Anonymising Elderly and Pathological Speech: Voice Conversion Using DDSP and Query-by-Example

Anonymous submission to Interspeech 2024

Abstract

Speech anonymisation aims to protect speaker identity by modifying personal identifiers in speech while retaining linguistic content. Current methods fail to retain prosody and unique speech patterns found in some domains, such as the elderly and pathological speech, essential for remote health monitoring. To address this gap, we propose a voice conversion-based method (DDSP-QbE) using differentiable digital signal processing and query-by-example. The proposed method trained with novel losses provides the inductive bias to disentangle linguistic, prosodic, and domain representations, improving the model's ability to adapt to rare speech patterns. Extensive objective evaluation shows that DDSP-QbE significantly outperforms the current voice conversion state-of-the-art in retaining prosody and domain traits across diverse datasets, pathologies, and speakers while maintaining quality and anonymity. Experts validate these outcomes by analysing twelve clinically pertinent domain attributes.
Index Terms: speech anonymisation, voice conversion, DDSP

1. Introduction

The widespread adoption of cloud-based speech technologies has made remote health monitoring more accessible for elderly individuals and those with speech disorders [1, 2]. However, since these speech recordings contain highly sensitive personal data, it becomes crucial to anonymise the speech before sharing the data across systems [3]. Speech anonymisation aims to conceal the speaker's identity in recordings while maintaining linguistic content. For elderly and pathological (non-standard) speech data, it is crucial to also preserve the prosody and unique speech patterns of the domain, such as the hoarseness in dementia patients [4], for further analysis, diagnosis, and tracking of age- or disease-related changes [5].

Voice conversion (VC) based methods [6] have been successful in producing anonymised speech, where a source utterance is modified to sound like a *target* speaker. Most VC methods are based on generative adversarial network (GAN), trained with cycle-consistency loss, allowing training on non-parallel datasets [7]. These methods generate a transformed spectrogram conditioned on the target speaker's embeddings, which are learnt jointly with the linguistic embeddings during training. GAN-based methods overcome the buzzy voice problems caused by spectrum over-smoothing in variational autoencoder approaches [8]. This is attributed to the GAN's discriminator, which ensures that the generator produces realistic conversions matching the target speaker's style. Conversely, techniques like KNN-VC [9] opt for a strategy that eschews the direct learning of speaker or phonetic embeddings. Instead, KNN-VC utilises self-supervised model-derived representations, where the features from the source are mapped to a target speaker using K-nearest neighbours. Similarly, another method [10] used pre-trained speaker embeddings as one of the inputs in the differentiable digital signal processing (DDSP) [11]-based framework. DDSP integrates traditional DSP elements, such as filters and synthesiser oscillators, into deep neural networks, with the neural network itself generating the parameters. To the best of the authors' knowledge, our work is pioneering DDSP-based VC evaluation compared to other VC or anonymisation methods.

Most VC methods [12] have primarily been evaluated on standard data, featuring speech from young and healthy adults, with an emphasis on the naturalness and intelligibility of conversions. However, the recently introduced any-to-many method Emo-StarGAN [13] extends focus to prosody, achieving emotion-preserving conversions by employing losses derived from both hand-crafted and deep learning (DL) generated para-linguistic features, as well as an adversarial emotion classifier. However, the method fails to preserve the atypical speech patterns seen in stuttering [14], a common form of non-standard data.

Thus, we propose 'DDSP-QbE', an any-to-many VC method focused on preserving prosody and domain characteristics in speech anonymisation, even for unseen speakers from non-standard data. Our method builds on recent advancements in query-by-example (QbE) [15] and DDSP [16]. Our approach uses a subtractive harmonic oscillator-based DDSP synthesiser [16], inspired by the human speech production model [17], to incorporate an inductive bias for effective learning with limited data. By leveraging QbE, we directly derive target phonetic representations from source speech, thus bypassing the need to learn these representations during training. We introduce an inductive bias for prosody preservation by: (i) employing a novel loss function that utilises emotional speech to facilitate the separation of prosodic and linguistic features, and (ii) adding supplementary hand-crafted and DL-generated input features to the network, which have prosodic knowledge from the source utterance. For domain preservation, we employ loss functions based on acoustic properties that are crucial in clinical evaluations of voice disorders [18]. In addition to an objective evaluation, we conduct an in-depth subjective assessment of domain preservation, with speech pathologists assessing the retention of twelve clinically recognised measures for voice disorder. This analysis offers key insights, highlighting which domain aspects are preserved and which are not. Both objective and subjective analyses demonstrate DDSP-QbE's ability to anonymise speech while retaining prosodic and clinically pertinent domain features.

