

PRIVACY VERSUS EMOTION PRESERVATION TRADE-OFFS IN EMOTION-PRESERVING SPEAKER ANONYMIZATION

Zexin Cai, Henry Li Xinyuan, Ashi Garg, Leibny Paola García-Perera, Kevin Duh,
Sanjeev Khudanpur, Nicholas Andrews, Matthew Wiesner

Human Language Technology Center of Excellence, Johns Hopkins University

ABSTRACT

Advances in speech technology now allow unprecedented access to personally identifiable information through speech. To protect such information, the differential privacy field has explored ways to anonymize speech while preserving its utility, including linguistic and paralinguistic aspects. However, anonymizing speech while maintaining emotional state remains challenging. We explore this problem in the context of the VoicePrivacy 2024 challenge. Specifically, we developed various speaker anonymization pipelines and find that approaches either excel at anonymization or preserving emotion state, but not both simultaneously. Achieving both would require an in-domain emotion recognizer. Additionally, we found that it is feasible to train a semi-effective speaker verification system using only emotion representations, demonstrating the challenge of separating these two modalities.

Index Terms— voice privacy, emotion recognition, speaker verification, speech anonymization, voice conversion, speech synthesis

1. INTRODUCTION

Voice-driven interaction has been integrated into various aspects of human life, making tasks more convenient and hands-free. This technology has seen significant growth in the modern era, with notable examples including virtual assistants on smart devices, wearable technology, and customer service applications. However, the increasing use of voice-driven interaction raises security and privacy concerns, particularly regarding the exposure of speech recordings to fraudsters and hackers when transmitted over untrusted public networks [1]. Consequently, the personally identifiable information in the raw speech signal can be susceptible to leakage or extraction [2].

To mitigate privacy concerns associated with the potential interception and misuse of speech data, speech anonymization is employed to protect the most sensitive information, speaker identity, within speech. Specifically, speech anonymization aims to suppress acoustic characteristics that could be used to identify the speaker while at the same time preserving other characteristics, chiefly linguistic content, within the speech. The field of speech anonymization is still nascent, with formal definitions and a comparison platform for solutions on standardized datasets and protocols recently established by the VoicePrivacy Challenge series [3].

Since speech anonymization inherently involves altering and transforming speech, most research has centered on techniques such as voice conversion (VC), speech synthesis, noise addition, and traditional signal processing methods to achieve anonymization [3]. Among the developed anonymization techniques, the x-vector-based method [4], used as the baseline for the VoicePrivacy challenge, offers a flexible choice of pseudo-speaker and achieves adequate performance in privacy and utility assessments. Essentially, the

x-vector-based method employs a framework similar to an any-to-any VC approach, synthesizing anonymized speech by conditioning the framework with x-vector [5] speaker representations to produce pseudo-speakers' voices. Several subsequent studies have improved the x-vector-based method from various angles to boost its privacy protection ability [6], such as constructing x-vectors via singular value modification [7] and using a generative model to sample pseudo-speakers in the x-vector space [8]. Beyond the approaches described above that achieve speech anonymization through acoustic models like those used in VC techniques, other research explores a speech synthesis-based method by cascading automatic speech recognition (ASR) and text-to-speech (TTS) systems [9, 10], which can significantly eliminate speaker identity footprints in speech.

Recent developments in the speech anonymization community have presented a more complex anonymization scenario. Besides preserving linguistic content and hiding speaker identity, an anonymization system should also maintain unchanged paralinguistic attributes [11]. Under these conditions, researchers struggle to conceal speaker identity while retaining paralinguistic attributes, highlighting the trade-off between utility (paralinguistic attributes) and privacy (speaker identification) in this setting [2]. The VoicePrivacy Challenge 2024 emphasizes the preservation of emotional state [12]. Additionally, the challenge recognizes the risk that an attacker could access anonymized data and train a new speaker verification model on it. Therefore, understanding how emotion and speaker information are entangled in speech signals is essential in this anonymization context to overcome the privacy-utility trade-off.

