Speaker anonymization with neural audio codec language models

Michele Panariello,[1] *Francesco Nespoli*,[2] *Massimiliano Todisco*[1] *and Nicholas Evans*[1]

[1]**EURECOM** - Sophia Antipolis, France

[2]**Microsoft UK** - London

Champion et al. [1], INTERSPEECH 2022:
Proposes speaker anonymization system in which speech features are **discretized** to reduce personal information leakage

Borsos et al. [2], TASLP 2023
Wang et al. [3], 2023
Audio generation by autoregressively modeling **discrete tokens** of a *neural audio codec* (NAC) [4]

Suno AI, 2023
Releases Bark,[1] open-source NAC-based TTS system inspired by [2,3].

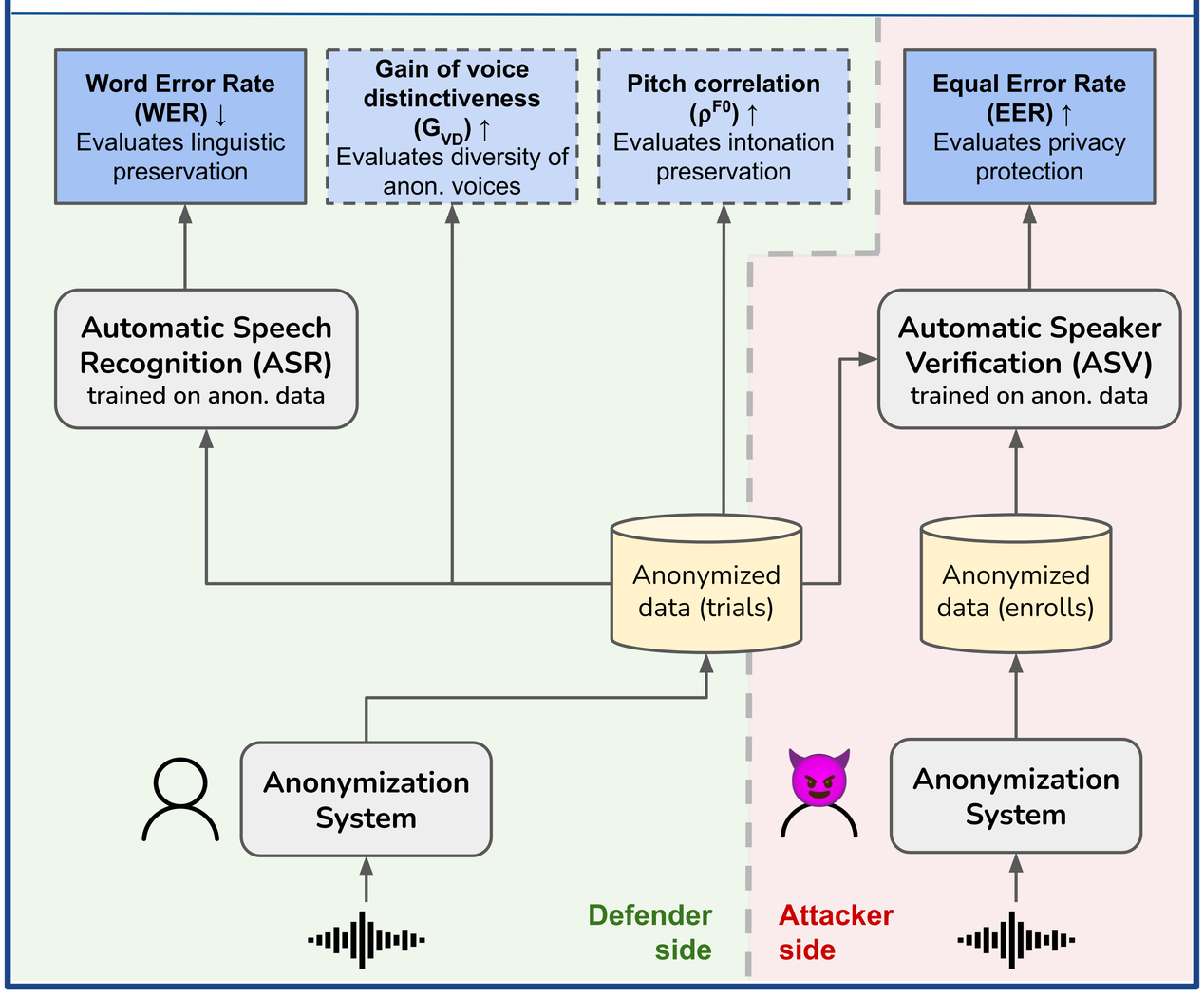
This paper:
Apply discrete NAC token modeling for speaker anonymization

