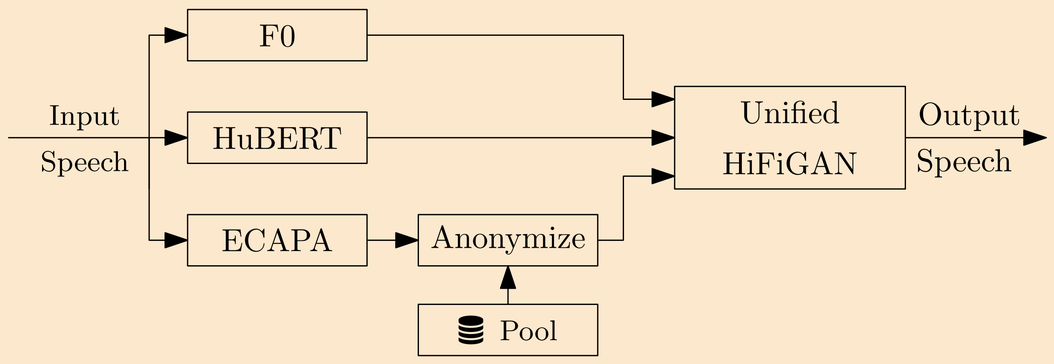
INTERNATIONAL AUDIO LABORATORIES ERLANGEN
A joint institution of Fraunhofer IIS and Universität Erlangen-Nürnberg

Why disentanglement-based speaker anonymization systems fail at preserving emotions?

Ünal Ege Gaznepoglu, Nils Peters (ege.gaznepoglu@audiolabs-erlangen.de)

1. Introduction

Disentanglement-based speaker anonymization is a popular approach, an example is SSL-SAS (signal flow diagram given below) [1]



1) Obtain a disentangled representation by, e.g., (F0, ECAPA embeddings, post-processed HuBERT embeddings)
2) Modify the speaker identity (ECAPA embeddings)
3) Synthesize speech using a neural vocoder (HiFiGAN etc.)

Often, emotion information is lost in the process.

2. Hypotheses

We identified three possible reasons:

? GAN-based vocoders cannot produce emotional speech
- Especially if trained on read speech (e.g., LibriTTS train-clean-100)
- Recent variants (e.g., BigVGAN) claim better OOD performance

? Speaker embeddings contain emotion information
- Initial experiments in [2] showed limited improvement

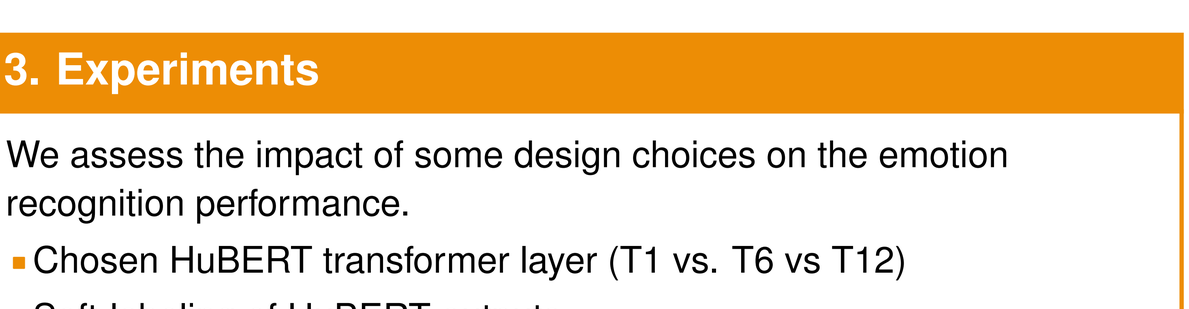
? Intermediate representation not containing emotion information

3. Experiments

We assess the impact of some design choices on the emotion recognition performance.
- Chosen HuBERT transformer layer (T1 vs. T6 vs T12)
- Soft-labeling of HuBERT outputs
- Discriminatively-learned speaker embeddings (ECAPA) vs. generatively learned (GST)
- Original vs. anonymized speaker embeddings

4. Evaluation

We use the VPC 2024 evaluation protocol [3]
- Unweighted average recall (UAR) computed by a speech emotion recognition (SER) system for emotion preservation (left)
- Word error rate (WER) computed by an automated speech recognition (ASR) system for intelligibility (right)