Earlier work on anonymization has explored the privacy-utility trade-off, but does not investigate its causes or potential solutions [2]. Aside from voice timbre, prosodic features such as melody, rhythm, and intensity—shaped by a speaker's social environment and critical learning period—also provide significant information about their identity [13]. This theory is supported by findings that the source speaker can be recognized to a certain degree in voice-converted speech [14]. Inspired by the above research, this paper delves into the factors that might cause the leakage of speaker identity and investigates the relationship between speaker and emotion in speech. To achieve this, we apply various VC-based and cascaded ASR-TTS methods to the anonymization task in the VoicePrivacy Challenge 2024. Our study reveals that speech emotion recognition (SER) and automatic speaker verification (ASV) systems rely on overlapping speech attributes. Disentangling identity from acoustic properties is a non-trivial task, as these properties are closely related. While we can minimize the trade-off between privacy and emotion preservation given prior knowledge of the corresponding in-domain emotion recognizer, the challenge of separating speaker and emotion information in speech remains significant. Finally, our results suggest that emotion recognizers can serve as a reliable objective evaluation metric in emotional speech synthesis.

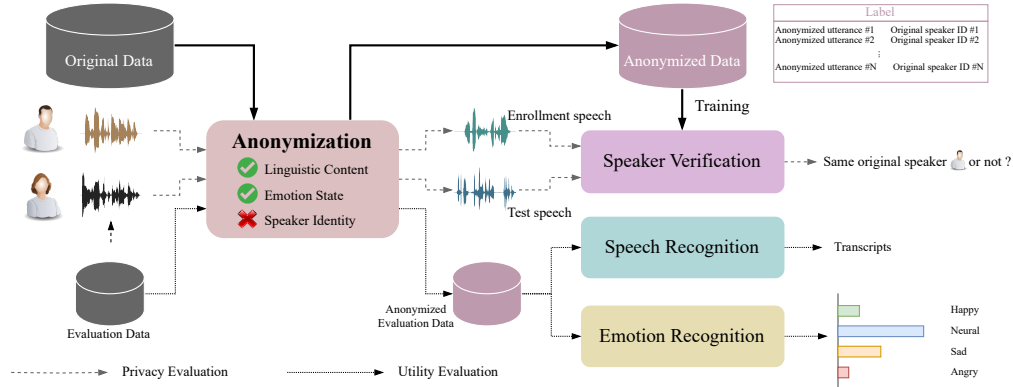


Fig. 1: Speech anonymization task and evaluation pipeline (w.r.t the VoicePrivacy 2024 Challenge)

2. METHOD

2.1. Task Definition and Evaluation Metrics

In the semi-informed speech anonymization task a user supplies speech data and attempts to protect their identity using a speech anonymization system. An attacker, who has access to the anonymized data, attempts to discover the speaker’s identity. A speech anonymization system is created to obscure the user’s identity while maintaining the linguistic content and paralinguistic attributes. In the VoicePrivacy Challenge 2024, the goal is to maintain the emotional state of the speech post-anonymization. As such, the anonymization performance is evaluated from two fronts: privacy evaluation and utility evaluation. The privacy metric measures how well the system conceals the original speakers’ identities, whereas the utility metrics access the retention of content and emotional state.

Figure 1 illustrates the anonymization and evaluation pipeline of this task. The core element is the anonymization module, which converts each original input audio into anonymized audio under the following conditions: 1. preserving linguistic content, 2. preserving emotional state, and 3. removing speaker identity information.

The privacy evaluation pipeline adheres to a standard speaker verification process. The verification model is trained on anonymized data labeled with the original speakers’ identities. An effective anonymization system should sufficiently distort and obscure the original identities at the waveform level, preventing the speaker verification system from identifying different speakers. During evaluation, pairs of source speech from the evaluation dataset are anonymized and treated as enrollment and test speech. The speaker verification model then assesses whether the two utterances originate from the same original speaker. With a perfect anonymization system, the verification system, acting as the attacker, performs no better than random guessing. The main metric for privacy evaluation is the equal error rate (EER), calculated based on similarity scores from pairs of utterances in the anonymized evaluation set, known as trials. A lower EER indicates a greater risk of speaker re-identification, thus a higher EER indicates better performance in preserving voice privacy.