2. Differentiable Digital Signal Processing

In the DDSP framework, a synthesiser generates speech, with its parameters predicted by a neural network, enabling end-to-end training. However, this necessitates that the synthesiser's components are differentiable to enable back-propagation. A subtractive-based synthesiser model [16] is a harmonic-plus-noise model, which decomposes a monophonic sound into two components, harmonic y_h and stochastic y_s, through two stages. In the first stage,

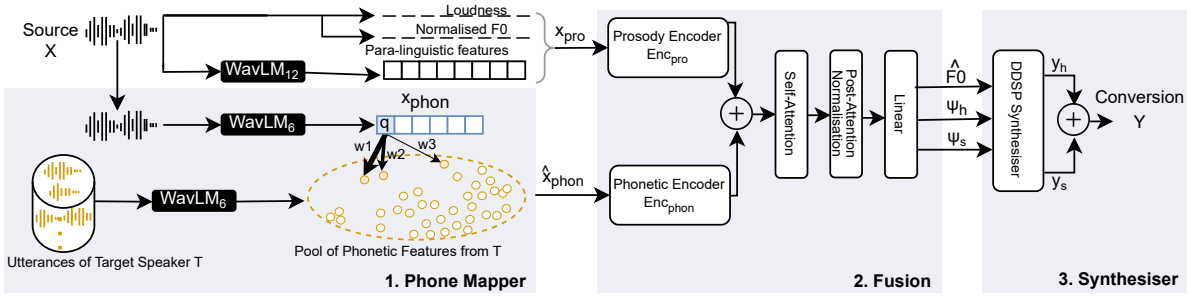

Figure 1: *The proposed framework DDSP-QbE . In the illustration, the number of matching candidates for computing $\hat{x}_{phon}$, is M=3.*

the synthesiser approximates y_h by a fundamental frequency (F0) constrained sawtooth signal, and derives the unfiltered harmonic component as $\tilde{y}_h(t) = \sum_{j=1}^{J} \frac{1}{j} sin(\phi_j(t))$, where $\phi_j(t)$ is the phase for the j^{th} harmonic at the t^{th} time instant. On the other hand, the unfiltered stochastic component $\tilde{y}_s$ is derived from a uniform noise signal $\eta \in [-1,1]$. In the second stage, y_h and y_s are acquired through individual spectral filtering using linear time-varying finite impulse response (LTV-FIR) filters $\psi_h \in \mathbb{R}^{F_h}$ and $\psi_s \in \mathbb{R}^{F_s}$, respectively [16]. The final audio is obtained as $Y = y_h + y_s$.

3. DDSP-QbE

The state-of-the-art VC approach Emo-StarGAN [13] successfully retains prosody during anonymisation for standard data. However, it struggles to preserve the distinctive speech patterns, or domain features characteristic of elderly and pathological speech, such as roughness, strain, or breathiness [4]. Our method aims to overcome these challenges by isolating these unique domain-specific patterns from personal speaker traits, even in situations with limited data.

3.1. Framework

As portrayed in Fig. 1, the proposed framework comprises three components: Phone Mapper, Fusion, and Synthesiser, with only the latter two requiring training.

Phone Mapper: Given a source utterance X, we obtain the phonetic representations $\hat{x}_{phon}$ for a *target* speaker from Phone Mapper. These representations maintain the linguistic content of the source utterance while sounding like the target speaker. Previous works [9, 19] have shown that representations from the 6th layer of the self-supervised model WavLM-Large (WLM$_6$) serve as a promising candidate to derive *latent* phonetic features as they achieved high performance in phone discrimination tasks [20]. Furthermore, similar-sounding phones tend to be closer together in this latent phone space [9]. With this understanding, we derive the target-phone features from the source-phone features through a QbE scheme, similar to the previous works [9, 15].

Initially, a pool of phone representations is generated per target speaker using WLM$_6$, computed frame-wise for all available utterances. For a source utterance X, phonetic representations x_{phon} are derived in a similar manner using WLM$_6$. Each source-phone representation acts as a query $q \in x_{phon}$ (shown in Fig. 1), which is replaced with $\hat{q} \in \hat{x}_{phon}$. $\hat{q}$ is computed as the weighted average of top-M phone representations similar to q from the target speaker's pool. Specifically, it is calculated as $\hat{q} = \frac{\sum_{i=1}^{M} m_i * w_i}{\sum_{i=1}^{M} w_i}$, where m_i represents the selected phone representation, and w_i is its corresponding weight. The weights $\{w_i\}_{i=1}^{M}$ are determined by applying a softmax function to the inverse of the cosine distance between q and the $\{m_i\}_{i=1}^{M}$ candidates. The weighting scheme is utilised to reduce the impact of outliers, thereby ensuring that phonetic representations closer in distance to q are given more importance.

