INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Emo-StarGAN: A Semi-Supervised Any-to-Many Non-Parallel Emotion-Preserving Voice Conversion

Suhita Ghosh[1*], *Arnab Das*[1,3*], *Yamini Sinha*[2], *Ingo Siegert*[2], *Tim Polzehl*[3], *Sebastian Stober*[1]

[1]Artificial Intelligence Lab (AILab), Otto-von-Guericke-University, Magdeburg, Germany
[2]Mobile Dialog Systems, Otto-von-Guericke-University, Magdeburg, Germany
[3]Speech and Language Technology, German Research Center for Artificial Intelligence (DFKI)

{suhita.ghosh,yamini.sinha,ingo.siegert,stober}@ovgu.de
{arnab.das,tim.polzehl}@dfki.de

Abstract

Speech anonymisation prevents misuse of spoken data by removing any personal identifier while preserving at least linguistic content. However, emotion preservation is crucial for natural human-computer interaction. The well-known voice conversion technique StarGANv2-VC achieves anonymisation but fails to preserve emotion. This work presents an any-to-many semi-supervised StarGANv2-VC variant trained on partially emotion-labelled non-parallel data. We propose emotion-aware losses computed on the emotion embeddings and acoustic features correlated to emotion. Additionally, we use an emotion classifier to provide direct emotion supervision. Objective and subjective evaluations show that the proposed approach significantly improves emotion preservation over the vanilla StarGANv2-VC. This considerable improvement is seen over diverse datasets, emotions, target speakers, and inter-group conversions without compromising intelligibility and anonymisation.
Index Terms: speech anonymisation, voice conversion, StarGAN

1. Introduction

The increasing use of cloud-based speech devices, such as smart speakers, raises concerns about the protection and confidentiality of the sensitive data being collected and used [1, 2]. In case of data compromise, the spoken data can be exploited to bypass the speaker verification systems or impersonate authorised users [3, 4]. This makes it crucial to anonymise the utterance before being shared across systems, such that the speaker cannot be traced back. Voice conversion (VC) achieves anonymisation by modifying the utterance of the source speaker to sound like another target speaker while preserving at least linguistic content. In cases where the response of a speech device is driven by the end-user's emotional state, the preservation of emotion also becomes pertinent, e.g., a digital assistant responding with comforting words when the user sounds sad.

Many VC approaches using parallel data have been proposed, such as parametric statistical modelling-based [5, 6], non-parametric exemplar-based [7, 8] and deep neural network-based [9]. Parallel data comprise utterances having the same linguistic content from both the source and target speakers, which is arduous and expensive to acquire. Therefore, recent works focus more on non-parallel data, as it is simpler to obtain and better represents real-life situations where any arbitrary speech requires anonymisation.

A few non-parallel VC approaches [10, 11] use phonetic posteriorgrams (PPGs) as one of the inputs to the encoder-decoder framework to generate translated acoustic features. These methods tend to produce mispronunciations due to alignment issues [12], resulting in degraded prosody, which provides cues about emotion [13]. The non-parallel variational autoencoder (VAE) approaches [14, 15] typically disentangle the content and speaker embeddings using a reconstruction loss and relevant constraints to remove speaker information. The VAE-based approaches are prone to spectrum smoothing, which leads to a buzzy-sounding voice, dampening the emotion [16]. A plethora of generative adversarial network (GAN) based VC approaches [16, 17] were proposed, which can use non-parallel data due to cycle-consistency loss [18]. GANs overcome the over-smoothing effect through a discriminator, which teaches the generator to produce natural sounding conversions. Recently, StarGANv2-VC [19], a non-parallel any-to-many GAN-based VC technique has been proposed. The method is attractive due to its fast real-time conversion and naturally sounding samples with high intelligibility. However, the model fails to preserve the source speaker's emotion, especially for diverse emotions and acoustic conditions such as high varying pitch.

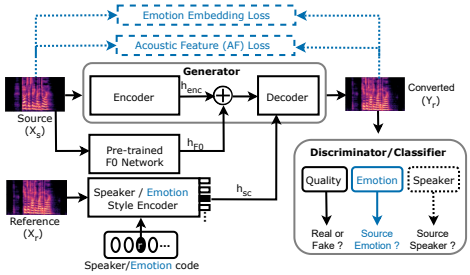

Figure 1: *The proposed framework adapted from StarGANv2-VC [19]. The blue components do not belong to StarGANv2-VC. In voice conversion, the style encoder captures speaker embeddings. The same framework is used for emotion conversion, where the style encoder learns emotion embeddings. The dashed components are not used in the emotion embedding training.*

Thus, we propose the novel "Emo-StarGAN" in this paper, which is an any-to-many semi-supervised *emotion-preserving* variant of StarGANv2-VC. Two kinds of emotion supervision are proposed: (i) *direct*: through an emotion classifier, which provides feedback to the generator when the emotion ground truth is available. (ii) *indirect*: through losses computed between source and conversions using emotion embeddings or acoustic descriptors correlated with emotion, improving the conversion quality for diverse target speakers. Extensive evaluation is conducted on three datasets, diverse target speakers, emotions, and various group conversions such as accent and gender. Both objective and subjective evaluations portray that Emo-StarGAN improves emotion preservation significantly over StarGANv2-VC for all cases, without hurting the naturalness, intelligibility and anonymisation.

