Improving Voice Quality in Speech Anonymization With Just Perception-Informed Losses

Suhita Ghosh*, Tim Thiele*, Frederic Lorbeer*, Sebastian Stober
Artificial Intelligence Lab (AILab), Otto-von-Guericke-University, Magdeburg, Germany
{suhita.ghosh, stober}@ovgu.de

Abstract

The increasing use of cloud-based speech assistants has heightened the need for effective speech anonymization, which aims to obscure a speaker's identity while retaining critical information for subsequent tasks. One approach to achieving this is through voice conversion. While existing methods often emphasize complex architectures and training techniques, our research underscores the importance of loss functions inspired by the human auditory system. Our proposed loss functions are model-agnostic, incorporating handcrafted and deep learning-based features to effectively capture representations about speech quality. Through objective and subjective evaluations, we demonstrate that a VQVAE-based model, enhanced with our perception-driven losses, surpasses the vanilla model in terms of naturalness, intelligibility, and prosody while maintaining speaker anonymity. These improvements are consistently observed across various datasets, languages, target speakers, and genders.

1 Introduction

Over the last few years, cloud-based speech devices, such as voice assistants, have become indispensable in life [1]. However, this also poses increasing privacy threats, as speech data contains sensitive information encompassing health, affiliations, and other private information about the speaker [2, 3]. Therefore, speech anonymization becomes pertinent, which hides the personal identifiers in the speech while retaining the linguistic content. Voice conversion (VC) is one of the ways to achieve speech anonymization, where the source utterance is modified to sound like another 'target' speaker. In cases where the response of a speech device is driven by the end-user's emotional state, preserving prosody becomes crucial, such as in health monitoring systems that adjust alert urgency based on detected stress or anxiety to ensure timely intervention.

Research on statistical modelling approaches [4, 5, 6, 7] provided the groundwork for the development of deep learning (DL)-based VC techniques, significantly advancing the state-of-the-art in VC research [8, 9]. Most of the VC methods are based on generative adversarial networks (GANs) [10], which produce natural-sounding conversions. This is due to the discriminator's role in guiding the generator to create conversions consistent with the target speaker's characteristics. Current GAN-based VC methods [11, 12] are trained with over seven losses in addition to adversarial losses, complicating training due to instability in the optimization process and heightened sensitivity to hyperparameter choices [13]. In contrast to GANs, variational autoencoders (VAEs) [14, 15] offer a clear advantage by providing a well-defined likelihood function, ensuring a more stable training than GANs. These methods typically disentangle the speaker and content embeddings using a reconstruction loss and relevant constraints to remove speaker information.

One notable variant of VAE, the vector-quantized VAE (VQVAE) [16], uses a discrete distribution over a codebook instead of a continuous distribution, which is potentially a more intuitive approach

*These authors contributed equally to this work.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

given that language is inherently discrete, such as speech being represented as a sequence of phonemes. Recent VQVAE-based approaches organize the latent embeddings by the phonetic content [17] or arrange them in a hierarchical manner [18], to capture the different semantic levels of speech across various temporal scales. However, VQVAE-based approaches have garnered limited traction compared to GANs, due to their tendency to produce averaged outputs, leading to a buzzy-sounding voice [8]. One reason for this issue is that VQVAEs typically use an element-wise loss function in the output space [19]. These losses do not penalize the regions which are pertinent to the human auditory system [20]. This leads to low-quality reconstructions with dampened prosody.

Thus, we propose novel 'perception-informed' losses to enhance the quality of speech produced by a VC model. These losses aim to introduce an inductive bias that may aid in achieving higher fidelity reconstructions. We propose two kinds of losses: handcrafted feature-based and representation-driven. The handcrafted feature-based loss is computed on formants, which represent the resonance frequencies of the vocal tract and are crucial for defining the characteristics of vowel sounds, playing a significant role in speech perception and identifying phonetic elements. Drawing from perceptual losses in audio enhancement tasks [21], our representation-driven losses emphasize the aspects of sound most critical to human listeners, thereby improving the perceptual fidelity of the generated speech. Although our proposed loss functions can be integrated into any model, we consider a VQVAE-based model to demonstrate their effectiveness, as training a VQVAE model is generally simpler than training a GAN-based model. We extensively evaluate our approach using various datasets, languages, target speakers, and genders. We demonstrate through objective and subjective tests that our proposed loss functions generalize well and significantly enhance speech quality across different scenarios.