Fusion: This component generates the parameters needed by the DDSP synthesiser, as shown in Fig. 1. It has been found that the phonetic representations from WLM$_6$ are intertwined with para-linguistic or prosodic cues [9]. This implies that during the *mapping phase*, the prosodic information is also replaced along with phonetic features, which is undesirable as we aim to preserve the prosody from the source. To address this 'prosody leakage' issue, we provide the network with prosodic features x_{pro}, extracted directly from the source utterance, along with the mapped phonetic features $\hat{x}_{phon}$ from the Phone Mapper, as shown in Fig. 1. The prosodic features x_{pro} are a combination of hand-crafted and DL-generated features, which are correlated with prosodic cues. The hand-crafted features considered are loudness and sample-wise z-normalised logarithmic F0 contour, aimed at capturing the speaker-independent pitch variations. Recent analyses [21, 22] have shown the middle (12th) layer of WavLM-Large (WLM$_{12}$) to perform well for para-linguistic related tasks, such as emotion classification. Therefore, we consider the representations from WLM$_{12}$ as the prosody-correlated deep feature as prosody provides important cues to emotion [23].

Initially, $\hat{x}_{phon}$ and x_{pro} are fed to their individual branches, phonetic encoder Enc$_{\text{phon}}$ and prosody encoder Enc$_{\text{pro}}$ respectively, as shown in Fig. 1. Each branch comprises two 1D convolutions with ReLU activation followed by group normalisation. The outputs from both branches are combined through element-wise addition and passed through a stack of three self-attention layers. Subsequently, this is followed by a shallow convolution stack with post-attention normalisation. Finally, a linear layer is used with dimensions matching the number of parameters Θ, required by the synthesiser. The architecture of the Fusion component is akin to the small Conformer architecture [24], which has demonstrated efficacy in capturing both local and global contexts in a sequence of acoustic features.

Synthesiser: We integrate the subtractive synthesiser proposed in [16] in our framework. The synthesiser produces the conversion Y using the parameters Θ derived from the Fusion module, where $\Theta = \{\widehat{F0}, \psi_h, \psi_s\}$. In our work, we incorporate the network predicted F0 ($\widehat{F0}$) unlike in the original DDSP work [11], which incorporates an additional inductive bias in the network and drives it to produce F0-consistent speech.

3.2. Domain and Prosody-Aware Losses

Relying solely on multi-resolution spectral losses, as employed in previous works [10, 11, 16], ensures high fidelity in the reconstruction, but fails to guarantee the preservation of non-linguistic features. To address this, we incorporate additional losses during training.

Jitter and Shimmer are clinically acclaimed indicators for assessing voice disorders [25]. Jitter refers to the variation or irregularity in the timing of consecutive periods of F0, reflecting the insta-

bility in vocal fold vibrations [26]. Therefore, jitter can be used to assess 'shakiness' or 'unsteadiness' correlates in the voice. We consider the jitter of the five-point period perturbation quotient (j_{ppq5}) as it is widely used in clinical studies for its ability to provide a more consistent assessment by considering neighbouring periods [27]. T_is are the extracted F0 period lengths and N is the number of F0 periods. We calculate jitter loss L_{jit} as the mean absolute error (MAE) between j_{ppq5} computed from the source X and the conversion Y.

$$j_{ppq5} = \frac{\frac{1}{N-1}\sum_{i=2}^{N-2}|T_i - \frac{1}{5}\sum_{n=i-2}^{i+2}T_n|}{\frac{1}{N}\sum_{i=1}^{N}T_i} * 100 \quad (1)$$

Shimmer, on the other hand, captures amplitude irregularities, which can be used to capture 'roughness' and 'breathiness' correlates in the voice [27, 28]. We consider the *local* shimmer s_{loc}, which is used in the clinically recognised measure to assess breathiness, acoustic breathiness index (ABI) [29]. s_{loc} is computed by taking the absolute average difference between the amplitudes of consecutive periods, and then dividing it by the average amplitude [27]. We calculate the shimmer loss L_{shim} as the MAE between the s_{loc} extracted from the source X and the converted speech Y samples.