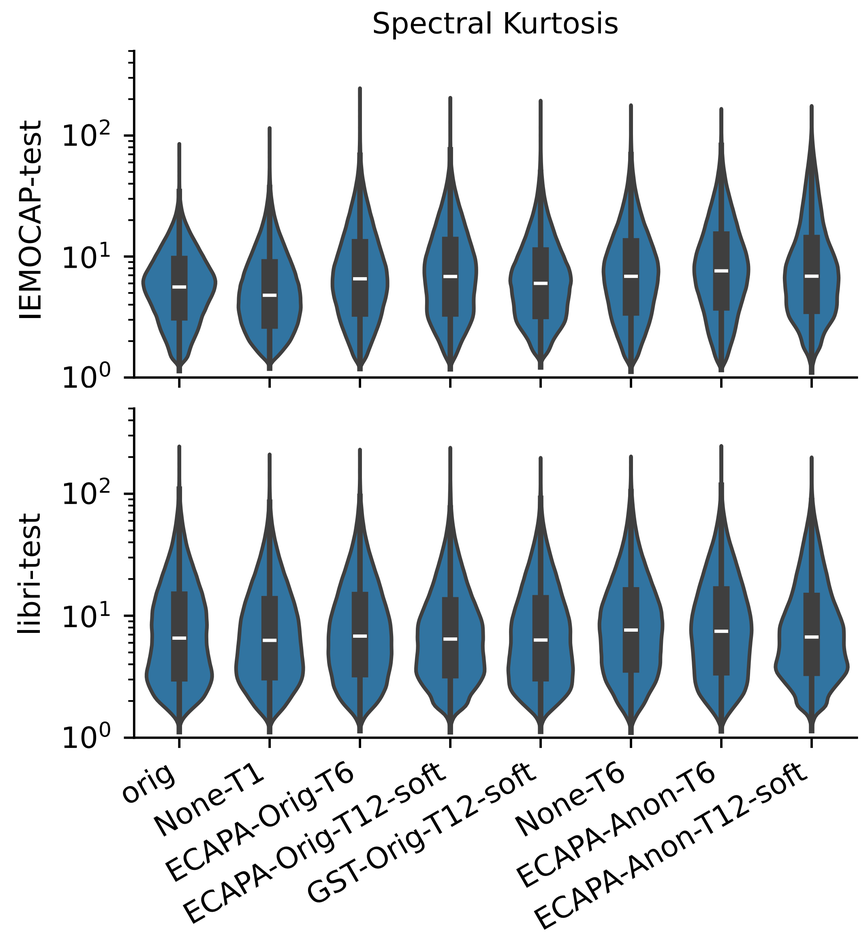
In addition, we compute three emotion-related acoustic features [4]:
- Spectral centroid
- Spectral kurtosis
- Loudness

5. Results

i) VPC 2024 Evaluations

		Libri-test	IEMOCAP-test				
			Sad	Neutral	Angry	Happy	All
Speaker emb.	HuBERT	WER [%] (↓)	Recall [%] (↑)				UAR [%] (↑)
Original data		1.84	25.39	72.58	71.66	72.82	67.19	71.06
B3		4.35	47.04	0.65	41.83	66.09	41.72	37.57
B4		5.90	50.45	11.26	46.68	61.54	54.64	42.78
B5		4.37	42.99	5.07	55.30	56.20	36.10	38.17
N/A	T1	2.11	**34.95**	**50.80**	62.04	72.58	73.69	**64.78**
ECAPA-Orig	T6	2.15	34.98	31.58	61.68	65.67	77.54	59.11
ECAPA-Orig	T12-soft	2.28	37.26	12.86	**69.20**	73.50	58.44	53.50
GST-Orig	T12-soft	2.30	37.46	29.28	75.26	60.00	48.85	53.35
N/A	T6	2.31	36.25	16.40	51.79	66.79	**79.95**	53.73
ECAPA-Anon	T6	**2.11**	35.51	25.77	64.80	63.88	70.59	56.26
ECAPA-Anon	T12-soft	2.42	38.85	3.42	46.71	**77.65**	52.98	45.19

ii) Impact on Emotion-related Acoustic Feature Distributions



For audio examples, please scan the QR code for the accompanying website

6. Conclusions

- **Primary reason: insufficient information in the bottleneck**
- Speaker embeddings learned in a generative context are more susceptible to capturing emotion information
- Vocoder impact is limited
- **UAR for emotion recognition performance is misleading**
- Synthesis artifacts increase spectral kurtosis, resulting in an increased angry and happy recall
- IEMOCAP utterances have quality issues (utterances without any linguistic content, with interfering speech, and background noise)

References

[1] X. Miao, X. Wang, E. Cooper, J. Yamagishi, and N. Tomashenko, "Language-independent speaker anonymization approach using self-supervised pre-trained models," in *Oddysey: Speaker and Lang. Recognition Workshop*, 2022.
[2] X. Miao, Y. Zhang, X. Wang, N. Tomashenko, D. C. L. Soh, and I. Mcloughlin, "Adapting general disentanglement-based speaker anonymization for enhanced emotion preservation," in *arXiv preprint*, 2024.
[3] Pierre Champion et al., *3rd VoicePrivacy challenge evaluation plan (version 2.0)*, 2024.
[4] S. Ghosh, A. Das, Y. Sinha, I. Siegert, T. Polzehl, and S. Stober, "Emo-StarGAN: A semi-supervised any-to-many non-parallel emotion-preserving voice conversion," in *Proc. Interspeech Conf.*, 2023.

Friedrich-Alexander-Universität Erlangen-Nürnberg
Fraunhofer IIS
Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin