

System Description: Speaker anonymization system with sentiment transfer and feature interpolation

Tao Tan¹, Shutao Liu¹, Yibo Duan², Sheng Zhao², Xi Shao³

¹Auditory Intelligence Computing Group(AIC), Nanjing University of Posts and Telecommunications, China

²Nanjing Longyuan Information Technology Co.Ltd, China

(1222013926, 1022010103, shaoxi)@njupt.edu.cn, (duanyibo, zhaosheng)@lyxxkj.com.cn

Abstract

This paper introduces a novel speaker anonymization approach that utilizes the WavLM model for robust feature extraction, alongside an advanced blending module and HiFi-GAN for speech synthesis. The proposed methodology is designed to obscure speaker identity while retaining both the linguistic and emotional attributes of the speech. Specifically, our system utilizes the WavLM model to extract speech features from the source speech, termed source query and from the anonymized speech, which are termed anon template. These extracted features represent the characteristics of the original and anonymized speech, respectively. This representation, enriched with the original emotional content, is then fed into the HiFi-GAN model, enabling the synthesis of speech that maintains the emotional nuances of the original recordings. Comprehensive evaluations show that our system not only meets but exceeds several baselines in speaker anonymization, Word Error Rate (WER), and Unweighted Average Recall (UAR) for emotion recognition. The findings confirm the efficacy of our approach in balancing effective privacy protection with the preservation of speech intelligibility and emotional fidelity.

Index Terms: voice privacy, speaker anonymization, voice conversion, feature interpolation, emotional characteristic

1. Introduction

With the rapid development of data mining, machine learning, and deep learning, privacy protection in speech data processing has attracted significant attention from researchers. The Voice Privacy Challenge 2024[1] aims to promote the development of speaker anonymization technologies that protect speaker identity while preserving the linguistic content and other paralinguistic attributes of speech[2].

In recent years, the focus on maintaining privacy in speech technology has led to the development of various anonymization techniques aimed at concealing speaker identity while preserving the utility of the speech. Traditional methods, such as adding noise, using speech synthesis, and employing voice conversion, have been widely explored. However, these approaches often struggle to find a balance between effectively protecting speaker privacy and retaining the intelligibility and quality of speech. The Voice Privacy Challenge[3] has emerged as a critical platform for evaluating these techniques, offering a structured approach to assess their effectiveness in hiding speaker identities, maintaining speech intelligibility, and preserving speech quality.

In alignment with the objectives of the Voice Privacy Challenge 2024, we have developed an innovative speech anonymization system named wav-salf, integrating the WavLM model[4], an emotion-enhanced HiFi-GAN[5], and the

wav2vec2-emotion model[6] for extracting speech emotional features. The system utilizes WavLM to extract raw speech sequence features, termed as query, and combines these with randomly selected features from an anonymization pool using a GMM-blender to produce anonymized feature sequences, referred to as anon. Concurrently, the wav2vec2-emotion model is employed to preserve the emotional features of the speech more effectively. These emotional and anonymized features are then processed by the emotion-enhanced HiFi-GAN, substituting the waveform features typically required by the original HiFi-GAN. This innovative approach enables the HiFi-GAN to synthesize speech that retains the emotional content of the original speech, thus preserving emotional information without compromising the speaker's privacy. This configuration not only addresses the challenge's privacy requirements but also ensures the retention of vital paralinguistic information, setting a new standard in the field of voice anonymization.

2. Method

The overall structure of the submitted anonymization system can be described using Figure 1. The system utilizes the WavLM model to extract speech features and employs a method called GMM-Blender to generate a rich array of anonymous templates from the anonymization pool. To preserve emotions, Wav2vec2-Emotion extraction module has been incorporated into the HiFi-GAN model. This configuration enables the synthesized speech to effectively conceal the identity of the target speaker while simultaneously ensuring the preservation of the speaker's emotional and textual information.

2.1. Feature extractor

Our model primarily utilizes the Gaussian Mixture Model (GMM) blender module[7] and the Modified HiFi-GAN module for anonymization purposes. However, it is important to discuss the feature extraction mechanisms employed. For emotion feature extraction, we used a pretrained Wav2vec2-Emotion model, and for speech feature extraction, we employed WavLM.