Prosody Leakage from the mapping phase is addressed by introducing a loss formulation using emotional utterances. The Enc_{phon} encoder is specifically designed not to capture prosodic representations but only phonetic representations that sound like the target speaker. Thus, Enc_{phon} should generate comparable representations for two utterances (X_1 and X_2) by the same speaker, containing the same linguistic content but delivered with different emotions. To quantify the disparity between these representations, we introduce the loss L_{pro}, as depicted in Equation 2. This loss guides the encoder Enc_{phon} to capture non-prosodic representations, thereby facilitating the disentanglement of prosodic and non-prosodic features.

$$L_{pro} = |\text{Enc}_{\text{phon}}(\text{WLM}_6(X_1)) - \text{Enc}_{\text{phon}}(\text{WLM}_6(X_2))| \quad (2)$$

Training Objectives: We train the DDSP-QbE model with the multi-resolution spectral loss L_s and F0-related loss L_{f0}, as done in [16], along with our proposed losses. Therefore, we train the model DDSP-QbE using the objective function shown in Equation 3, where λ_s,λ_{jit},λ_{shim},λ_{pro} and λ_{f0} are hyper-parameters.

$$Loss = \lambda_s L_s + \lambda_{jit} L_{jit} + \lambda_{shim} L_{shim} + \lambda_{pro} L_{pro} + \lambda_{f0} L_{f0} \quad (3)$$

4. Experiment and Results

Due to the dearth of anonymisation methods for non-standard data, we use Emo-StarGAN (denoted as 'Emo'), known for emotion preservation as our baseline, which has been evaluated on German stuttering speech [14].

Three English datasets are considered for training and evaluation: (i) ESD [30]: standard data having annotations for five emotion classes, (ii) ADReSS_0 [31]: speech from healthy and dementia-diagnosed elderly speakers, and (iii) Sep-28k [32]: stuttering speech, containing annotations for 4 types of stuttering: block, interjection, word and sound repetitions. All utterances are resampled to 16 kHz and randomly split into training, validation, and test sets in proportions 0.8/0.1/0.1, respectively, ensuring no overlap of speakers across the sets. For a fair comparison, we train and evaluate all the models on the same splits. For ESD, we consider utterances from 6 speakers for training. For the non-standard datasets, we consider 20 speakers each from ADReSS_0 and Sep-28k, totalling $\approx$2.5 hours of non-standard data. Only excerpts featuring elderly and pathological speech are retained, guided by the dataset annotations. Speaker selections across all datasets are randomised, ensuring an even distribution of healthy, non-healthy, and various pathologies, as well as gender balance. Each model is trained using Adam optimiser with a learning rate of 0.002 for 150 epochs and batch size of 128 on an Nvidia A100 80 GB GPU. The WavLM representations for training are produced from 2-second audio. We use J=150 harmonics and the filter lengths as F_h = 176 and F_s = 80, and 5 resolutions $R = \{2^i\}_{i=6}^{10}$ for spectral loss L_s, with 75% overlapping among neighbouring frames. We set λ_{ms}=λ_{f0}=1.0, λ_{jit}=10, λ_{shim} = 0.1 and λ_{pro}=0.1. DDSP-QbE generated conversions faster than real-time, considering M=4 candidates from a phonetic feature pool, which is created for each target speaker from around 5 minutes of their utterances. We train a HiFiGAN vocoder on the training split, as described in [13], which is utilised by Emo to produce conversions. The remaining intricate training details will be open-sourced on GitHub alongside our code upon acceptance, including the dataset splits.

Evaluation Setup: We perform both objective and subjective evaluations for 4 source $\rightarrow$ target scenarios: (i) Elderly+Healthy $\rightarrow$ SD, (ii) Elderly+Dementia $\rightarrow$ SD, (iii) Stuttering $\rightarrow$ SD, and (iv) SD $\rightarrow$ SD, where SD denotes standard data. In all scenarios, the target speaker is chosen from SD, mirroring real-life situations due to the widespread availability of standard data. For evaluation, we use source utterances from the test split and randomly select 1000 conversions for each scenario, ensuring a balanced distribution across genders and types of pathologies.

