System Description for VoicePrivacy Challenge 2024

Tao Tan[1], Shutao Liu[1], Yibo Duan[2], Sheng Zhao[2], Xi Shao[3]
[1]Auditory Intelligence Computing Group(AIC), Nanjing University of Posts and Telecommunications, China
[2]Nanjing Longyuan InformationTechnology, China

Introduction

Speech information:

- Speech data contains a lot of personal information.

- Therefore, different solutions have been proposed to protect the speaker's privacy, and one of the main approaches is speaker anonymization.

age

Identity

race

geographic background

others

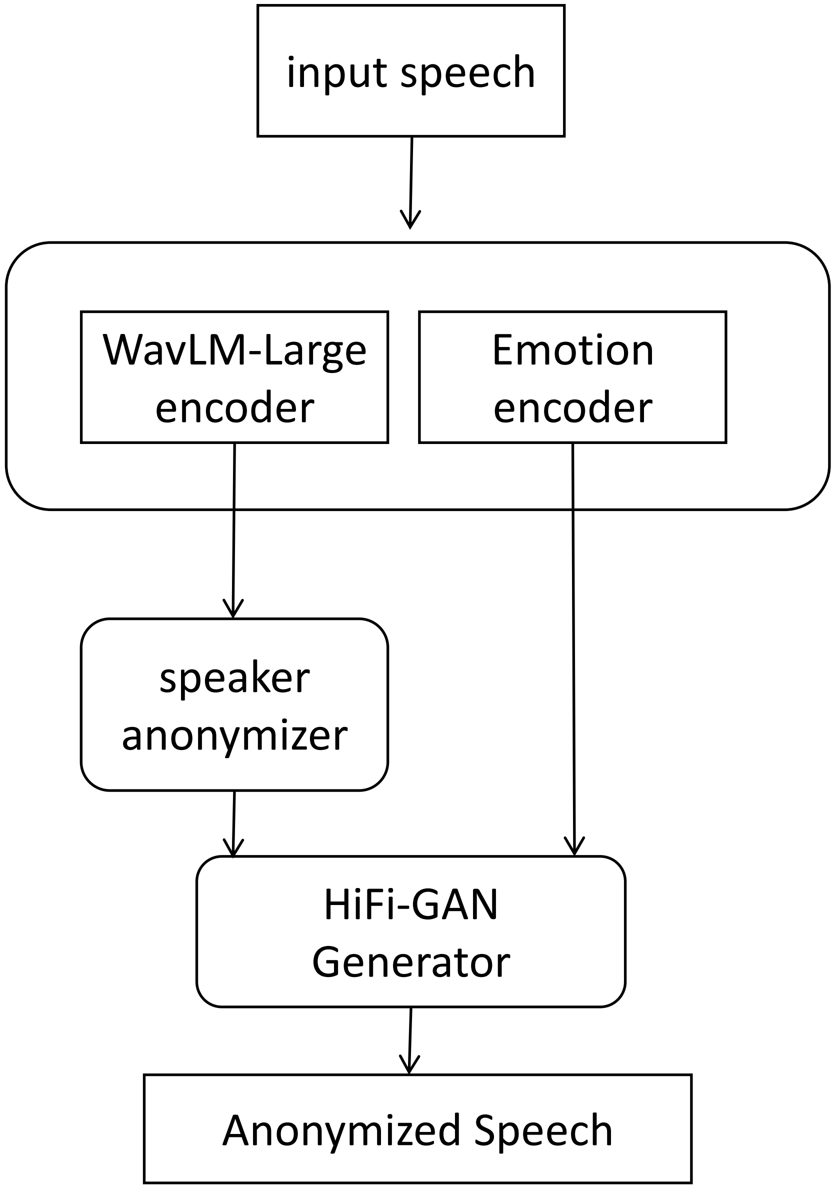
speaker anonymization:

- The task is to develop a voice anonymization system for speech data.