2. StarGANv2-VC Architecture

Our method is based on the StarGANv2-VC architecture, as shown in Figure 1. A *single* generator G is trained to convert a source

*These authors contributed equally to this work

10.21437/Interspeech.2023-191

utterance X_s to the target utterance Y_r, conditioned on the speaker style embedding h_{sc}. The speaker style embedding h_{sc} represents *speaker characteristics*, such as accent. The speaker style-encoder SE produces the speaker style embedding h_{sc} using the target speaker's mel-spectrogram X_r having the style information and, target speaker's code r (one-hot encoding). SE comprises multiple convolutional layers which are shared by all the speakers, followed by a speaker-specific linear projection layer, which outputs an embedding h_{sc} for each target speaker. A mapping network M having the same architecture as SE is trained along with it, which inputs a random latent vector instead of a reference mel-spectrogram, providing diverse style representation for all speakers. The converted sample produced by the generator $Y_r = G(X_s,h_{F0},h_{sc})$ captures the style of the target speaker-code r and has the linguistic content of the source utterance X_s. In order to produce F0-consistent conversions, the generator is fed with source-pitch embedding h_{F0} along with source utterance X_s and style representation h_{sc}. The pitch embedding h_{F0} is derived from the convolutional outputs of a pre-trained F0 network [20]. The framework consists of one discriminator D and one adversarial source speaker classifier C_s. D is the typical adversarial discriminator, which encourages the generator to produce plausible conversions. C_s has the same architecture as D, which is trained to enforce the generator to produce conversions having no details about the source speaker.

3. Emo-StarGAN

Recent VC works [21] including StarGANv2-VC have primarily focused on generating naturally sounding voices with correct linguistic content, and not much on emotion preservation. The proposed Emo-StarGAN aims to anonymise an utterance by modifying the source speaker's timbre, while preserving the source's linguistic and *emotional* content e_s.

3.1. Direct Emotion Supervision

Our framework uses an additional emotion classifier C_e which provides *direct* emotion supervision for utterances having emotion labels, as shown in Figure 1. C_e encourages the generator to produce *emotion-consistent* samples, such that the source and target samples have the same emotion. When C_e is trained, the generator weights are fixed, and the emotion classifier is trained to ascertain the emotion of the source utterance through the classification loss L_{emod}.

$$L_{emod} = \mathbb{E}_{X_s,e_s}\left[CrossEntropy(C_e(X_s),e_s)\right] \quad (1)$$

In contrast, during the training of the generator, C_e weights are fixed, and the generator is encouraged to produce samples having the same emotion as the source through the loss L_{emog}.

$$L_{emog} = \mathbb{E}_{X_s,e_s,h_{sc}}\left[CrossEntropy(C_e(G(X_s,h_{sc})),e_s)\right] \quad (2)$$

3.2. Indirect Emotion Supervision

Incorporation of explicit emotion supervision for the converted samples becomes challenging due to the unavailability of the emotion labels. Therefore, it becomes pertinent to measure the emotion discrepancy between the source and the converted samples through representations of emotion. To this end, we propose two ways to measure discrepancies of the emotional content: acoustic features correlated to emotion and deep emotion embeddings.

3.2.1. Emotion-aware Acoustic Feature Loss

We propose acoustic feature loss L_{af}, an unsupervised loss computed between the acoustic descriptors of the source and converted samples, as shown in Equation 3, where AF denotes an acoustic feature. The acoustic features are correlated with emotion and require

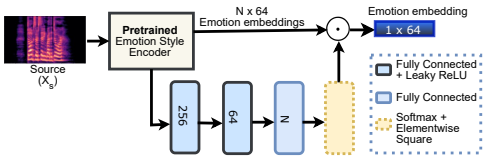

Figure 2: *Automatic emotion embedding extraction. N denotes the number of emotions classes.*

$$L_{af} = \mathbb{E}_{X_s,h_{sc}}\left[\|AF(X_s) - AF(G(X_s,h_{sc}))\|_1\right] \quad (3)$$

being differentiable to provide feedback to the network. Based on [22], the acoustic descriptors can be categorised into two groups, spectral and non-spectral. Spectral features add additional information about higher-level harmonics to that already existing in pitch, which provides pertinent cues for the emotional state [23]. Many works [23, 24] report spectral features to be better discriminators in between emotions that have different degree of polarity (valence) but similar intensity (arousal), such as anger and happiness. The non-spectral features are energy or voicing-related, which are typically prosodic and arousal indicative [25]. We consider two descriptors from each of the two categories. All descriptors are extracted over voiced segments using 50% overlapping windows, to capture the local transients.

- **Spectral centroid:** Higher spectral centroid values indicate emotions positioned in the upper-right quadrant of the valence-arousal 2D plane, such as *excited* or *happy* [26]. Lower values indicate subdued emotions, such as *sad*.
- **Spectral kurtosis:** Spectral kurtosis shows the existence of increased energy concentration within specific frequency ranges. Further, it can detect the series of transients [27], which can make it a good indicator of emotions, especially the ones having subtle intonation changes, such as in the emotion *surprise*.
- **Loudness:** Loudness is an arousal indicative non-spectral feature, which correlates stronger to emotion than root-mean-square energy due to the perceptual A-weighting [28]. Louder sounds elicit stronger emotional responses (high arousal), and vice-versa.
- **Change in F0 (ΔF0):** ΔF0 is a prosodic non-spectral feature, which captures change in intonation, where a considerable change implies stronger emotions, such as *anger* or *excited* [28].

3.2.2. Emotion Embedding Loss

Another way of incorporating indirect emotion supervision is through latent emotion representations. The emotion embedding loss L_{embed} penalises the discrepancy between the latent emotional content of the source and converted samples.

$$L_{embed} = \mathbb{E}_{X_s,h_{sc}}\left[\|Emb(X_s) - Emb(G(X_s,h_{sc}))\|_1\right] \quad (4)$$

The emotion embedding is obtained by a two-stage training on categorical emotion-labelled data. At Stage I, the vanilla StarGANv2-VC model is trained for emotion conversion task rather than voice conversion, as shown in Figure 1. The emotion style-encoder learns $N \times 64$ embeddings of emotion classes, where N denotes the number of emotion classes. However, this framework cannot be used in the VC training, as an emotion label (code) is required to generate the emotion embeddings, which is unknown for the converted samples. Therefore, the pre-trained emotion style-encoder from Stage I is fine-tuned for automatic embedding extraction, as shown in Figure 2. At Stage II, the pre-trained emotion style-encoder is extended with fully-connected layers and a softmax

distribution is generated over all emotions. Further, the softmax score is element-wise squared to encourage sparsity. Finally, a dot product between the sparse $1 \times N$ score and the encoder output is performed to produce a 1×64 dimensional latent emotion representation. This fine-tuned model is used in the VC training to extract the emotion embeddings from both source and converted samples.

3.3. Training Objectives

The components in Emo-StarGAN are trained with the proposed emotion-aware losses along with the losses from StarGANv2-VC. The generator is trained with loss L_G (Equation 5) comprising the proposed emotion classification loss L_{emod}, unsupervised emotion-aware losses (L_{af} and L_{embed}), and losses from StarGANv2-VC.

$$L_G = \min_{G,SE,M} L_{adv} + \lambda_{af} L_{af} + \lambda_{embed} L_{embed} + \lambda_{emog} L_{emog} + \lambda_{aspk} L_{aspk} + \lambda_{sty} L_{sty} - \lambda_{ds} L_{ds} + \lambda_{F0} L_{F0} + \lambda_{asr} L_{asr} + \lambda_{cyc} L_{cyc} \quad (5)$$

The losses from StarGANv2-VC: L_{adv} is the typical GAN adversarial loss, L_{aspk} is the adversarial source speaker classification loss, L_{sty} ensures that the style representations can be recreated from the generated samples, L_{ds} is maximised to ensure samples generated from different speaker style-codes sound different, L_{F0} encourages the generator to produce F0-consistent samples, L_{asr} ensures source and converted samples have the same linguistic content and L_{cyc} is the cyclic consistency loss, which preserves the non-timbre features of the source. λ_{af}, λ_{embed}, λ_{emog}, λ_{aspk}, λ_{sty}, λ_{ds}, λ_{F0}, λ_{asr} and λ_{cyc} are hyperparameters of the corresponding losses. The discriminator and classifiers are trained using the objective function shown in Equation 6, where λ_{spk} and λ_{emod} are hyperparameters for the source speaker classification loss L_{spk} and emotion classification loss L_{emod}, respectively.

$$L_D = \min_{D,C_e,C_s} -L_{adv} + \lambda_{emod} L_{emod} + \lambda_{spk} L_{spk} \quad (6)$$

4. Experiment and Results

4.1. Dataset and Training details

English utterances from VCTK [29], emotional speech dataset (ESD) [30] and Ryerson audio-visual database of emotional speech and song (RAVDESS) [31] datasets are considered. VCTK has no emotion ground truth, whereas ESD and RAVDESS are labelled with categorical emotions, where we consider five emotion classes common to both, $e \in \{$happy, sad, anger, neutral, surprise$\}$. The utterances are re-sampled to 24 kHz and randomly split as 0.8/0.1/0.1 (train/validation/test). All VC models are trained on 10 randomly selected speakers from VCTK and ESD each.