2.1.1. Wav2vec2-Emotion

To efficiently extract emotion features, we employ the wav2vec2-large-robust-12-ft-emotion-ms-dim model, which is included in the official roster of sanctioned pre-trained models for this task. The emotion vector extracted by it can well preserve the emotional features in the speech signal. This is particularly critical in ensuring the preservation of emotional information during the speech anonymization process, allowing the system to protect the identity of the speaker while maintaining the richness of emotional content in the speech.

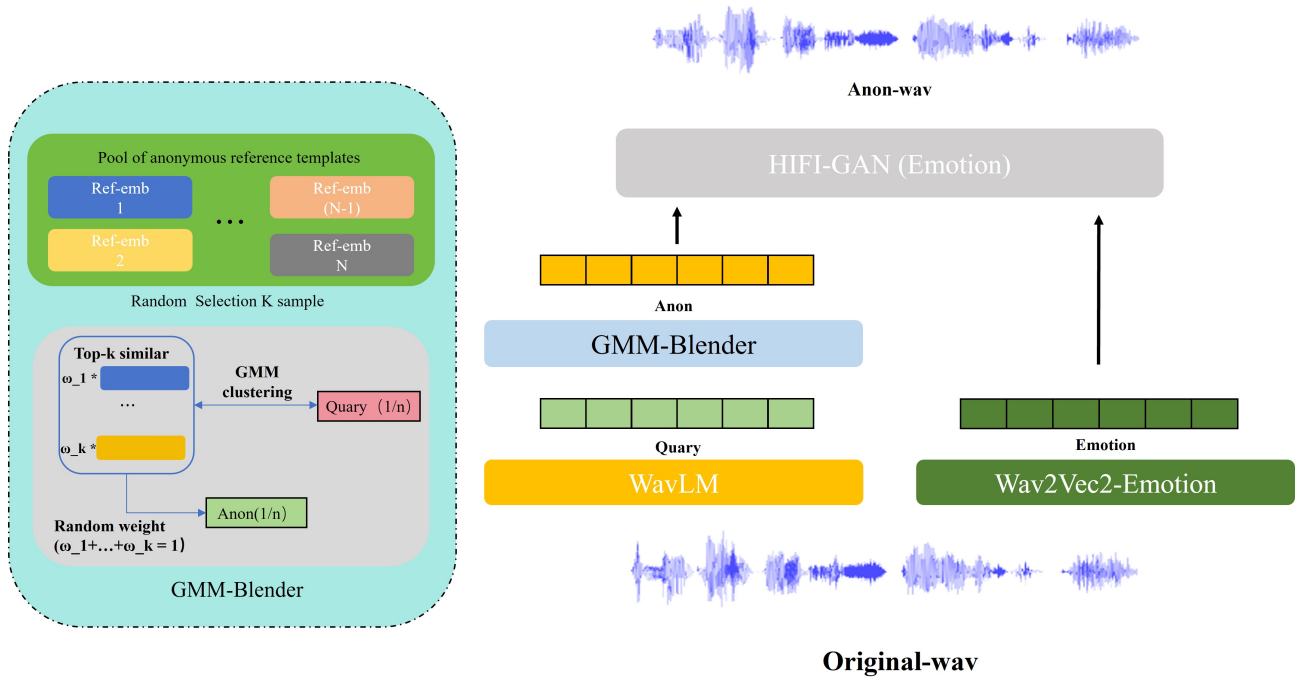


Figure 1: Architecture of the proposed Anonymization System: This schematic illustrates the overall structure of the anonymization system, highlighting the implementation process of the GMM-Blender, as depicted by the dashed box on the left.

2.1.2. WavLM

we employ the WavLM model, which is approved for use according to the official documentation. Developed through large-scale, domain-agnostic, self-supervised learning on a vast corpus of unsupervised speech data, WavLM is adept at extracting a comprehensive array of speech features. This encompasses not only speaker identity traits but also textual characteristics embedded within the speech signals. The extensive training dataset empowers WavLM to capture nuanced speech features, enhancing its versatility across various speech processing applications. The WavLM model has yielded promising outcomes in downstream tasks such as speaker verification (SV)[8] and automatic speech recognition (ASR), prompting us to adopt it as our speech feature extractor.

2.2. GMM Blender

This module is crucial for ensuring the anonymization performance of our model. Our experiments with various methods to generate anonymous pools revealed that larger pools will produce higher-quality speech. And a concise but more varied pool will enhance the performance of anonymization effectiveness. This improvement in speech quality is beneficial for reducing the Word Error Rate (WER) and enhancing the auditory perception of speech. To generate an anonymous pool that is sufficiently extensive and distinct from the original speaker’s voice, we implemented the procedure illustrated in figure 1.

Our anonymous pool is derived from speakers in the Lib-rispeech dataset. Each time anonymization is required, we randomly select M speakers from this pool, who then serve as the basis for the anonymization process. After applying Gaussian Mixture Model (GMM) clustering[9] to these selected speakers. We then select 4 samples that exhibit the smallest distances be-

tween these 2000 cluster centers. The speech from the speakers at each selected cluster center is mixed to create a composite cluster center, which is subsequently utilized as a speech synthesis template. Our steps are as follows:

Every time the system initiates an anonymization request, we generate an anonymous pool. Initially, we randomly select M speakers from the entire dataset. For each speech to be anonymized, we then randomly pick 4 speaker from the pool to provide speech embeddings. From each selected speaker, we randomly choose 50 speech. These speech constitute the initial anonymous pool, each speech is represented by a_i .

$$A = a_1, a_2, a_3, \dots, a_i, \dots, a_N, i \in [1, 200]$$

Then we extract the speech embedding using WavLM model:

$$e_i = WavLM(a_i), i \in [1, 200]$$

$$E_i = \{e_1, e_2, \dots, e_j, \dots, e_l\}, j \in [1, l_i]$$

$$\mathbb{E} = \{E_1, E_2, E_3, \dots, E_{200}\}$$

Where $E_i \in \mathbb{R}^{l_i \times 1024}$, l_i is the length of speech embedding, We assume the total length of the embeddings is L , where $L = \sum l_i$. \mathbb{E} stands for the total assembly of all the speech embedding extracted by the WavLM model. For the subsequent Gaussian Mixture Model (GMM) clustering operation, we process the embeddings of each speech clip E_i by concatenating them along the time dimension, to form a speech embedding matrix, which shape is $L \times 1024$.

Subsequently, we will perform clustering to obtain 2000 cluster centers, denoted as $C = \{c_1, c_2, \dots, c_j, \dots, c_{2000}\}$, along with the sample ec_i^j in each class. Then, we selected the four features nearest from the cluster center for these 2000 classes

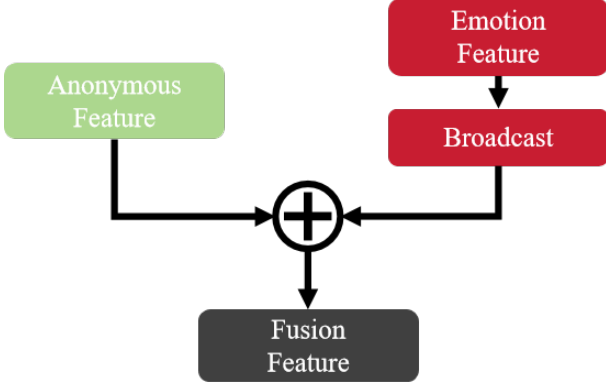


Figure 2: Feature fusion method: The emotion features are broadcast into the speech feature length and then they are added together

and operated on them. Each sample ec_i^j is divided into 32 segments and added together with other samples' segments according to randomly assigned weights to acquire a new clustering center:

$$c_{j,m} = \sum_{n=1}^4 w_{n,m}^j \cdot ec_{n,m}^j$$

Where $w_{n,m}^j$ denotes the random weight of the j th cluster center, the n th sample and the m th segment, as well as $ec_{n,m}^j$. A limitation is added such that $\sum_{n=1}^4 w_{n,m}^j = 1$.