For utility evaluation, the anonymized evaluation data is transcribed using a speech recognition system. The performance in preserving content is assessed by comparing these transcripts to the ground truth content from the source data and measuring the word error rate (WER). Similarly, an emotion recognizer is employed on the anonymized data to determine the emotional state of the anonymized speech. In this case, four emotion states—Happy, Neutral, Sad,

and Angry—are evaluated. An anonymization system demonstrates good emotion preservation performance if the emotional state of the anonymized speech matches that of the original speech. Preservation of emotion state is measured using the Unweighted average recall (UAR) [12]. In general, A lower WER denotes superior preservation of linguistic content, whereas a higher UAR indicates superior preservation of emotion states.

2.2. Anonymization Approaches

We employ two primary synthesis approaches to achieve anonymization for the described task. The speech anonymization process is shown in Figure 2, where one method is based on voice conversion (VC) models, and the other employs a cascaded ASR-TTS pipeline.

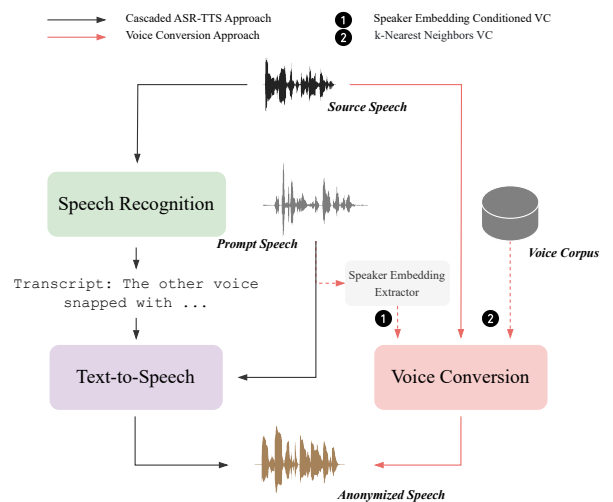


Fig. 2: VC-based and cascaded ASR-TTS anonymization process

VC is a method that changes the voice of the source speech to match that of a target speaker, preserving the content and most prosodic features. This technique aligns closely with the objectives of the anonymization task. We explore two VC-based systems that can convert source utterances to a variety of target speakers.

The first VC-based anonymization system, similar to the x-vector-based system from the VoicePrivacy Challenge, conditions on speaker representations to convert the source voice to a target

speaker’s voice. The model uses the content representation extracted by a pre-trained self-supervised learning (SSL) model called ContentVec [15] as input. This approach uses a transformer-based VC system [16] to convert the input content representation into the target speaker’s Mel-spectrogram by conditioning on the target speaker’s representation vector. The audio waveform is then reconstructed from the Mel-spectrogram using a HiFi-GAN vocoder [17].

Another VC-based solution utilizes kNN VC [18], functioning at the WavLM-feature [19] level. The kNN-VC system maps the WavLM features of the source utterance to those of the target speaker using k-nearest neighbor regression. Each frame from the source speech is replaced by the average of the k-nearest neighbor target WavLM features, followed by a HiFi-GAN vocoder to synthesize the target utterance. This approach, unlike the previous VC method, necessitates a target speech corpus for conversion.

While VC-based systems can effectively modify the acoustic characteristics related to the timbre of source speakers, certain prosodic features, reflecting the speakers’ habitual speaking styles, remain unchanged. These prosodic features could be used to identify the speaker. To mitigate this, we employ a cascaded ASR-TTS pipeline to enhance anonymization by modifying the speaking style of the source speech. As shown in Figure 2, we first transcribe the source utterance using an ASR system. Subsequently, a multi-speaker TTS system generates the anonymized utterance, cloning the voice and speaking style from a prompt utterance.

3. EXPERIMENTS

In this section, we explore the relationship between emotion and speaker information using the anonymization pipeline from the VoicePrivacy 2024 challenge, focusing on English corpora. We anonymized datasets using the systems in Section 2.2 and evaluated their privacy and utility performance. For the cascaded approach, we used various utterance prompts, from randomized speech to audio containing some original speaking style, to study the impact on speaker identity exposure. After identifying the trade-off between privacy and emotion preservation, we explored strategies to mitigate it. Additionally, we analyzed the extraction of speaker information from emotion embeddings to understand their overlap and the challenges in disentangling them.