- Specifically, according to the VoicePrivacy 2024 Challenge, the speaker anonymity system needs to satisfy: (i) output a speech waveform; (ii) conceal the speaker identity on the utterance level; (iii) not distort the linguistic and emotional content.

Proposed Method

- System overview
  - Our anonymization system consists of four modules：
  - (a) SSL-based feature extractor
  - (b) Emotional feature extractor
  - (c) Anonymous pools
  - (d) Vocoder



Proposed Method

Methods for building anonymous pools:

- datasets: LibriSpeech train-other-500
- pseudo-speaker:

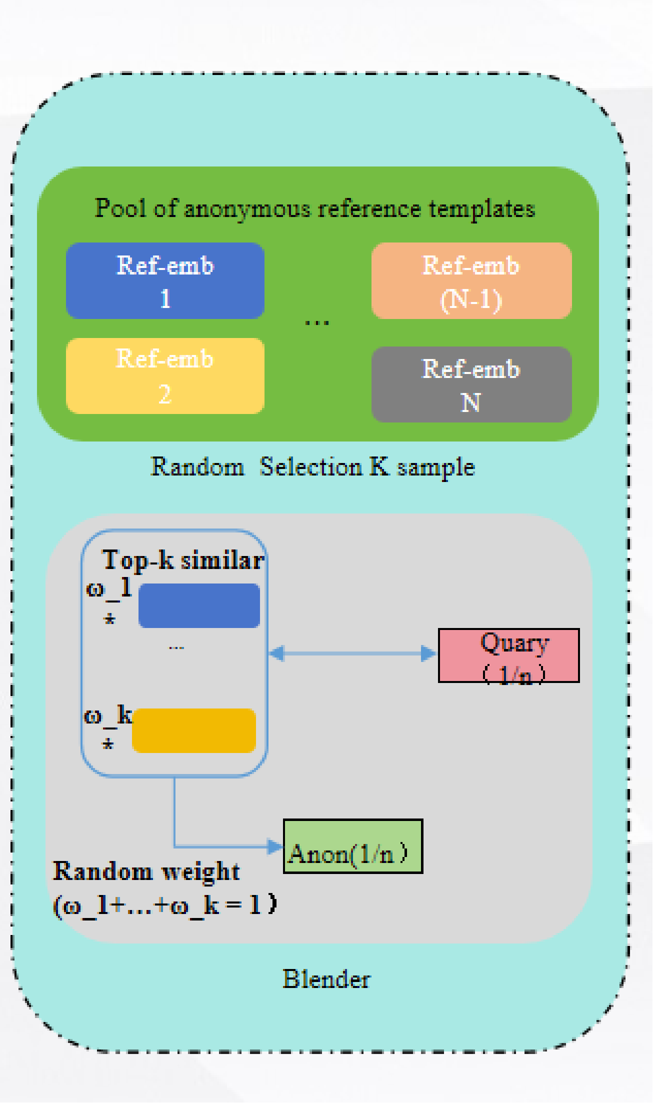
  Randomly select 50 audio files from the dataset to form a pseudo-speaker reference audio.
- Anonymised processes:
  - We adopt the KNN-VC speech transformation method to extract the speech embeddings using the WavLM model and obtain the features of the layer 6 output of WavLM-Large. This step can be described as:

$$D_{spk} = KNN(x, R_{spk}, k)$$

where kNN(x, R, k) means find k nearest vectors to vector x in set R.

In order to generate the final pseudo-speaker, we need to randomly combine the representations of these target speakers. Let the randomly generated speaker weight vector $w = (w_1, w_2, \ldots, w_k)$ and constrain the sum of the weights to 1.