After completing the aforementioned procedure, we obtained 2000 speech embeddings, referred to as 'anon presentations'. These embeddings will be utilized as synthesis templates for the target speech.

Then, we will match each speech feature to be anonymized with a feature template, and use the nearest template as the anonymization result for that frame.

2.3. Modified HiFi-GAN

HiFi-GAN[5] is an adversarial generative network that can transform input speech features into speech signals that humans can understand. This model originally takes mel-spectrograms as input. For better performance, we used speech features extracted from WavLM, a pretrained model as mentioned before.

To enhance the model's ability to synthesize speech emotions, an emotion vector was incorporated into the input of HiFi-GAN. This emotion vector was derived from the previously mentioned wav2vec variant. To fuse the input speech features with the emotion vector features, we employed the synthesis method as shown in the following figure 2.

In addition to utilizing a fused input of speech and emotional features, we also implemented minor adjustments to the loss function. We input the generated speech into an emotional feature extractor, a network designed to capture the input emotional characteristics, to derive emotional features. We compare the synthesized speech emotion vector with the emotional vector of the original speech, thereby compelling the model to produce the same emotion in response to the input. The loss function we employ is structured as follows:

$$Loss = L_{HiFi-GAN} + \lambda \cos_sim(emo_{orig}, emo_{gen})$$

Where emo_{orig} and emo_{gen} represent the emotional characteristics of the input speech and the emotional characteristics of

the synthesized speech, respectively. And the parameter λ controls the weight of the emotional loss function. $L_{HiFi-GAN}$ is the loss function of original HiFi-GAN model, which helps the model to generate a speech close to the original speech.

3. Experiment

We implemented our method using PyTorch and conducted experiments on x86 Linux machines equipped with 2 NVIDIA 3090 GPUs.

3.1. Datasets

3.1.1. Training Set

For training our model, we utilize two officially sanctioned datasets: LibriSpeech and the ESD[10]. LibriSpeech[11], renowned for its diverse range of speakers and speaking styles, is employed not only for training but also for constructing an anonymization pool using the train-other-500 and train-clean-360 subset. This ensures a robust adaptation of our model to a variety of speech patterns and linguistic contents. The ESD is specifically leveraged to enhance our model's capability in handling and synthesizing emotional speech.

3.1.2. Testing Set

Our testing framework is designed to evaluate the model across two main dimensions: emotion recognition and speech recognition accuracy. For emotion recognition, we use the development and test sets from the IEMOCAP dataset[12], which is specifically tailored for emotion analysis in spoken content. This dataset provides a rigorous benchmark for assessing the effectiveness of our model in maintaining emotional attributes post-anonymization. For assessing speech recognition accuracy, we employ the dev and test sets from LibriSpeech, referred to as libri-dev and libri-test.

3.2. Model Setup

Our system employs architectural details as outlined in the Table 1 below.

3.2.1. speech feature extractor

The two speech feature extractors used to extract original-wav and ref-wav are both WavLM models. They share the same structure and weights. We used the speech features output by the official WavLM-large pre-trained model at layer 6. WavLM model takes raw speech waves as input and output a matrix as speech embedding. We adopt the setting of WavLM-large which set each frame of speech embedding to be 1024.

3.2.2. Emotion feature extractor

As mentioned before we utilized the wav2vec2-emotion model fine-tuned on the MS-podcast as an emotion feature extractor. It extracts 1024-dimensional vectors for each speech input to describe emotional features.