Our model has the same number of trainable model parameters as StarGANv2-VC. Each model is trained on log mel-spectrograms derived from 2 second audio samples, for 100 epochs with a batch size of 16. Each training takes around 36 hours on average to complete on A100 (80 GB). We use pre-trained F0 and automatic speech recognition models from [19]. AdamW optimizer [32] is used with a learning rate of 10^{-4}. We set $\lambda_{aspk} = 0.1, \lambda_{emod} = 0.01, \lambda_{emog} = 0.01, \lambda_{sty} = 1, \lambda_{ds} = 1, \lambda_{F0} = 5, \lambda_{asr} = 1, \lambda_{cyc} = 1, \lambda_{embed} = 2$ and $\lambda_{af} = 2$. A HiFiGAN [33] vocoder is trained on the mentioned datasets, which generates one-minute long waveform from the converted mel-spectrogram in 0.1 seconds on the A100. The emotion conversion model is trained using cross-validation only on ESD, using training split 0.9/0.1 (train/validation), and using the same setup as the VC models. The best model is selected based on the lowest mean absolute error (MAE). To evaluate emotion preservation, a Support Vector Machine (SVM) based emotion classifier is trained as in [34] on *source* utterances of ESD and RAVDESS.

4.2. Evaluation Setup

We evaluate our approach using both objective and subjective measures. We consider StarGANv2-VC as the *baseline*. Further, we perform experiments to find the best emotion-preserving acoustic feature AF_{best}. We train our model Emo-StarGAN using the combination of emotion classifier loss, emotion embedding loss and acoustic feature loss using AF_{best}. For all experiments, an equal number of female (F) and male (M) speakers are randomly selected as source and target. From each of the three datasets, 10 source speakers are considered. For ESD and RAVDESS, 5 utterances for each source speaker and each emotion in e are selected. We convert source utterances from ESD using ESD target speakers (ESD→ESD) for *within-corpus* and RAVDESS→ESD for *cross-corpus* scenarios. We select 6 target speakers from ESD, leading to 1500 conversions for each scenario. For the *inter-accent* conversion use case, VCTK→VCTK conversion is performed, where 10 utterances from each source speaker and accent group (British, American, and Canadian) and 6 target speakers having British accent are considered, leading to 1800 conversions.

Objective Evaluation: Emotion preservation is evaluated in four ways: (i) Acc_{orig}: SVM classification accuracy, considering the emotion labels of the source utterance provided in the dataset, (ii) Acc_{svm}: SVM classification accuracy, considering SVM prediction on source utterance as the emotion ground truth, (iii) Embedding MAE: mean absolute error between the source and converted emotion embedding outputs, (iv) Pitch correlation coefficient (PCC): measures the degree of intonation preservation [35], which provides cues to emotion preservation [36]. The voice quality is measured by predicted mean opinion score (pMOS) [37]. We report the character error rate (CER) using the transcriptions from Whisper *medium-english* model [38]. Equal error rate (EER) measures anonymisation using the state-of-the-art speaker verification model ECAPA-TDNN [39]. For the metrics Acc_{orig}, Acc_{svm}, PCC, pMOS, and EER higher values indicate better performance, and for Embedding MAE and CER lower values are preferable.

Subjective evaluation: We consider 100 randomly selected conversions for subjective evaluation as it is expensive and time-consuming to perform for all. 138 online subjects participated in the user study through the Crowdee[1] platform. For emotion preservation assessment, subjects were presented with the source utterance and two options: conversions from baseline and Emo-StarGAN. Further, they were asked to choose one of the options having similar rhythm, intonation, pauses, stresses and intensity as the source, irrespective of voice quality and the linguistic content. The subjects were asked to rate on a 5-point scale for naturalness (1: bad to 5: excellent). For speaker anonymisation, the raters were presented with the converted sample and another utterance from the source speaker, and were asked to rate on a 5-point scale (1: different to 5: similar). At least three subjects rated each task. The raters were not informed whether the samples are original or converted. They were further provided with anchoring examples and hidden trapping questions. Raters caught in the latter twice were rejected from evaluations.