3.1. Dataset

The VoicePrivacy 2024 challenge uses subsets from the LibriSpeech [20] and IEMOCAP [21] corpora for development and evaluation. More details can be found in the data description section of the challenge’s evaluation plan [12]. There are 10 subsets specifically designated for the evaluation process. The subsets `libri-dev-asr` and `libri-test-asr` are used for ASR evaluation. The subsets `libri-dev-enrolls`, `libri-dev-trials-f`, `libri-dev-trials-m`, `libri-test-enrolls`, `libri-test-trials-f`, and `libri-test-trials-m` are used for evaluating privacy (speaker verification) performance. The subset `libri-train-clean-360` is employed for training the speaker verification system following anonymization. For emotion preservation performance, the subsets `IEMOCAP-dev` and `IEMOCAP-test` are utilized.

We also incorporate the LibriTTS [22] speech synthesis dataset in our experiments. This dataset comprises 585 hours of clean speech data at 24kHz from 2,456 speakers. We also utilize the VoxCeleb1 [23] dataset in our study, with the training data including 148,642 utterances from 1,211 speakers and the test set comprising 4,874 utterances from 40 speakers. There is no overlap between the training and testing data in both the source and target datasets.

3.2. Experimental Details

We train the speaker embedding-conditioned VC model outlined in Section 2.2 using the LibriTTS training sets. Content features are extracted with the pre-trained ContentVec.legacy-500 model,¹ and speaker embeddings are obtained from an ECAPA-TDNN model [24] by SpeechBrain [25]. The VC conversion system, including the feature transformation and vocoder modules, is trained on audio recordings at a 24kHz sample rate. After anonymization, we downsample the synthesized audio to 16kHz for evaluation. For the kNN-VC method, k is set to 4.

In the cascaded ASR-TTS approach, we employ the ‘medium-en’ model from Whisper² [26] as our ASR system to transcribe the source utterance. The Whisper model achieves a WER of 3.38% on the `libri-dev-asr` set and 3.29% on the `libri-test-asr` set. For the study of privacy and emotion preservation, we choose the open-source synthesis model XTTS,³ which is a generative TTS model providing high-fidelity synthesis and capable of voice and style cloning based on a prompt audio segment.

3.3. Anonymization Performance

We anonymized the utterances from the datasets selected by the VoicePrivacy challenge⁴ using various approaches detailed in Section 2.2. The anonymization performance of different systems is summarized in Table 1, with the corresponding systems annotated as follows:

- **Origin:** The original, unanonymized speech.
- **ConVec2Mel-VC:** The speaker embedding-conditioned VC system we developed. During anonymization, the target embedding is extracted from a randomly selected utterance from LibriTTS.
- **kNN-VC:** The kNN-based VC method. For each utterance, the WavLM feature pool is obtained from a randomly chosen target speaker in LibriTTS, with the target pool comprising at least 5 minutes of audio.
- **ConVec2Mel-VC-XTTS:** This system utilizes the cascaded ASR-TTS method. For each source utterance, the prompt speech is the corresponding anonymized speech from ConVec2Mel-VC.
- **kNN-VC-XTTS:** This system follows the cascaded ASR-TTS approach. For each source utterance, the prompt speech is the corresponding anonymized speech from the kNN-VC system.
- **XTTS:** The cascaded ASR-TTS anonymization method, where the prompt utterance during inference is randomly chosen from the LibriTTS dataset.

As indicated in the table, VC-based anonymization systems perform better in emotion preservation. Both ConVec2Mel-VC and kNN-VC show comparable performance, with an average UAR of around 49%, implying that the speech attributes retained by these systems support the emotion recognizer in identifying the target emotion. Nevertheless, some hidden speech characteristics tied to speaker identity remain unanonymized, allowing the speaker verification model to detect these patterns, resulting in an average EER of less than 20%. Specifically, ConVec2Mel-VC achieves an EER of 9.7% in privacy evaluation, while the kNN-VC approach attains an average EER of around 15.19%. Therefore, although the VC anonymization systems preserve some level of emotion, they

¹<https://github.com/auspicious3000/contentvec>

²<https://github.com/openai/whisper>

³<https://github.com/coqui-ai/TTS>

⁴<https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2024>

Table 1: Privacy and utility performance of various anonymization approaches (darker color indicates better performance)