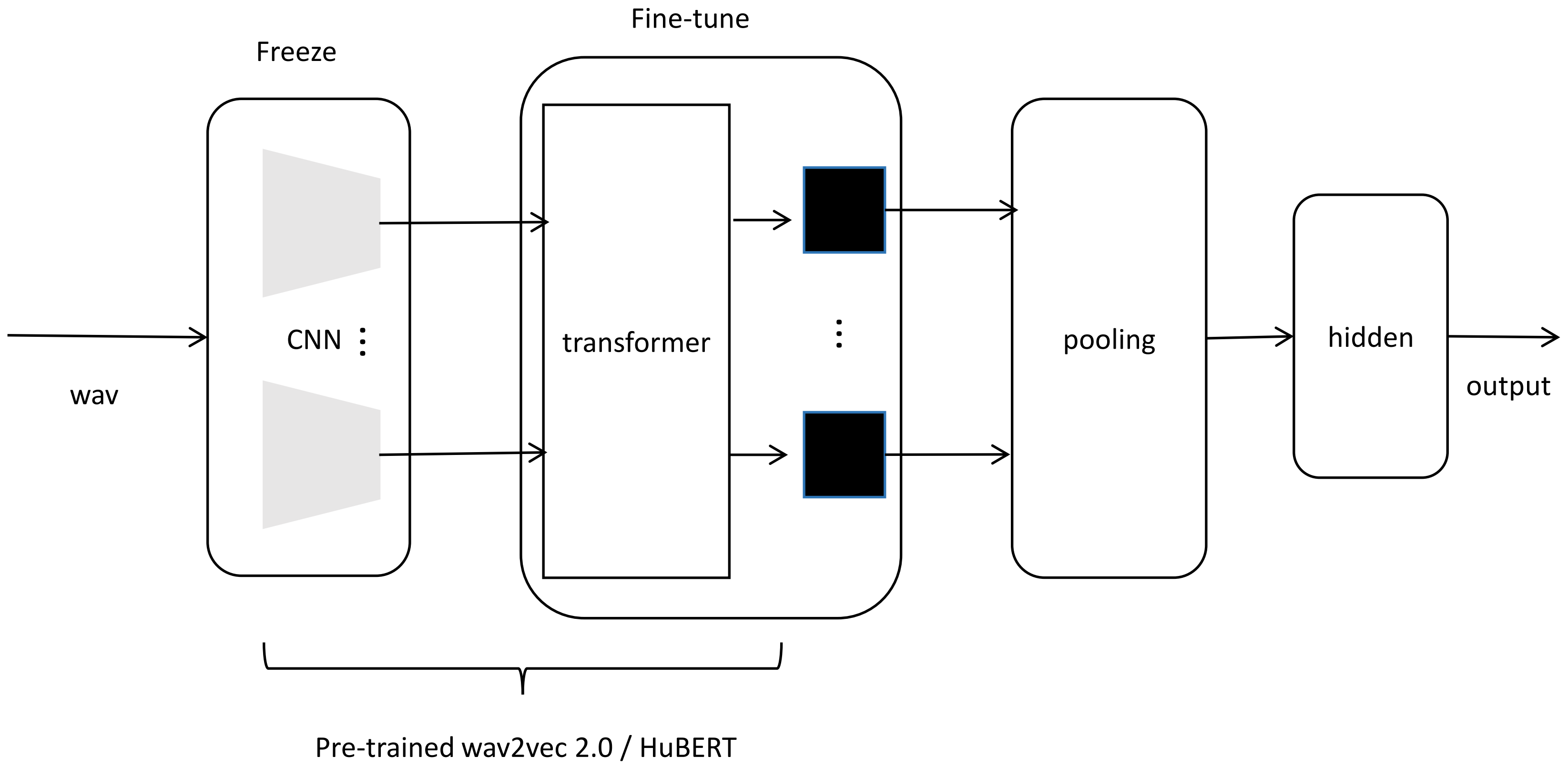
$$D = \sum wD_{spk}$$



Proposed Method

- Emotion encoder:

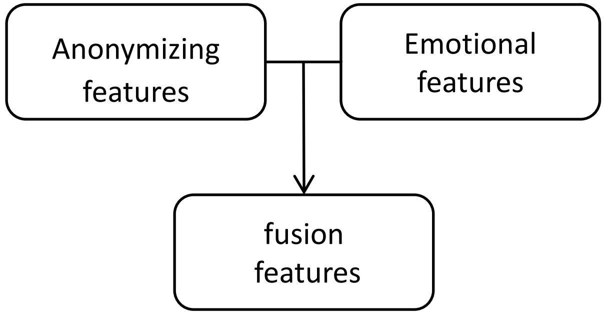
We use wav2vec2-large-robust-12-ft-emotion-ms-dim to extract emotion information, which is a self-supervised learning model that is pre-trained with a large amount of unlabelled speech data and fine-tuned on the labelled data to improve the performance of the emotion recognition task.