The speaker anonymization task

Formalized in the VoicePrivacy initiative [5], speaker anonymization is the task of taking an input speech signal and processing it so that

- The **liguistic** (spoken) content is preserved
- The **paralinguistic** content (emotion, intonation) is preserved
- The **identity** of the speaker is concealed

The output is a new, anonymized waveform.



Proposed anonymization system

- **Semantic encoder (HuBERT):** encodes audio to anonymize in discrete semantic tokens ***s*** from a dictionary of size N_S

$$\mathbf{s} \in \{1, \dots, N_S\}^{T_S}$$

- **NAC encoder (EnCodec):** encodes random pseudo-speaker prompt in discrete acoustic tokens $\tilde{\boldsymbol{a}}$ using Q hierarchical dictionaries, each of size N_Q. Lower-level dictionaries encode coarser features

$$\tilde{\mathbf{a}} \in \{1, \dots, N_Q\}^{Q \times T_A}$$

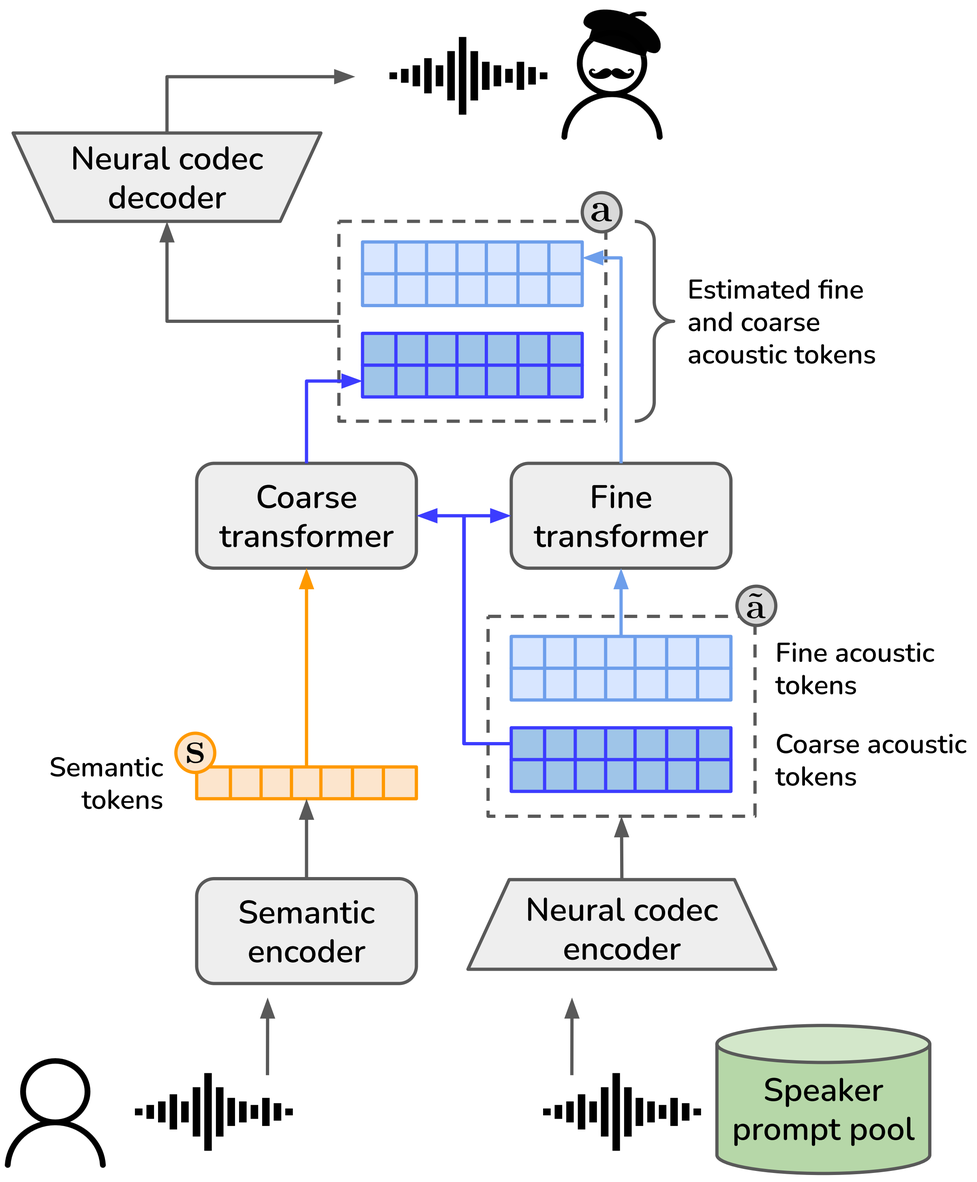
- **Coarse transformer (GPT-like):** estimates the first Q_C levels of output acoustic tokens ***a*** from ***s*** and $\tilde{\boldsymbol{a}}$. They follow the semantics in ***s*** and the speaking style in $\tilde{\boldsymbol{a}}$