System	Privacy - EER (%) \uparrow					Utility - UAR mean (%) \uparrow			Utility - WER (%) \downarrow		
	libri-dev-f	libri-dev-m	libri-test-f	libri-test-m	avg.	IEMOCAP-dev	IEMOCAP-test	avg.	libri-dev	libri-test	avg.
Origin	10.511	0.931	8.761	0.418	5.16	69.0796	71.0618	70.07	1.807	1.844	1.83
ConVec2Mel-VC ¹	15.342	7.451	10.444	5.57	9.70	50.7706	48.3282	49.55	2.157	2.269	2.21
kNN-VC ¹	18.351	13.663	16.239	12.496	15.19	47.7042	50.6086	49.16	2.991	2.962	2.98
ConVec2Mel-VC-XTTS ²	39.775	31.056	36.817	30.959	34.65	45.2988	40.6318	42.97	3.999	4.329	4.16
kNN-VC-XTTS ²	44.034	44.567	43.939	46.135	44.67	36.7774	38.0922	37.43	4.758	4.069	4.41
XTTS ³	48.143	48.769	47.040	47.660	47.90	34.3710	32.9232	33.65	4.869	4.537	4.70
Emo _{MSP} -XTTS ⁴	44.034	37.888	46.899	45.637	43.61	36.8728	37.0036	36.94	4.834	3.898	4.37
Emo _{IEMOCAP} -XTTS ⁴	43.751	44.100	45.256	47.834	45.24	52.0652	52.8012	52.43	4.520	3.967	4.24

¹alter the original voice while leaving some prosodic features, such as phoneme durations, unchanged

²clone the anonymized voice and speaking style from the anonymized speech

³fully anonymize the voice by cloning both the voice and speaking style from a random utterance

⁴replicate a different speaker's voice and speaking style, yet expressing the same emotion

also leak speaker information. Both systems maintain the original content well, as reflected in the WER results.

The XTTS system achieves the highest privacy performance among all systems, with an EER close to 50%, by cloning a random voice and speaking style from another utterance. However, when the XTTS system is conditioned on an utterance with a modified voice but preserved prosodic attributes, resulting anonymization systems like ConVec2Mel-VC-XTTS and kNN-VC-XTTS can achieve higher emotion preservation scores. For instance, the ConVec2Mel-VC-XTTS system, which clones the speaking style from the ConVec2Mel-VC system, achieves an average UAR of 42.97%. This is lower than the UAR of ConVec2Mel-VC but higher than the XTTS approach, which has a UAR of 33.65%. Notably, the emotion preservation performance of kNN-VC-XTTS, while better than XTTS, is not significantly higher. This might be due to the lack of temporal coherence in the kNN-VC, leading to a distorted distribution compared to normal speech and causing the XTTS system to struggle with cloning the corresponding speaking style.

This again leads to speaker identity leakage from these preserved attributes. Regarding privacy preservation, the ConVec2Mel-VC-XTTS system achieves an EER of about 34.65%, which is lower than the XTTS approach. This suggests that attributes other than voice timbre can reveal speaker information and are helpful in emotion recognition.

3.4. Achieving the best of both worlds

The systems discussed above demonstrate a clear trade-off between privacy and emotion preservation performance. The results are shown in Figure 3. As emotion preservation performance rises, speaker information leakage takes place, leading to a decrease in privacy performance.

Based on the above results, we propose that randomly cloning a speaker's voice with a different speaking style expressing the same emotion could break this trade-off. To test this, we use emotion embeddings extracted from emotion recognizers as a proxy to find a target utterance for voice and style cloning in the cascaded ASR-TTS approach. This study examines two types of emotion embeddings. The first is a concatenated emotion representation from embeddings extracted by five pre-trained emotion recognizers provided by the challenge. This in-domain representation, derived from systems trained with IEMOCAP, represents the optimal emotion preservation achievable when the anonymization system has access to the

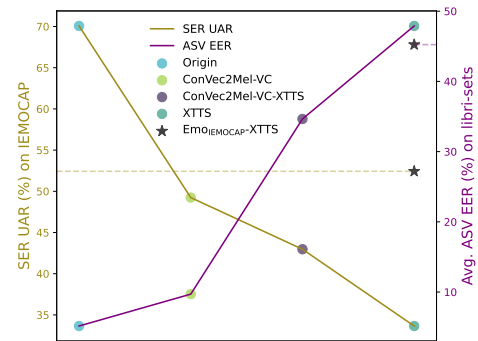


Fig. 3: Privacy-emotion preservation trade-off

SER system. The second embedding is obtained from an out-of-domain extractor⁵ [27] trained with the MSP-Podcast dataset [28].