Proposed Method

- vocoder: HiFi-GAN

The vocoder translates the converted features into an audio waveform. Instead of conditioning on spectrograms, we adapt a conventional vocoder to take self-supervised features as input.



- Pre-matching training method:

During the training process, the features in the training set are mapped to the nearest neighbor features in other speech segments by k-nearest neighbor regression, and then these “pre-matched” features are used to train the vocoder.

Experiments

- Training datasets: LibriSpeech train-clean-360 and ESD
- Evaluations: Attackers were assumed to have access to the un-anonymized speech and anonymized speech utterances.
- Calculate the assessment metrics (EER, WER, UAR) for the development and assessment sets using the provided scripts.

Results

Table 2: *Privacy results of different systems on the libri-test and libri-dev sets. EER(%) achieved by* ASV_{eval}^{anon} *on data processed by anonymization systems*

Split Gender	dev F	dev M	test F	test M
B1	10.937	7.454	7.474	4.675
B2	12.910	2.045	7.483	1.557
B3	28.426	28.426	28.426	26.724
B4	34.378	31.056	29.378	29.378
B5	**35.816**	32.918	33.496	34.729
B6	25.141	20.961	21.146	21.137
Our	32.671	**34.192**	**34.126**	**36.080**

Results

Table 4: *Comparison of different systems across various emotions in terms of accuracy(%).*

Split	**System**	UAR	SAD	NEU	ANG	HAP
dev	Orig	69.08	63.63	65.97	79.78	66.95
dev	B1	42.71	0.26	34.03	78.88	57.67
dev	B2	55.61	32.96	57.97	64.44	67.09
dev	B3	38.09	0.73	34.45	70.54	46.63
dev	B4	41.97	9.03	41.88	63.22	53.74
dev	B5	38.08	7.54	49.11	62.05	33.62
dev	B6	36.39	2.58	15.25	49.77	77.96
dev	Our	**60.69**	36.99	64.79	75.51	65.45
test	Orig	71.06	72.58	71.66	72.82	67.19
test	B1	42.78	2.78	37.97	72.51	57.85
test	B2	53.49	32.78	66.23	56.97	57.98
test	B3	37.57	0.65	41.83	66.09	41.72
test	B4	42.78	11.26	46.68	61.54	51.64
test	B5	38.17	5.07	55.30	56.20	36.10
test	B6	36.13	1.59	24.49	46.72	71.71
test	Our	**60.95**	40.99	69.79	69.18	63.85

- This result shows that our system outperforms other ranges of scores even if they are not in the same error rate range.

Results

Table 3: *Word Error Rate (WER(%)) comparison of different systems between libri-dev and libri-test sets.*

System	Split	WER	Split	WER
Orig	dev	1.81	test	1.84
B1	dev	3.07	test	2.91
B2	dev	10.44	test	9.95
B3	dev	4.28	test	4.35
B4	dev	6.14	test	5.89
B5	dev	4.73	test	4.37
B6	dev	9.69	test	9.09
Our	dev	**2.33**	test	**2.37**

We compare our model with systems identical to those used in the privacy results, as shown in Table 3. The WER (Word Error Rate) results were computed using the libri-dev and libri-test datasets. Among all the systems evaluated, the system we developed achieved the best WER score.

Conclusions

- In summary, we utilize a self-supervised pre-trained model for the anonymization task.
- Our anonymization system is superior in maintaining both verbal and non-verbal content compared to the baseline system.

THANKS