / trained using vocoded data Ft.VOC	Whisper large v3	0.87	7.22	0.52
	ResNet-152 (AVDeepfake 1M++ top-1, Klein 2025)	-	7.15	3.81

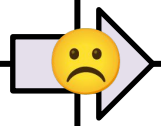
RQ1: it is possible that a fine-tuned Whisper can detect synthetic (vocoded) words while transcribing

Nicholas Klein, Hemlata Tak, Ja
Pindrop it! Audio and visual deepfake countermeasures for robust detection and fine-grained localization. In *Proc. ACM MM* (*Mm '25*), 2025. Association for Computing Machinery, Dublin, Ireland, 13700–13706.

Results – Audiobook test data

The lower the better

		Evaluate on vocoded data (E.VOC)		*Evaluate on TTS data (E.TTS)*	
		FAR (%)	FRR (%)	FAR (%)	FRR (%)
Fine-tuned / trained using vocoded data Ft.VOC	Whisper large v3	7.22	0.52	76.16	9.01
	ResNet-152 (AVDeepfake 1M++ top-1, Klein 2025)	7.15	3.81	83.04	10.14



Detecting unseen synthetic methods?

Most of the fake words are mis-classified as REAL

Results – Audiobook test data

RQ2: cross-dataset generalization is an issue for both models

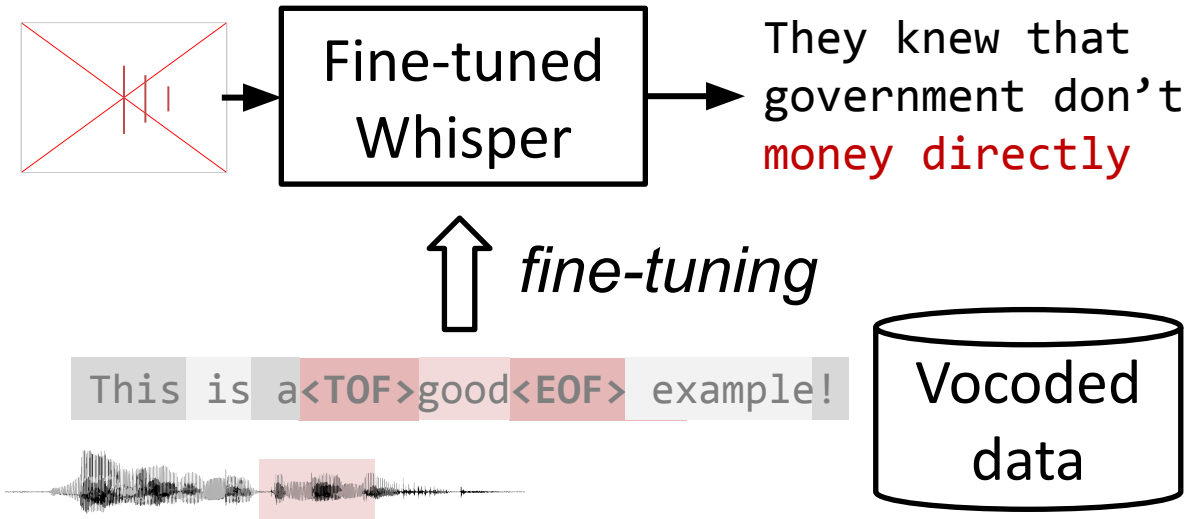
The lower the better

		Evaluate on vocoded data (E.VOC)		*Evaluate on TTS data (E.TTS)*	
		FAR (%)	FRR (%)	FAR (%)	FRR (%)
Fine-tuned / trained using vocoded data Ft.VOC	Whisper large v3	7.22	0.52	76.16	9.01
	ResNet-152 (AVDeepfake 1M++ top-1, Klein 2025)	7.15	3.81	83.04	10.14
Fine-tuned / trained using TTS data Ft.TTS	Whisper large v3	18.86	78.59	1.38	1.79
	ResNet-152	76.42	9.08	0.15	3.13