For each utterance to be anonymized, the prompt audio is selected from the LibriTTS dataset using the following steps: 1. randomly select 5000 utterances from the dataset, 2. calculate the cosine similarity between the emotion representations of source utterance and all 5000 utterances, 3. randomly choose one target utterance from the top 10 utterances with the highest similarity scores.

The last two lines of Table 1 present the results of the anonymization system that employs the emotion-proxy anonymization strategy. Emo_{MSP}-XTTS is based on the out-of-domain emotion recognizer, while Emo_{IEMOCAP}-XTTS relies on the in-domain emotion recognizer. Emo_{IEMOCAP}-XTTS achieves strong emotion preservation performance with a UAR of 52.43% and, simultaneously, high privacy performance with an EER of 45.24%. This system is marked in Figure 3 as the ideal system, breaking the privacy-emotion preservation trade-off shown earlier. However, this assumes that the anonymization system has prior knowledge of the emotion recognition system.

In the alternative scenario, where the anonymization system possesses out-of-domain prior emotion knowledge, Emo_{MSP}-XTTS achieves privacy performance comparable to Emo_{IEMOCAP}-XTTS but struggles to find the best match for emotion, with an emotion preservation UAR of 36.94%. Despite this, the emotion preservation score is higher than that of the XTTS system, suggesting that the emotion-proxy strategy is effective in the anonymization task.

⁵<https://huggingface.co/audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim>

3.5. Speaker-Identifying Information in Emotion Embeddings

We extract emotion embeddings by models trained with IEMOCAP for 10 randomly selected speakers from libri-dev set and plot the embedding space by projecting it to 2D space using t-SNE. As observed from Figure 4, although the embeddings are learned by training models to classify four emotions, embeddings from distinct speakers are distributed apart, while embeddings from the same speaker are clustered together in the representation space. This suggests that emotion embeddings carry a certain amount of speaker information, leading to the trade-off in speech anonymization.

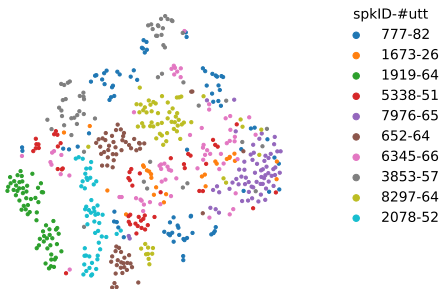


Fig. 4: t-SNE visualization on libri-dev emotion embedding space

To explore how much SER and ASV systems depend on the same speech characteristics for recognition, we employ the emotion recognizer as an utterance-level representation extractor and train a speaker verification model solely using the emotion representations. The emotion embeddings serve as input, followed by a hidden layer with 192 neurons, a dropout rate of 0.5, and the ReLU activation function to map the input to speaker embeddings. A speaker classifier is then employed to predict the target speaker. The training architecture and hyperparameter settings adhere to the challenge’s recipes, except that the input features are changed to emotion embeddings, and the speaker embedding extractor is simplified to a hidden layer instead of using the ECAPA-TDNN structure.

The experiment is conducted on two datasets: libri-train-360 and VoxCeleb1. For both datasets, we split the data into training and validation subsets with a ratio of 0.9 and 0.1, respectively. For the model trained on the libri-train-360 dataset, we set the epochs to 50, while for the VoxCeleb1 dataset, we set the epochs to 200. After training, we select the model with the best performance on the validation set for evaluation. For each experiment, we use 5 different random seeds, training and evaluating the model on the corresponding data.