4.3. Results and Discussion

Selection of Acoustic Feature (AF) and Ablation: In order to get AF_{best}, we train the baseline with acoustic feature loss, where the AF is replaced with one of the acoustic features mentioned in Section 3.2.1. Among all acoustic features, *spectral kurtosis* preserves emotion the most (30.1% Acc_{orig}, 80.4 PCC), also outperforming the baseline (19% Acc_{orig}, 78.1 PCC). The PCC values of the other acoustic features are similar, having range 80.0 to 80.4. Acc_{orig} for

[1] https://www.crowdee.com/

Table 1: *Objective evaluation. Mean and standard deviation (in brackets) reported. Emo-SG denotes Emo-StarGAN. 'All Conv.' includes all conversions. Type column denotes special cases, such as source-emotion, source-accent → target-accent, source and target same genders (M→M, F→F), source and target different genders (M→F, F→M) or, 'All' including all sub-groups.*

Source - Target	Type	Acc_{orig} [%] ↑		Acc_{svm} [%] ↑		Embedding MAE [$\times 10^2$] ↓		PCC [$\times 10^2$] ↑		pMOS ↑		CER [%] ↓		EER [%] ↑	
		Baseline	Emo-SG	Baseline	Emo-SG	Baseline	Emo-SG	Baseline	Emo-SG	Baseline	Emo-SG	Baseline	Emo-SG	Baseline	Emo-SG
All Conv.	All	20.2	**72.4**	39.4	**87.1**	48.9 (11.1)	**40.8** (10.9)	78.9 (12.9)	**84.3** (10.6)	3.68 (0.41)	**3.72** (0.44)	3.42 (8.39)	**2.57** (7.08)	49.63	**49.64**
ESD ↓ ESD	All	19.1	**68.9**	20.1	**94.7**	43.4 (17.0)	**31.5** (14.4)	78.1 (14.0)	**84.9** (10.4)	3.75 (0.43)	**3.90** (0.4)	4.27 (8.79)	**3.56** (7.75)	**45.86**	45.45
	Happy	10.7	**69.7**	13.1	**85.7**	43.3 (17.5)	**31.0** (14.9)	76.7 (16.7)	**84.0** (11.7)	3.60 (0.45)	**3.76** (0.41)	5.21 (9.36)	**4.92** (8.67)	-	-
	Sad	15.7	**96.2**	15.7	**96.2**	40.7 (15.3)	**32.4** (14.5)	82.3 (12.7)	**87.5** (10.0)	3.76 (0.39)	**4.07** (0.38)	2.47 (6.15)	**2.38** (6.0)	-	-
	Surprise	0.0	**16.0**	2.1	**97.9**	47.1 (15.5)	**29.5** (11.6)	77.6 (11.2)	**85.4** (7.4)	3.64 (0.38)	**3.74** (0.35)	7.43 (11.17)	**5.16** (9.3)	-	-
	Angry	10.6	**94.0**	10.6	**95.0**	43.7 (17.8)	**31.1** (14.5)	76.1 (14.8)	**85.6** (9.6)	3.77 (0.39)	**3.86** (0.34)	3.53 (8.02)	**2.96** (7.01)	-	-
	Neutral	79.3	**99.3**	79.7	**99.3**	40.7 (18.8)	**34.5** (16.5)	78.0 (12.8)	**80.4** (12.6)	4.08 (0.37)	**4.19** (0.34)	**1.63** (5.73)	1.69 (5.93)	-	-
	Different gender	18.6	**63.3**	19.6	**94.2**	50.05 (15.7)	**38.8** (14.0)	76.9 (14.7)	**86.7** (9.0)	3.71 (0.44)	**3.92** (0.37)	4.82 (9.09)	**2.53** (5.70)	-	-
	Same gender	19.5	**78.6**	21.1	**96.3**	37.6 (15.1)	**25.35** (10.3)	79.1 (13.2)	**82.7** (11.6)	3.78 (0.41)	**3.80** (0.39)	3.78 (8.48)	**3.27** (7.02)	-	-
RAVDESS ↓ ESD	All	27.8	**49.2**	41.4	**76.0**	52.5 (7.41)	**44.7** (7.66)	86.2 (10.8)	**88.0** (9.3)	3.44 (0.41)	**3.49** (0.41)	4.57 (9.89)	**4.52** (5.7)	50.19	**50.44**
	Happy	0.0	**14.0**	63.0	**96.0**	51.2 (6.5)	**44.2** (5.8)	89.0 (7.2)	**90.7** (6.4)	3.35 (0.35)	**3.40** (0.33)	5.19 (9.42)	**2.95** (6.49)	-	-
	Sad	0.0	**59.0**	9.0	**73.0**	55.1 (9.5)	**44.8** (6.5)	76.1 (16.2)	**81.0** (15.0)	**3.20** (0.47)	3.13 (0.45)	13.11 (12.08)	**8.96** (2.31)	-	-
	Surprise	**0.0**	**0.0**	4.0	**34.0**	52.3 (6.1)	**41.2** (6.4)	**89.2** (6.2)	**89.2** (7.5)	**3.37** (0.25)	3.35 (0.34)	4.60 (8.35)	**3.71** (5.45)	-	-
	Angry	51.0	**93.0**	51.0	**93.0**	50.8 (7.5)	**44.6** (9.1)	90.7 (5.4)	**91.8** (4.2)	3.74 (0.33)	**3.84** (0.31)	**2.00** (8.50)	5.65 (3.02)	-	-
	Neutral	80.0	**88.0**	80.0	**88.0**	53.1 (6.2)	**48.8** (8.1)	86.1 (8.2)	**87.7** (5.1)	3.34 (0.29)	**3.48** (0.37)	2.63 (9.07)	**1.28** (3.86)	-	-
	Different gender	33.7	**43.1**	46.8	**70.0**	54.7 (5.2)	**32.2** (6.3)	87.2 (7.1)	**88.6** (6.9)	**3.46** (0.42)	**3.46** (0.40)	5.74 (11.26)	**3.75** (2.50)	-	-
	Same gender	24.2	**52.9**	38.1	**72.9**	51.6 (7.1)	**43.1** (6.1)	85.6 (12.5)	**87.6** (10.5)	3.46 (0.42)	**3.51** (0.40)	**3.73** (8.86)	4.97 (2.35)	-	-
VCTK ↓ VCTK	All	-	-	56.8	**90.6**	50.8 (13.0)	**46.4** (12.1)	78.4 (11.8)	**83.1** (10.8)	3.51 (0.36)	**3.57** (0.37)	3.27 (8.31)	**1.63** (5.54)	**50.13**	49.90
	British → British	-	-	48.9	**91.6**	51.9 (13.4)	**46.9** (12.6)	77.5 (10.8)	**82.5** (9.7)	3.53 (0.35)	**3.62** (0.36)	4.16 (9.5)	**2.19** (6.45)	-	-
	American → British	-	-	66.5	**89.9**	50.0 (12.3)	**46.3** (11.7)	80.0 (11.6)	**84.4** (10.4)	3.55 (0.38)	**3.49** (0.37)	2.58 (7.59)	**1.25** (5.07)	-	-
	Canadian → British	-	-	53.3	**89.9**	50.0 (13.3)	**45.9** (12.0)	77.2 (13.8)	**81.5** (13.1)	3.52 (0.36)	**3.49** (0.37)	2.85 (6.86)	**1.28** (4.24)	-	-
	Different gender	-	-	55.8	**90.4**	48.0 (12.0)	**44.0** (10.0)	79.1 (10.8)	**83.5** (9.9)	3.45 (0.36)	**3.53** (0.36	3.62 (8.53)	**1.76** (5.59)	-	-
	Same gender	-	-	59.5	**91.0**	51.0 (10.5)	**46.1** (9.1)	77.9 (12.5)	**82.7** (11.4)	3.55 (0.36)	**3.61** (0.38)	2.99 (8.11)	**1.53** (5.50	-	-