Objective Evaluation: We assess domain preservation by classifying stuttering types and dementia. The classifiers are trained only on the original utterances in the training split, as done in [13]. We compare the class predicted for the converted speech with that of the original speech, considering the latter as the ground truth. Specifically for SD$\rightarrow$SD, we analyse whether pathologies such as dementia or stuttering are inadvertently introduced during the conversion process. For other performance metrics, we follow the methods detailed in previous work [13]: assessing overall quality through the predicted mean opinion score (pMOS) [33], prosody preservation via the pitch-correlation coefficient (PCC), intelligibility by measuring the character error rate (CER) using transcriptions from the Whisper medium-English model [34], and the strength of anonymisation through the equal error rate (EER).

Subjective Evaluation: We embrace two kinds of user studies[1], considering 160 randomly selected conversions per model due to the extensive time and cost involved in evaluating all of them. Each audio clip lasts for 4-13 seconds. The raters for the studies were unaware whether the samples were original or synthetically generated.

In the first study, two speech-language pathologists assessed the preservation of 12 domain attributes. These measures are typically used clinically to detect speech disorders [35, 36]: (i) Dysphonia: following the GRBAS scale [37], which provides an assessment of the severity level of a speech disorder, (ii) roughness: measures raspiness or harshness in voice, (iii) breathiness: measures lack of clarity in phonation, (iv) abnormal respiration, (v) articulation error, (vi) word repetition such as 'I will [will] go', (vii) sound repetition such as 'I am [pr-pr-pr-]prepared', (viii) omission or made-up words, (ix) block: unnatural pause or gasps of air, (x) interjection or filler-words such as 'um' or 'uh', (xi) strain: excessive effort or tension in phonation, and (xii) asthenia: lack of strength in the voice.

In the second study, 87 English-speaking participants assessed prosody, naturalness and anonymisation, as done in [13]. For prosody preservation, subjects compared the rhythm and intonation of a source utterance to the conversions by Emo and DDSP-QbE (ABX test), disregarding quality and content, and choosing the more similar or 'both equal' option. They rated naturalness on a 5-point MOS scale from 'bad' to 'excellent'. For anonymisation, raters marked speaker similarity on a 5-point scale (1: different, 5:

[1] conducted on Crowdee: `https://www.crowdee.com`

Table 1: *Objective evaluation results with 95% confidence intervals are presented. 'Domain Pr.' indicates Domain Preservation. The 'Type' column specifies special cases, such as the source speaker's domain, source and target gender groups, or 'All', which includes all sub-groups.*

Source	Type	Domain Pr. [%] ↑		PCC [$\times 10^2$] ↑		pMOS ↑		CER [%] ↓		EER [%] ↑	
		Emo	DDSP-QbE	Emo	DDSP-QbE	Emo	DDSP-QbE	Emo	DDSP-QbE	Emo	DDSP-QbE
All conversions	All	67.8±1.2	**78.1**±2.3	68.1±0.7	**77.4**±0.6	2.41±0.02	**3.39**±0.02	18.15±1.00	**1.04**±0.09	**50.16**±0.03	48.98±0.07
	Different gender	64.7±0.9	**76.2**±2.5	67.6±1.0	**75.4**±0.9	2.41±0.03	**3.36**±0.03	22.57±1.41	**1.05**±0.12	-	-
	Same gender	70.9±1.4	**80.0**±2.1	68.6±0.9	**79.4**±0.8	2.42±0.03	**3.41**±0.02	13.73±1.43	**1.02**±0.12	-	-
Elderly → SD	All	61.3±0.7	**79.7**±1.7	55.1±1.2	**70.6**±1.4	2.28±0.04	**3.24**±0.03	24.8±1.61	**2.41**±0.21	40.62±0.11	**43.41**±0.08
	Dementia	63.3±0.9	**78.9**±2.5	53.5±1.7	**68.4**±2.0	2.31±0.05	**3.29**±0.04	23.36±2.26	**2.40**±0.29	38.24±0.08	**40.41**±0.05
	Healthy	59.3±0.8	**80.5**±0.9	56.6±1.7	**72.8**±1.8	2.24±0.05	**3.20**±0.05	26.24±2.29	**2.42**±0.31	43.40±0.06	**44.39**±0.08
Stuttering → SD	All	64.3±2.7	**77.8**±2.2	80.0±0.7	**84.2**±0.5	2.41±0.04	**3.33**±0.03	19.07±4.01	**0.13**±0.05	**48.99**±0.02	48.85±0.05
	Block	59.7±3.7	**78.3**±2.2	80.3±1.7	**83.8**±1.0	2.41±0.08	**3.37**±0.07	16.55±4.00	**0.08**±0.08	-	-
	Word Repetition	73.2±1.7	**76.3**±2.1	78.9±1.6	**83.6**±0.9	2.38±0.09	**3.34**±0.06	16.91±3.93	**0.15**±0.10	-	-
	Sound Repetition	49.8±4.7	**78.0**±2.2	80.3±1.6	**84.6**±1.0	2.37±0.10	**3.22**±0.07	22.50±3.94	**0.20**±0.12	-	-
	Interjection	74.5±0.2	**78.2**±2.3	79.1±1.7	**83.6**±1.1	2.47±0.08	**3.37**±0.07	20.32±3.99	**0.08**±0.08	-	-
SD → SD	All	**77.8**±1.6	76.8±5.1	79.9±1.0	**81.4**±0.4	3.55±0.04	**3.88**±0.02	10.58±0.85	**0.57**±0.09	**52.29**±0.05	51.66±0.04