Summary

Summary

- **Detecting fake words via next-token prediction**
  - create vocoded training data & add special tokens
  - fine-tune Whisper



- **It is feasible to detect synthetic words via next-token prediction**
- **Room for improvement**
  - Unseen attacks
  - Unseen domains (Youtube, studio) → see paper
  - Diverse training data may be the key

Related work

Partial fake (interval) detection

- Metrics: counting (Yi 2022), intersection of union (Li 2025), range-EER (Zhang 2023)
- Models: single-resolution (Yi 2022, Zhang 2022, Klein 2025), multi-resolution (Zhang 2022)
- Datasets: PartialSpoof (Zhang 2022), LlamaPartialSpoof (Luong 2025), PartialEdit (Zhang 2025)

Using Whisper or large language model for fake detection

- Whisper as feature extractor for utterance-level detection (Kawa 2023)
- Large language model (LLM) fake word detection (Lin 2026)
- Many other latest work using LLMs

Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang, Zhengkun Tian, Ye Bai, Cunhang Fan, Shan Liang, Shiming Wang, Shuai Zhang, Xinrui Yan, Le Xu, Zhengqi Wen, and Haizhou Li. 2022. ADD 2022: the first Audio Deep Synthesis Detection Challenge. In *Proc. ICASSP*, May 2022. 9216–9220.
Yupei Li, Li Wang, Yuxiang Wang, Lei Wang, Rizhao Cai, Jie Shi, Björn W. Schuller, and Zhizheng Wu. 2025. DFALLM: Achieving Generalizable Multitask Deepfake Detection by Optimizing Audio LLM Components.
Lin Zhang, Xin Wang, Erica Cooper, Junichi Yamagishi, Jose Patino, and Nicholas Evans. 2021. An Initial Investigation for Detecting Partially Spoofed Audio. In *Proc. Interspeech*, August 2021. 4264–4268.
Lin Zhang, Xin Wang, Erica Cooper, Nicholas Evans, and Junichi Yamagishi. 2022. The PartialSpoof Database and Countermeasures for the Detection of Short Fake Speech Segments Embedded in an Utterance. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2022), 1–13.
Lin Zhang, Xin Wang, Erica Cooper, Nicholas Evans, and Junichi Yamagishi. 2023. Range-Based Equal Error Rate for Spoof Localization. In *Proc. Interspeech*, August 20, 2023. ISCA, 3212–3216.
Piotr Kawa, Marcin Plata, Michał Czuba, Piotr Szymański, and Piotr Syga. 2023. Improved DeepFake Detection Using Whisper Features. In *Proc. Interspeech*, August 20, 2023. ISCA, 4009–4013.
Lin Zhang, Thomas Thebaud, Zexin Cai, Sanjeev Khudanpur, Daniel Povey, Leibny Paola Garcia-Perera, Matthew Wiesner, and Nicholas Andrews. 2026. Can LLMs Help Localize Fake Words in Partially Fake Speech?
Hieu-Thi Luong, Haoyang Li, Lin Zhang, Kong Aik Lee, and Eng Siong Chng. 2025. LlamaPartialSpoof: An LLM-Driven Fake Speech Dataset Simulating Disinformation Generation. In *Proc. ICASSP*, January 06, 2025. arXiv.
Nicholas Klein, Hemlata Tak, James Fullwood, Krishna Regmi, Leonidas Spinoulas, Ganesh Sivaraman, Tianxiang Chen, and Elie Khoury. 2025. Pindrop it! Audio and visual deepfake countermeasures for robust detection and fine-grained localization. In *Proc. ACM MM* (*Mm '25*), 2025. Association for Computing Machinery, Dublin, Ireland, 13700–13706.

Thank you



Code & checkpoint

https://github.com/nii-yamagishilab/Whisper-deepfake-word-detection