3.2.3. GMM-Blender

As the core step of our proposed method, we experimented with multiple parameters. After weighing the impact of various parameters, we have set the following parameters:

Table 1: *Modules and training corpora for proposed system*

index	Module	Description	Output features	Data
1	speech Feature Extractor	WavLM pre-trained model[4]: Input: Raw speech waveform Output: WavLM latent speech feature from layer No.6n	Original speech versus anonymous template speech features	Pre-trained model
2	Emotion Feature Extractor	Wav2vec2-large-robust-12-ft-emotion-ms-dim model Input: Raw speech waveform Output: Latent speech feature from the last layer	Raw speech Emotion Features	Pre-trained model (Finetuned on MSP-Podcast corpus[13])
3	GMM-Blender	GMM Input:Anonymous pool randomly selected anonymous templates and source speech. Output:Anonymous pools and blend the features of each frame of speech according to randomly generated weights.	Blended speech frame embeddings	LibriSpeech:train-clean-360
4	Vocoder	Modified HiFi-GAN Input:Blended speech frame embeddings and Raw speech Emotion Features. Anonymous speech features are fused with emotion features.	Anon speech waveform	ESD, LibriSpeech:train-clean-360, train-other-500.

1. For anonymous pools, we select 4 speakers to form an anonymous pool, and then we select four speakers for each speech.
2. For the GMM clustering method, we selected 2000 clustering centers to ensure sufficient speech embeddings for the HiFi GAN module to reference. We select 4 speech frame embeddings near each cluster center (if not enough, select all) to ensure a certain level of randomness. Our mixed weight matrix adopts a random selection method based on reference time.

3.2.4. HiFi-GAN

For the HiFi-GAN model, we used its original structure without making any changes except for adding a feature aggregation operation before feature input. For the weight of the loss function, we set it to 0 in the initial 30 rounds and gradually increase it to 1 in the following 30 rounds, then keep it unchanged until the end of training. During training, we set the lr and batch sizes to 1e-3 and 64, respectively, using the Adam optimizer and warmupscheduler.

Further architectural details of the implementation are provided in Table 1

3.3. Inference

Before conducting actual anonymization, we have extracted all the features of the reference speech. In this way, there is no need to extract features from the speech in the anonymous pool during actual anonymization, which greatly reduces the time and computational power required for inference. But for the anonymized speech that will be generated in real time, we have to feed it into two feature extractors for feature extraction. For speech to be anonymized over a period of time, we use the same anonymous pool. That is to say, an anonymous pool will only be used for a specific period of time. We set this time to 20 minutes. During these 20 minutes, all the anonymized features we obtained are stored in memory without the need for additional

calculations. Simply select the cluster center that is closest to the feature to be anonymized.

4. Evaluations and results

4.1. Evaluations

The Voice Privacy 2024 Challenge focuses on the retention of emotional states, which is evaluated objectively. Three complementary metrics are used: the Equal Error Rate (EER) for privacy metrics and two utility metrics, the Word Error Rate (WER) for Automatic Speech Recognition and the Unweighted Average Recall (UAR) for Speech Emotion Recognition. To weigh the privacy utility, the results are divided into four intervals based on the minimum target EERs. Within each EER interval – [10,20), [20,30), [30,40), [40,100) – systems will be ranked separately in order of (1) increasing WER and (2) decreasing UAR. The most critical capability of an anonymization system is to perform anonymization, and we used the ASV method for verification. The ability of the anonymization system to keep linguistic content undistorted was evaluated using the ASR system, which was fine-tuned from wav2vec2-large-960h-lv60-self [14] using the SpeechBrain[15] recipe on LibriSpeech-train-960. This ASR evaluation model is fixed and trained and fine-tuned based on raw data.

4.2. Results

4.2.1. ASV evaluation

The privacy results are presented in Table 2, where the outcomes for baseline systems 1, 2, 5, and 6 are denoted as B1, B2, B5, and B6, respectively, while the results from our system are labeled as 'Ours'. All EER (Equal Error Rate) results were computed using anno-enroll and anno-trial datasets. As evident from the table, our system's results surpass those of B5 in all cases except for female speakers on the libri-dev set. Compared

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	B1	B2	B5	B6	Ours
dev	F	10.937	12.91	35.816	25.141	32.671
dev	M	7.454	2.05	32.918	20.961	34.192
test	F	7.474	7.48	33.496	21.146	34.126
test	M	4.675	1.56	34.729	21.137	36.080

to other baseline systems, our system demonstrates a substantial improvement. This significant discrepancy underscores the effectiveness of our anonymization approach, which can effectively obliterate speaker-specific information.

4.2.2. WER evaluation

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
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.