The verification performance on the evaluation sets is shown in Table 2. Speakers from the test sets are distinguishable with models trained solely on emotion embeddings. Specifically, the model trained on libri-train-360 achieves an EER of 19.28% on the lib-dev-f set and EERs of less than 10% on other evaluation sets. The model trained with VoxCeleb1 achieves an EER of approximately 12.68% on the corresponding test set. These results indicate that a certain level of speaker information is embedded in the emotion embeddings. Such information can be extracted and learned by a single linear layer, highlighting the challenge of disentangling speaker and emotion attributes to fully conceal speaker identity while preserving the emotional state in speech anonymization.

4. DISCUSSIONS AND CONCLUSIONS

To explore factors that expose speaker identity, we use VC approaches without paralinguistic attribute control in our study. Future

Table 2: Speaker verification results

Train Set	Test Set	#trials	ASV EER (%)		
			mean	min	max
libri-train-360	lib-dev-f	15270	19.280 ± 3.414	16.620	23.131
	lib-dev-m	13440	4.996 ± 0.583	4.502	5.433
	lib-test-f	11744	9.922 ± 1.768	8.395	12.042
	lib-test-m	9906	5.883 ± 0.724	5.120	6.412
VoxCeleb1-dev	vox1-test	37611	12.686 ± 0.214	12.515	12.940
	vox1-test-f	7036	16.919 ± 0.660	16.351	17.781
	vox1-test-m	22483	15.621 ± 0.257	15.349	15.892

work will investigate VC systems [29, 30] that incorporate speaker-emotion disentanglement abilities in the anonymization setting. The XTTS system’s ability to clone the target speaker’s voice and speaking style, along with the better emotion preservation performance of the cascaded ASR-TTS system when cloning voice-converted utterances rather than random utterances, demonstrates the effectiveness of emotion cloning. Since there is no current standard for objective evaluation of emotion cloning in the speech synthesis field, our results indicate that the emotion recognition pipeline from the VoicePrivacy 2024 challenge could be well-suited for this purpose.

While our experiments used prompt speakers and utterances from the LibriTTS dataset, using anonymized voices from more expressive corpora with richer emotions could provide a clearer insight into the entanglement between speaker information and emotion. Additionally, our study did not address the degree of speaker information exposure between speaking style and timbre, which remains unknown. The emotion recognizer trained with the IEMOCAP dataset, which has a limited number of speakers, contains retrievable speaker information. It would be interesting to see whether an emotion recognizer trained on a dataset with a larger number of speakers, like MSP-Podcast, retains more speaker information. This would be useful in understanding the speech attributes influencing data-driven speaker and emotion recognizers.

In summary, this paper explores the entanglement between speaker and emotion in speech. Our experimental results on the speech anonymization task demonstrate that enhancing privacy preservation performance results in decreased emotion preservation, highlighting the trade-off between these two attributes. However, this trade-off can be overcome if the anonymization system incorporates a robust emotion recognizer. Furthermore, emotion recognizers also retain speaker information, suggesting that speaker and emotion recognizers depend on similar speech characteristics for recognition. Thus, disentangling emotion and speaker attributes from speech remains a challenging and significant task to address.

Acknowledgement

This work was supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the ARTS Program under contract D2023-2308110001. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