the other acoustic features are, spectral centroid (24.1%), loudness (15.4%) and $\Delta F0$ (20.1%), which portrays the spectral features to be more emotion preserving than the non-spectral ones, compliant with [23]. The ablation study (Table 2) shows that the unsupervised loss L_{embed} contributes the most to emotion preservation, even more than the direct supervision by emotion classifier C_e, this might be attributed to C_e suffering from confirmation bias on noisy emotion labels. Further, we observe that each individual proposed technique preserves emotion more than the baseline.

Table 2: *Ablation results. Mean and standard deviation (in brackets) reported. L_{af} uses spectral kurtosis as the acoustic feature. Baseline is trained with 'none' of the emotion-aware losses.*

Method	Acc_{orig} [%] ↑	PCC [$\times 10^2$] ↑	pMOS ↑	CER [%] ↓	EER [%] ↑
Baseline	20.2	78.9 (12.9)	3.68 (0.41)	3.42 (8.39)	49.63
Emo-StarGAN	**72.4**	**84.3** (10.6)	**3.72** (0.44)	**2.57** (7.08)	**49.64**
L_{embed}	**51.0**	**81.3** (12.2)	**3.90** (0.40)	**3.12** (7.97)	**48.09**
C_e	49.3	81.0 (12.3)	3.50 (0.46)	3.50 (7.76)	45.83
L_{af}	30.1	80.4 (11.8)	3.89 (0.37)	5.52 (11.68)	47.64

