Emotion-Enhanced SpeakerAnonymisation Using the FreeVC Framework

Yuqi Li—Fudan University& Qifu Techonology
Yuanzhong Zheng—Qifu Techonology
Jingyi Fang—Qifu Techonology
Jinming Chen—Qifu Techonology

CONTENTS

01 BACKGROUND

02 PROPOSED METHOD

03 EXPERIMENT

04 RESULT

BACKGROUND

Definition

Modifying vocal attributes to prevent speaker identification while maintaining speech intelligibility and naturalness.

Importance

Ensures privacy and data security across various digital platforms.

Applications

•**Mobile Health**: Secure storage and querying of encrypted voice data in healthcare systems.
•**Journalistic Protection**: Anonymization in sensitive communication scenarios to protect source identities.
•**Consumer Technology**: Enhancing security in voice-enabled IoT devices, preventing eavesdropping and data exploitation.

Voice Anonymization
Challenges
Innovative Approaches
Applications

Trade-offs

Balancing between effective anonymization and the naturalness and intelligibility of speech.

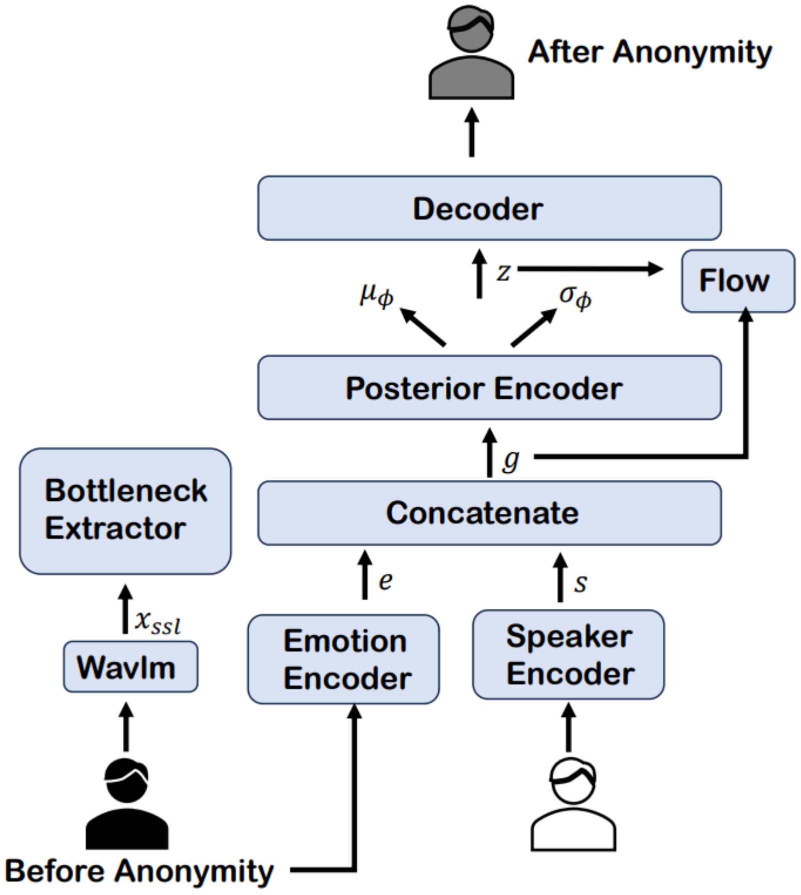
Technological Demands

Evolving speaker recognition technologies necessitate advanced anonymization techniques.

VoicePrivacy Challenge

In response to the call of VPC, a number of good methods that combine innovation and practicality have emerged

PROPOSED METHOD



Bottleneck Extractor

•**Function**: Compresses input features into a condensed representation, reducing dimensionality from 1024 to 2*d.

•**Purpose**: Retains essential content, minimizes speaker-specific characteristics to protect privacy.

•**Technology**: Uses convolutional layers with dilated structure to capture broad acoustic contexts.

Emotion Encoder

•**Approach 1**: Utilizes openSMILE for acoustic feature extraction, supplemented by LSTM for deep feature analysis.

•**Approach 2**: Employs Wav2vec 2.0 fine-tuned on MSP-Podcast for emotion recognition, extracting high-dimensional emotion embeddings.

Speaker Encoder

•**Description**: Encodes target speaker characteristics to maintain naturalness of the transformed speech while ensuring anonymity.

Concatenation and Processing

•**Flow**: Concatenates encoded features (content, emotion, speaker characteristics) and processes through Decoder and Posterior Encoder.

•**Output**: Anonymized speech output preserving linguistic content and emotional states.