4.2.3. UAR evaluation

The effectiveness of our anonymization system in preserving emotional content without distortion was assessed using a SER (Speech Emotion Recognition) system. This system was developed using the SpeechBrain recipe for SER on the IEMOCAP dataset, utilizing a wav2vec2-based model.

As indicated in Table 4, when compared to the baseline system B5—which also falls within the [30-40] interval of Equal Error Rate—our system demonstrates notable improvements in Word Error Rate (WER) and sentiment retention capabilities on both development and test datasets.

Table 4 clearly demonstrates that after four verification rounds, our system achieved the highest Unweighted Average Recall (UAR) score. Notably, this score surpasses that of the B5 system, which is within the same error rate range, and also outperforms scores in other ranges. The data reveal that our anonymization system is particularly effective in retaining the emotional features associated with sadness, resulting in superior UAR scores compared to other baseline systems. This underscores the system’s robust ability to preserve emotional content. We anticipate that the adoption of an enhanced feature fusion method could lead to further improvements in performance.

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	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	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

5. Conclusion

Our proposed speech anonymization model exhibits robust performance, effectively preserving speaker privacy while maintaining high utility for Automatic Speech Recognition (ASR) and Speech Emotion Recognition (SER). This balance between anonymization and utility is highlighted by our results in the Voice Privacy 2024 Challenge.

During the Automatic Speaker Verification (ASV) evaluation, our system achieved a competitive Equal Error Rate (EER), thereby effectively anonymizing speaker information. Furthermore, the model’s Word Error Rate (WER) performance exceeded that of the baseline systems, indicating minimal distortion in linguistic content. Additionally, the model demonstrated significant capability in retaining emotional content, as reflected by high Unweighted Average Recall (UAR) scores across various emotional states.

These outcomes validate our methodological approach involving sophisticated feature extraction, blending, and anonymization techniques, aligning with the challenge’s objectives. Looking forward, we plan to refine our feature fusion methods and explore more efficient anonymization techniques to further enhance performance.

6. References

- [1] N. Tomashenko, X. Miao, P. Champion, S. Meyer, X. Wang, E. Vincent, M. Panariello, N. Evans, J. Yamagishi, and M. Todisco, “The voiceprivacy 2024 challenge evaluation plan,” *arXiv preprint arXiv:2404.02677*, 2024.
- [2] A. S. Shamsabadi, B. M. L. Srivastava, A. Bellet, N. Vauquier, E. Vincent, M. Maouche, M. Tommasi, and N. Papernot, “Differentially private speaker anonymization,” *arXiv preprint arXiv:2202.11823*, 2022.
- [3] N. Tomashenko, X. Wang, E. Vincent, J. Patino, B. M. L. Srivastava, P.-G. Noé, A. Nautsch, N. Evans, J. Yamagishi, B. O’Brien *et al.*, “The voiceprivacy 2020 challenge: Results and findings,” *Computer Speech & Language*, vol. 74, p. 101362, 2022.
- [4] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [5] J. Kong, J. Kim, and J. Bae, “Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis,” *Advances in neural information processing systems*, vol. 33, pp. 17022–17033, 2020.

- [6] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [7] M. Baas, B. van Niekerk, and H. Kamper, "Voice conversion with just nearest neighbors," *arXiv preprint arXiv:2305.18975*, 2023.
- [8] V. Miara, T. Lepage, and R. Dehak, "Towards supervised performance on speaker verification with self-supervised learning by leveraging large-scale asr models," *arXiv preprint arXiv:2406.02285*, 2024.
- [9] L. Jiao, T. Denœux, Z.-g. Liu, and Q. Pan, "Egmm: An evidential version of the gaussian mixture model for clustering," *Applied Soft Computing*, vol. 129, p. 109619, 2022.
- [10] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, vol. 137, pp. 1–18, 2022.
- [11] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.
- [12] S. Tripathi, S. Tripathi, and H. Beigi, "Multi-modal emotion recognition on iemocap dataset using deep learning," *arXiv preprint arXiv:1804.05788*, 2018.
- [13] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017.
- [14] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.
- [15] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong *et al.*, "Speechbrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.