Comparison with Baseline: Our method Emo-StarGAN outperforms the baseline with respect to emotion preservation for all scenarios (Table 1), which is also statistically significant ($p < 0.001$ for paired t-test on PCC and Embedding MAE columns). The subjective evaluation (Table 3) also shows that our model is voted more emotion preserving (72%) compared to the baseline (28%). *Surprise* is reported as one of the most difficult emotions in speech emotion recognition tasks [40]. Our method also achieves lower accuracy for 'surprise' compared to other emotions, where Acc_{orig} scores for ESD and RAVDESS are only 16% and 0% respectively. However, preservation seems much higher considering Acc_{svm} scores, 97.9% ESD and 34% for RAVDESS. Our framework improves emotion preservation significantly for the cross-corpus (RAVDESS→ESD) scenario with respect to all metrics, especially for *sad*, where the emotion preservation improves from 0% to 59% (Acc_{orig}), 9% to 73% (Acc_{svm}), 55.1 to 44.8 (Embedding MAE) and 76% to 81% (PCC). Considering inter-accent cases, our model produces a high Acc_{svm} score of 89.9% for both American → British and Canadian → British conversions, and also improves other quality metrics. For both gender conversion cases, similar observations are made. Our method outperforms the baseline mostly with respect to voice quality, intelligibility, and anonymisation, which is further supported by the subjective results. The code and demo audio samples can be found online[2].

Table 3: *Results of subjective evaluation. Mean and standard deviation (in brackets) reported. Emo-SG denotes Emo-StarGAN. Emotion v. column denotes the number of times a model is preferred over the other. Higher Speaker diss. indicates better anonymisation.*

Type	MOS ↑		Emotion V. ↑		Speaker Diss. ↑	
	Baseline	Emo-SG	Baseline	Emo-SG	Baseline	Emo-SG
All	4.09 (0.93)	**4.20** (0.93)	327	**840**	2.4 (1.4)	**2.6** (1.5)
Different gender	4.20 (0.94)	**4.24** (0.93)	154	**429**	2.7 (1.4)	**2.9** (1.5)
Same gender	4.05 (0.94)	**4.19** (1.01)	173	**411**	1.9 (1.3)	**2.3** (1.5)

5. Conclusions

To the best of our knowledge, we propose the first emotion-preserving any-to-many semi-supervised voice conversion framework Emo-StarGAN. We introduce novel unsupervised acoustic descriptor-based and deep emotion losses, which can be used with any other framework. Extensive experiments show that Emo-StarGAN preserves emotion significantly better than the state-of-the-art VC method StarGANv2-VC over seen source speakers, cross-corpus conversions, different genders, accents and emotions. Subjective results show that our method even achieves higher MOS and anonymisation scores. As future work, we plan to improve the emotion preservation for complex emotions by incorporating losses beneficial to a specific emotion. Further, we would like to extend the method with emotion embeddings learned from multi-label and arousal-valence labelled datasets.

6. Acknowledgements

This research has been partly funded by the Federal Ministry of Education and Research of Germany in the project Emonymous (project number S21060A) and partly funded by the Volkswagen Foundation in the project AnonymPrevent (AI-based Improvement of Anonymity for Remote Assessment, Treatment and Prevention against Child Sexual Abuse).

[2]https://github.com/suhitaghosh10/emo-stargan.git

7. References

[1] C. Wienrich, C. Reitelbach, and A. Carolus, "The trustworthiness of voice assistants in the context of healthcare investigating the effect of perceived expertise on the trustworthiness of voice assistants, providers, data receivers, and automatic speech recognition," *Frontiers in Computer Science*, vol. 3, p. 685250, 2021.

[2] M. Haase, J. Krüger, and I. Siegert, "User perspective on anonymity in voice assistants," in *Proc. of the HCI International*, 2023, p. s.p.

[3] D. Kumar, R. Paccagnella, P. Murley, E. Hennenfent, J. Mason, A. Bates, and M. Bailey, "Skill squatting attacks on Amazon Alexa," in *27th USENIX Security Symposium*, Baltimore, USA, 2018, pp. 33–47.

[4] N. Zhang, X. Mi, X. Feng, X. Wang, Y. Tian, and F. Qian, "Dangerous skills: Understanding and mitigating security risks of voice-controlled third-party functions on virtual personal assistant systems," in *IEEE Symposium on Security and Privacy*, 2019, pp. 1381–1396.

[5] K. Kobayashi, S. Takamichi, S. Nakamura, and T. Toda, "The NU-NAIST Voice Conversion System for the Voice Conversion Challenge 2016," in *Proc. Interspeech 2016*, 2016, pp. 1667–1671.

[6] E. Helander, H. Silén, T. Virtanen, and M. Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 3, pp. 806–817, 2011.

[7] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion using sparse representation in noisy environments," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 96, no. 10, pp. 1946–1953, 2013.

[8] Z. Jin, A. Finkelstein, S. DiVerdi, J. Lu, and G. J. Mysore, "Cute: A concatenative method for voice conversion using exemplar-based unit selection," in *Proc. of the IEEE ICASSP*. IEEE, 2016, pp. 5660–5664.

[9] R. Liu, X. Chen, and X. Wen, "Voice conversion with transformer network," in *Proc. of the IEEE ICASSP*, 2020, pp. 7759–7759.

[10] Z. Lian, Z. Wen, X. Zhou, S. Pu, S. Zhang, and J. Tao, "ARVC: An auto-regressive voice conversion system without parallel training data." in *INTERSPEECH*, 2020, pp. 4706–4710.