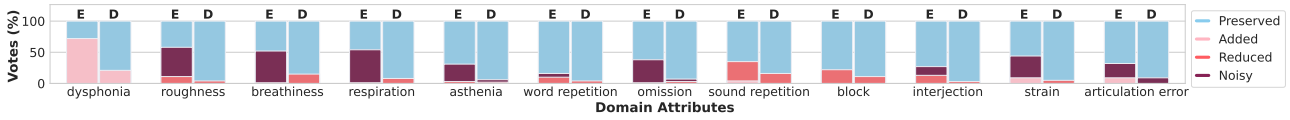

Figure 2: *Subjective evaluation results for domain preservation. Illustrates the percentage of each domain attribute being 'preserved', 'added' (or increased), 'removed' (or reduced), or 'noisy' (could not be assessed), for each model. 'E' represents Emo, and 'D' stands for DDSP-QbE.*

similar), after listening to a converted sample and another utterance from the source speaker. Each test was assessed by at least 3 raters, with those failing hidden traps twice excluded from analysis.

4.1. Results and Discussion

Table 1 indicates that DDSP-QbE surpasses Emo concerning all metrics except for anonymisation, where the EER scores are comparable. However in subjective assessment (refer to Table 2), DDSP-QbE conversions were perceived as less similar to the source speaker compared to those from Emo. DDSP-QbE significantly surpassed Emo in all metrics for objective and subjective evaluations, specifically for non-standard utterances. Concerning intelligibility, DDSP-QbE outperforms Emo significantly, as indicated by the CER scores ($p < 0.001$, paired t-test). Emo's reduced intelligibility

Table 2: *Subjective MOS and speaker similarity values with 95% confidence intervals. The Prosody column indicates the percentage of votes each model received. Source utterances' mean MOS is 3.4.*

Type	MOS ↑		Prosody [%] ↑		Speaker Similarity ↓	
	Emo	DDSP-QbE	Emo	DDSP-QbE	Emo	DDSP-QbE
All	3.54±0.17	**3.56**±0.13	34.8	**65.2**	3.01±0.38	**2.51**±0.41
Different gender	3.49±0.18	**3.55**±0.25	25.7	**74.3**	2.70±0.54	**2.21**±0.57
Same gender	3.53±0.24	**3.64**±0.18	43.9	**56.1**	3.22±0.60	**2.98**±0.35

stems from its challenge in adapting to uncommon speech patterns, such as irregular fricatives or plosives leading to prolongations, and repetitions of sounds like '[pr-pr-pr-]prepared', resulting in less clear distinctions between vowels and consonants. This issue also adversely affects Emo's scores for prosody preservation, which can also be seen in subjective results in Table 2. Further, Emo's preservation of prosody in elderly speech was much less, with a mean PCC of 55.1, compared to 80.0 for stuttering, likely due to more jitter and vocal tremors in elderly speech [38]. Emo was more inclined than DDSP-QbE to substitute the domain traits, such as roughness, breathiness, respiration, sound repetition, and blocks, with noise. This impacted Emo's Dysphonia score (see Fig. 2), indicating that pathological identifiers in Emo's conversions were enhanced, causing raters to perceive a greater severity of voice disorder compared to the original speech. DDSP-QbE successfully maintains most domain attributes but faces difficulties with breathiness, block and sound repetition, as seen in Fig. 2. Specifically, DDSP-QbE occasionally fails to accurately simulate the airflow release interruption in plosive sounds /p/, /b/, /t/, /d/, typical of stuttering speech. This suggests a need for advanced modelling techniques, incorporating metrics like cepstral peak prominence (CPP) and Acoustic Breathiness Index (ABI) [29].