EXPERIMENT

VCTK

110 speakers, multiple accents, used for phonetic diversity in voice conversion models.

RAVDESS

Audio and video from 24 actors, emotions like happiness and sadness, used for emotion recognition enhancement.

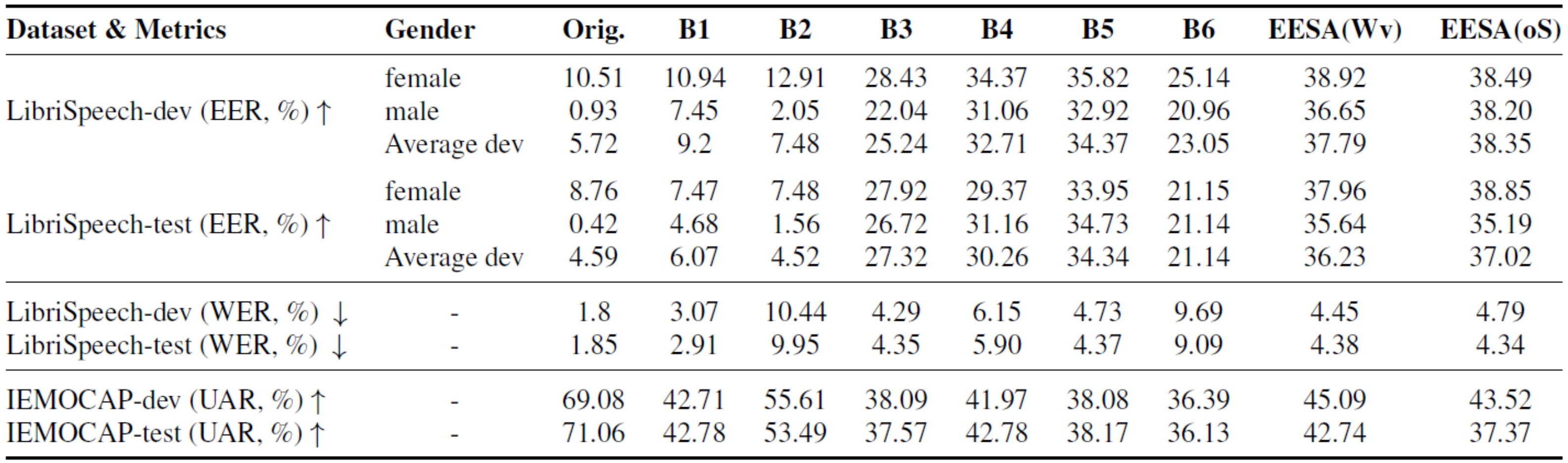
LibriSpeech

1,000 hours of diverse American English, crucial for ASR systems benchmarking.

IEMOCAP

12 hours of audiovisual data, capturing varied emotions, pivotal for emotion detection in HCI.

RESULT



Dataset & Metrics	Gender	Orig.	B1	B2	B3	B4	B5	B6	EESA(Wv)	EESA(oS)
LibriSpeech-dev (EER, %) ↑	female	10.51	10.94	12.91	28.43	34.37	35.82	25.14	38.92	38.49
	male	0.93	7.45	2.05	22.04	31.06	32.92	20.96	36.65	38.20
	Average dev	5.72	9.2	7.48	25.24	32.71	34.37	23.05	37.79	38.35
LibriSpeech-test (EER, %) ↑	female	8.76	7.47	7.48	27.92	29.37	33.95	21.15	37.96	38.85
	male	0.42	4.68	1.56	26.72	31.16	34.73	21.14	35.64	35.19
	Average dev	4.59	6.07	4.52	27.32	30.26	34.34	21.14	36.23	37.02
LibriSpeech-dev (WER, %) ↓	-	1.8	3.07	10.44	4.29	6.15	4.73	9.69	4.45	4.79
LibriSpeech-test (WER, %) ↓	-	1.85	2.91	9.95	4.35	5.90	4.37	9.09	4.38	4.34
IEMOCAP-dev (UAR, %) ↑	-	69.08	42.71	55.61	38.09	41.97	38.08	36.39	45.09	43.52
IEMOCAP-test (UAR, %) ↑	-	71.06	42.78	53.49	37.57	42.78	38.17	36.13	42.74	37.37

FEEDBACK

1. Due to network issues in China, it is not very convenient to directly access Google. Can you add some other ways to upload data?
2. The amount of data to be uploaded is large, and it is easy to fail to upload within the specified time. Can it be optimized?
3. The bit rate of the submitted audio needs to be limited, which does not seem to be stated in the requirements

4. Thank you very much for your patient responses and timely problem solving during the competition

THANKS