[11] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2016, pp. 1–6.

[12] S. Liu, Y. Cao, D. Wang, X. Wu, X. Liu, and H. Meng, "Any-to-many voice conversion with location-relative sequence-to-sequence modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1717–1728, 2021.

[13] H. Cao, Š. Beňuš, R. C. Gur, R. Verma, and A. Nenkova, "Prosodic cues for emotion: analysis with discrete characterization of intonation," *Speech prosody (Urbana, Ill.)*, vol. 2014, p. 130, 2014.

[14] D.-Y. Wu, Y.-H. Chen, and H.-y. Lee, "VQVC+: One-shot voice conversion by vector quantization and U-Net architecture," *Proc. Interspeech 2020*, pp. 4691–4695, 2020.

[15] H. Tang, X. Zhang, J. Wang, N. Cheng, and J. Xiao, "Avqvc: One-shot voice conversion by vector quantization with applying contrastive learning," in *Proc. of the IEEE ICASSP*, 2022, pp. 4613–4617.

[16] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 132–157, 2020.

[17] S. Sakamoto, A. Taniguchi, T. Taniguchi, and H. Kameoka, "StarGAN-VC+ASR: StarGAN-Based Non-Parallel Voice Conversion Regularized by Automatic Speech Recognition," in *Proc. Interspeech 2021*, 2021, pp. 1359–1363.

[18] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," in *2018 26th IEEE EUSIPCO*, 2018, pp. 2100–2104.

[19] Y. A. Li, A. Zare, and N. Mesgarani, "StarGANv2-VC: A Diverse, Unsupervised, Non-Parallel Framework for Natural-Sounding Voice Conversion," in *Proc. Interspeech 2021*, 2021, pp. 1349–1353.

[20] S. Kum and J. Nam, "Joint detection and classification of singing voice melody using convolutional recurrent neural networks," *Applied Sciences*, vol. 9, no. 7, p. 1324, 2019.

[21] T. Walczyna and Z. Piotrowski, "Overview of voice conversion methods based on deep learning," *Applied Sciences*, vol. 13, no. 5, p. 3100, 2023.

[22] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proc. Interspeech 2009*, 2009, pp. 312–315.

[23] A. Tursunov, S. Kwon, and H.-S. Pang, "Discriminating emotions in the valence dimension from speech using timbre features," *Applied Sciences*, vol. 9, no. 12, p. 2470, 2019.

[24] C. Busso and T. Rahman, "Unveiling the acoustic properties that describe the valence dimension," in *Thirteenth Annual Conference of the International Speech Communication Association*, September 2012, pp. 1179–1182.

[25] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the acoustics of emotion in audio: what speech, music, and sound have in common," *Frontiers in psychology*, vol. 4, p. 292, 2013.

[26] B. Schuller, M. Wöllmer, F. Eyben, and G. Rigoll, "Prosodic, spectral or voice quality? feature type relevance for the discrimination of emotion pairs," in *The role of prosody in affective speech*, S. Hancil, Ed., 2009, pp. 285–307.

[27] J. Antoni, "The spectral kurtosis: a useful tool for characterising non-stationary signals," *Mechanical systems and signal processing*, vol. 20, no. 2, pp. 282–307, 2006.

[28] S. Frühholz, W. Trost, and S. A. Kotz, "The sound of emotions—towards a unifying neural network perspective of affective sound processing," *Neuroscience & Biobehavioral Reviews*, vol. 68, pp. 96–110, 2016.

[29] J. Yamagishi, C. Veaux, K. MacDonald *et al.*, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2019.

[30] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and ESD," *Speech Communication*, vol. 137, pp. 1–18, 2022.

[31] S. R. Livingstone and F. A. Russo, "The Ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[32] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.

[33] 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. 17 022–17 033, 2020.

[34] T. Seehapoch and S. Wongthanavasu, "Speech emotion recognition using support vector machines," in *2013 5th international conference on Knowledge and smart technology (KST)*. IEEE, 2013, pp. 86–91.

[35] 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. Visited on 2023-03-03. [Online]. Available: https://arxiv.org/pdf/2203.12468.pdf

[36] E. Rodero, "Intonation and emotion: influence of pitch levels and contour type on creating emotions," *Journal of voice*, vol. 25, no. 1, pp. e25–e34, 2011.

[37] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, "MOSNet: Deep learning-based objective assessment for voice conversion," *Proc. Interspeech 2019*, pp. 1541–1545, 2019.

[38] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," OpenAI: San Francisco, CA, USA, Tech. Rep., 2022.

[39] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *Proc. Interspeech 2020*, pp. 3830–3834, 2020.

[40] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: Features and classification models," *Digital signal processing*, vol. 22, no. 6, pp. 1154–1160, 2012.