$$p\left(\mathbf{a}_{q,t} | \mathbf{s}, \tilde{\mathbf{a}}_{<Q_C,:}, \mathbf{a}_{<Q_C,<t}, \mathbf{a}_{<q,t}\right)$$

- **Fine transformer (GPT-like):** estimates remaining Q - Q_C levels of acoustic tokens ***a*** from $\tilde{\boldsymbol{a}}$

$$p\left(\mathbf{a}_{q,:} | \tilde{\mathbf{a}}, \mathbf{a}_{<q,:}\right)$$

- **NAC decoder (EnCodec):** converts ***a*** to waveform.



Results

System	LibriSpeech				VCTK			
	EER (%)	WER (%)	G_{VD}	ρ^{F0}	EER (%)	WER (%)	G_{VD}	ρ^{F0}
Original data	4.4	4.2	0	1	3.2	12.8	0	1
B1b [6]	8.6	4.4	-5.8	**0.78**	9.7	10.7	-7.1	**0.81**
T11 [7]	20.6	**3.9**	-19.0	0.68	39.7	**7.9**	-18.4	0.73
Champion et al. [1]	23.4	4.6	n.a.	0.52	40.8	10.3	n.a	0.60
Ours	**28.5**	7.5	**-1.5**	0.68	**45.5**	18.9	**-2.1**	0.74

References

[1] P. Champion et al., “Are disentangled representations all you need to build speaker anonymization systems?,” in Interspeech 2022, ISCA, pp. 2793–2797.
[2] Z. Borsos et al., “AudioLM: A Language Modeling Approach to Audio Generation,” in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 2523-2533, 2023.
[3] C. Wang et al., “Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers.” arXiv, Jan. 05, 2023.
[4] A. Défossez et al. “High Fidelity Neural Audio Compression,” in Transactions on Machine Learning Research, 2023.
[5] N. Tomashenko et al., “Introducing the VoicePrivacy Initiative,” in Interspeech 2020, ISCA, pp. 1693-1697, 2020.
[6] N. Tomashenko et al., “The VoicePrivacy 2022 Challenge evaluation plan,” arXiv, Sep. 28, 2022.
[7] J. Yao et al., “NWPU-ASLP System for the VoicePrivacy 2022 Challenge,” in Proc. 2nd Symposium on Security and Privacy in Speech Communication, 2022.

[1] https://github.com/suno-ai/bark