The ablation study presented in Table 3 shows that removing

Table 3: *Ablation results with 95% confidence intervals shown.*

Method	Domain Pr. [%] ↑	PCC [$\times 10^2$] ↑	pMOS ↑	CER [%] ↓	EER [%] ↑
Full DDSP-QbE	78.1±2.3	77.4±0.6	3.39±0.02	1.04±0.09	48.98±0.07
$\lambda_{pro}=0$	73.7±1.3	67.9±0.6	3.40±0.02	4.56±0.53	50.99±0.09
$\lambda_{jit}=0$	69.7±0.9	74.1±0.6	3.36±0.02	6.81±0.51	49.88±0.05
$\lambda_{shim}=0$	70.7±1.1	73.6±0.6	3.35±0.02	7.89±0.54	50.60±0.03

L_{pro} diminishes prosody preservation, while the individual removal of jitter and shimmer losses compromises the model's domain preservation capabilities. Interestingly, the absence of each component reduces the intelligibility of the conversions as well.

5. Conclusion

We propose the first speech anonymisation technique that successfully maintains the prosody and distinct speech characteristics prevalent in the elderly and pathological speech. Our approach utilises a subtractive DDSP synthesiser combined with query-by-example (QbE), possesses only 0.4% of the trainable parameters of Emo, and is trained on just ≈2.5 hours of non-standard data. Despite this, it shows a superior ability to generalise to rare speech patterns, showing the effectiveness of the proposed inductive biases. The detailed subjective assessments, including the one focusing on clinically relevant attributes, indicate that DDSP-QbE substantially surpasses the baseline in preserving both prosody and domain-specific traits across diverse speech patterns seen in non-standard data, target speakers, and genders. Looking ahead, our goal is to improve the retention of complex features such as breathiness and to adapt our approach to additional languages and speech disorders.

6. References

[1] P. Kulkarni, O. Duffy, J. Synnott, W. G. Kernohan, R. McNaney *et al.*, "Speech and language practitioners' experiences of commercially available voice-assisted technology: web-based survey study," *JMIR Rehabilitation and Assistive Technologies*, vol. 9, no. 1, p. e29249, 2022.

[2] S. Ahmed, M. Qaosar, R. W. Sholikah, and Y. Morimoto, "Early dementia detection through conversations to virtual personal assistant," in *2018 AAAI Spring Symposium Series*, 2018.

[3] E.-M. Schomakers and M. Ziefle, "Privacy concerns and the acceptance of technologies for aging in place," in *International Conference on Human-Computer Interaction*, 2019, pp. 313–331.

[4] S. Taylor, C. Dromey, S. L. Nissen, K. Tanner, D. Eggett, and K. Corbin-Lewis, "Age-related changes in speech and voice: spectral and cepstral measures," *Journal of Speech, Language, and Hearing Research*, vol. 63, no. 3, pp. 647–660, 2020.

[5] X. Liang, J. A. Batsis, Y. Zhu, T. M. Driesse, R. M. Roth, D. Kotz, and B. MacWhinney, "Evaluating voice-assistant commands for dementia detection," *Computer Speech & Language*, vol. 72, p. 101297, 2022.

[6] 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.

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

[8] M.-A. Georges, J.-L. Schwartz, and T. Hueber, "Self-supervised speech unit discovery from articulatory and acoustic features using VQ-VAE," in *Proc. INTERSPEECH 2022*. ISCA, 2022.

[9] M. Baas, B. van Niekerk, and H. Kamper, "Voice Conversion With Just Nearest Neighbors," in *Proc. INTERSPEECH 2023*, 2023, pp. 2053–2057.

[10] S. Nercessian, "End-to-end zero-shot voice conversion using a DDSP vocoder," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2021, pp. 1–5.

[11] J. Engel, C. Gu, A. Roberts *et al.*, "DDSP: Differentiable digital signal processing," in *International Conference on Learning Representations*, 2019.

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

[13] S. Ghosh, A. Das, Y. Sinha, I. Siegert, T. Polzehl, and S. Stober, "Emo-StarGAN: A semi-supervised any-to-many non-parallel emotion-preserving voice conversion," in *Proc. INTERSPEECH 2023*. ISCA, 2023.

[14] J. Hintz, S. Bayerl, Y. Sinha, S. Ghosh, M. S. Sebastian, K. R. Stober, and I. Siegert, "Anonymization of stuttered speech–removing speaker information while preserving the utterance," in *Proc. 3rd Symposium on Security and Privacy in Speech Communication*, 2022, pp. 41–45.