5. REFERENCES

- [1] A. Nautsch, A. Jiménez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado, M. Todisco, M. A. Hmani, A. Mtibaa *et al.*, “Preserving Privacy in Speaker and Speech Characterisation,” *Computer Speech & Language*, vol. 58, pp. 441–480, 2019.
- [2] S. Zhang, Z. Li, and A. Das, “Voicepm: A Robust Privacy Measurement on Voice Anonymity,” in *Proc. 16th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, 2023, pp. 215–226.
- [3] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco, “Introducing the VoicePrivacy Initiative,” in *Proc. Interspeech 2020*, pp. 1693–1697.
- [4] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, “Speaker Anonymization Using X-vector and Neural Waveform Models,” in *Proc. 10th ISCA Workshop on Speech Synthesis*, 2019, pp. 155–160.
- [5] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-Vectors: Robust DNN Embeddings for Speaker Recognition,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 5329–5333.
- [6] X. Miao, X. Wang, E. Cooper, J. Yamagishi, and N. Tomashenko, “Speaker Anonymization Using Orthogonal Householder Neural Network,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.
- [7] C. O. Mawalim, K. Galajit, J. Karnjana, and M. Unoki, “X-Vector Singular Value Modification and Statistical-Based Decomposition with Ensemble Regression Modeling for Speaker Anonymization System,” in *Proc. Interspeech 2020*, pp. 1703–1707.
- [8] H. Turner, G. Lovisotto, and I. Martinovic, “Speaker Anonymization with Distribution-Preserving X-Vector Generation for the VoicePrivacy Challenge 2020,” *arXiv preprint arXiv:2010.13457*, 2020.
- [9] S. Meyer, F. Lux, J. Koch, P. Denisov, P. Tilli, and N. T. Vu, “Prosody is Not Identity: A Speaker Anonymization Approach using Prosody Cloning,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.
- [10] S. Meyer, F. Lux, P. Denisov, J. Koch, P. Tilli, and N. T. Vu, “Speaker Anonymization with Phonetic Intermediate Representations,” in *Proc. Interspeech 2022*, pp. 4925–4929.
- [11] N. Tomashenko, X. Wang, X. Miao, H. Nourtel, P. Champion, M. Todisco, E. Vincent, N. Evans, J. Yamagishi, and J.-F. Bonastre, “The VoicePrivacy 2022 Challenge Evaluation Plan,” *arXiv preprint arXiv:2203.12468*, 2022.
- [12] 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.
- [13] L. Mary and B. Yegnanarayana, “Prosodic Features for Speaker Verification,” in *Ninth International Conference on Spoken Language Processing*, 2006.
- [14] D. Cai, Z. Cai, and M. Li, “Identifying Source Speakers for Voice Conversion based Spoofing Attacks on Speaker Verification Systems,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.
- [15] K. Qian, Y. Zhang, H. Gao, J. Ni, C.-I. Lai, D. Cox, M. Hasegawa-Johnson, and S. Chang, “Contentvec: An Improved Self-Supervised Speech Representation by Disentangling Speakers,” in *International Conference on Machine Learning*, 2022, pp. 18 003–18 017.
- [16] Z. Cai, Y. Yang, and M. Li, “Cross-Lingual Multi-Speaker Speech Synthesis with Limited Bilingual Training Data,” *Computer Speech & Language*, vol. 77, p. 101427, 2023.
- [17] J. Kong, J. Kim, and J. Bae, “HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 17 022–17 033.
- [18] M. Baas, B. van Niekerk, and H. Kamper, “Voice Conversion with Just Nearest Neighbors,” in *Proc. Interspeech 2023*, pp. 2053–2057.
- [19] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, “WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, p. 1505–1518, 2022.
- [20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR Corpus Based on Public Domain Audio Books,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 5206–5210.
- [21] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, “IEMOCAP: Interactive Emotional Dyadic Motion Capture Database,” *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.
- [22] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, “LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech,” in *Proc. Interspeech 2019*, pp. 1526–1530.
- [23] A. Nagrani, J. S. Chung, and A. Zisserman, “VoxCeleb: A Large-Scale Speaker Identification Dataset,” in *Proc. Interspeech 2017*, pp. 2616–2620.
- [24] B. Desplanques, J. Thienpondt, and K. Demuyne, “ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification,” in *Proc. Interspeech 2020*, pp. 3830–3834.
- [25] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, “SpeechBrain: A General-Purpose Speech Toolkit,” *arXiv preprint arXiv:2106.04624*, 2022.
- [26] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision,” in *International Conference on Machine Learning*, 2023, pp. 28 492–28 518.
- [27] 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*, vol. 45, no. 9, pp. 10 745–10 759, 2023.
- [28] 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, 2019.
- [29] X. Chen, X. Xu, J. Chen, Z. Zhang, T. Takiguchi, and E. R. Hancock, “Speaker-Independent Emotional Voice Conversion via Disentangled Representations,” *IEEE Transactions on Multimedia*, vol. 25, pp. 7480–7493, 2022.
- [30] Z. Du, B. Sisman, K. Zhou, and H. Li, “Disentanglement of Emotional Style and Speaker Identity for Expressive Voice Conversion,” in *Proc. Interspeech 2022*, pp. 2603–2607.