[15] C. Lea, D. Yee, J. Narain, Z. Huang, L. Tooley, J. P. Bigham, and L. Findlater, "Latent phrase matching for dysarthric speech," in *Proc. INTERSPEECH 2023*. ISCA, 2023.

[16] D.-Y. Wu, W.-Y. Hsiao, F.-R. Yang, O. Friedman, W. Jackson, S. Bruzenak, Y.-W. Liu, and Y.-H. Yang, "DDSP-based singing vocoders: A new subtractive-based synthesizer and a comprehensive evaluation," in *ISMIR 2022*, 2022.

[17] G. Fant, "The source filter concept in voice production," *STL-QPSR*, vol. 1, no. 1981, pp. 21–37, 1981.

[18] K. Nishikawa, H. Kawano, R. Hirakawa, and Y. Nakatoh, "Analysis of prosodic features and formant of dementia speech for machine learning," in *2022 5th International Conference on Information and Computer Technologies (ICICT)*. IEEE, 2022, pp. 173–176.

[19] 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.

[20] E. Dunbar, N. Hamilakis, and E. Dupoux, "Self-supervised language learning from raw audio: Lessons from the Zero Resource Speech Challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1211–1226, 2022.

[21] Z. Zhu and Y. Sato, "Deep investigation of intermediate representations in self-supervised learning models for speech emotion recognition," in *2023 IEEE International conference on acoustics, speech, and signal processing workshops (ICASSPW)*. IEEE, 2023, pp. 1–5.

[22] Y. Li, Y. Mohamied, P. Bell, and C. Lai, "Exploration of a self-supervised speech model: A study on emotional corpora," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 868–875.

[23] 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.

[24] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.

[25] N. Sripriya, S. Poornima, R. Shivaranjani, and P. Thangaraju, "Non-intrusive technique for pathological voice classification using jitter and shimmer," in *2017 International Conference on Computer, Communication and Signal Processing (ICCCSP)*. IEEE, 2017, pp. 1–6.

[26] J. P. Teixeira and A. Gonçalves, "Algorithm for jitter and shimmer measurement in pathologic voices," *Procedia Computer Science*, vol. 100, pp. 271–279, 2016.

[27] B. Barsties v. Latoszek, J. Mayer, C. R. Watts, and B. Lehnert, "Advances in clinical voice quality analysis with VOXplot," *Journal of Clinical Medicine*, vol. 12, no. 14, p. 4644, 2023.

[28] C. B. Barcelos, P. A. L. Silveira, R. L. V. Guedes, A. N. Gonçalves, L. D. S. Slobodticov, and E. C.-d. Angelis, "Multidimensional effects of voice therapy in patients affected by unilateral vocal fold paralysis due to cancer," *Brazilian journal of otorhinolaryngology*, vol. 84, pp. 620–629, 2018.

[29] B. B. v. Latoszek, Y. Maryn, E. Gerrits, and M. De Bodt, "The Acoustic Breathiness Index (ABI): a multivariate acoustic model for breathiness," *Journal of Voice*, vol. 31, no. 4, pp. 511–e11, 2017.

[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. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Detecting cognitive decline using speech only: The ADReSSo challenge," *medRxiv*, 2021.

[32] C. Lea, V. Mitra, A. Joshi, S. Kajarekar, and J. P. Bigham, "SEP-28k: A dataset for stuttering event detection from podcasts with people who stutter," ICASSP 2021.

[33] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, "HiFi++: A unified framework for bandwidth extension and speech enhancement," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[34] 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*. PMLR, 2023, pp. 28 492–28 518.

[35] I. Martínez-Nicolás, T. E. Llorente, F. Martínez-Sánchez, and J. J. G. Meilán, "Ten years of research on automatic voice and speech analysis of people with alzheimer's disease and mild cognitive impairment: a systematic review article," *Frontiers in Psychology*, vol. 12, p. 620251, 2021.

[36] H. R. Perez and J. H. Stoeckle, "Stuttering: clinical and research update," *Canadian family physician*, vol. 62, no. 6, pp. 479–484, 2016.

[37] M. Hirano, "Clinical examination of voice," *Disorders of human communication*, vol. 5, pp. 1–99, 1981.

[38] B. G. Schultz, S. Rojas, M. St John, E. Kefalianos, and A. P. Vogel, "A cross-sectional study of perceptual and acoustic voice characteristics in healthy aging," *Journal of Voice*, vol. 37, no. 6, pp. 969–e23, 2023.