2 Vanilla VQVAE

VQVAE achieves VC by transforming the source mel-spectrogram x using the speaker embedding of the target speaker, typically learned during training [18]. There are three key components of VQVAE:

1. **Encoder** Enc takes a mel-spectrogram x and maps it to a discrete latent variable $z = Enc(x)$, which is received by the vector quantization layer.
2. **Vector Quantization** layer, also known as the codebook C_c, sits between the encoder Enc and the decoder Dec. This layer consists of learnable vectors representing embeddings that capture speaker-independent content information. The encoder's output z is used to select the most similar vector q from this codebook based on Euclidean distance. This selected vector q is then passed to the decoder in place of z. Since the nearest vector selection is non-differentiable, the straight-through re-parameterization trick is applied to compute the discrete latent vector q_{st}, as $q_{st} = z + sg(q - z)$, where sg is the stop-gradient operator [16].
3. **Decoder** Dec receives two inputs: the content embedding q_{st} and a speaker embedding e_s, selected from speaker codebook $C_s = \{e_s\}_{s=1}^{S}, s \in 1...S$. The speaker codebook C_s is jointly optimized with the other model parameters during training through back-propagation. Using both of these inputs q_{st} and e_s, the decoder generates the transformed mel-spectrogram $x_{dec} = Dec(q_{st}, e_s)$. Therefore, VC using VQVAE can be achieved by just replacing the source speaker embedding with the target speaker embedding.

We use a hierarchical-based VQVAE [18] as our baseline, which employs L=3 levels of vector quantization layers to capture speech representations at varying semantic depths (e.g., phoneme, syllable, word), enhancing reconstruction quality. The model is trained with the loss functions shown in Equation 1: reconstruction loss for preserving linguistic content, codebook loss to ensure that the encoded representations remain close to the discrete codebook vectors, and commitment loss ensures latent representations remain consistent with specific codebook vectors [16]. Each loss is weighed by hyperparameter λ.

$$L_{\text{vanilla}} = \lambda_{\text{recon}} \|x - x_{dec}\|_2^2 + \lambda_{\text{code}} \sum_{l=1}^{L} \|sg[z_l] - q_l\|_2^2 + \lambda_{\text{com}} \sum_{l=1}^{L} \|z_l - sg[q_l]\|_2^2 \tag{1}$$

3 Perception-Informed Losses

Recent VC research has mainly focused on enhancing architectures to improve synthesized speech quality, often leading to complex models and overfitting [8]. In contrast, our approach introduces novel loss functions, applicable to any model, that aim to capture speech quality in line with human perception.

Handcrafted Feature-Based Loss: Formants serve as a concise descriptor of the spectral content of vowels, efficiently capturing important speech features with minimal parameters [22]. Phonetically, formants are resonant frequencies that are characteristic of the shape of the human vocal tract during speech production [23] and are also affected by prosody [24]. F1 is the lowest frequency formant, followed by F2, F3, and so on. Typically, F1 and F2 suffice for vowel identification [25]. However, F3 adds an important layer of detail that enhances the precision of vowel identification [26], aids in consonant distinction and provides critical information for speaker identification [27] and speech intelligibility [28]. We compute the formant loss $L_{formant}$ as shown in Equation 2, where $\Phi^k(.)$ represents the kth formant. Here, K and N denote the total number of formants and frames, respectively.

$$L_{\text{formant}} = \frac{1}{K \times N} \sum_{k=1}^{K} \sum_{n=1}^{N} (\Phi^k(x_n) - \Phi^k(x_{dec_n}))^2 \tag{2}$$

Representation-Driven Losses: Intermediate representations from self-supervised deep learning models capture a wide range of speech features, such as tonal quality, prosody, clarity, and background noise [29], which are vital for assessing speech quality. Different network layers capture varying levels of abstraction, from basic acoustic features to more abstract representations like phonemes [30]. Embeddings from supervised models trained for quality-related tasks offer richer information than standard element-wise loss functions [31]. Consequently, we compute the quality discrepancy as shown in Equation 3, which is a general representation-driven loss, calculated on the activations α^j from the jth layer of a quality-based perceptual network.

$$L_{\text{DL}} = \frac{1}{|J|} \sum_{j \in J} \frac{1}{N} \sum_{n=1}^{N} (\alpha^j(x_n) - \alpha^j(x_{dec_n}))^2 \tag{3}$$

We consider two kinds of representation-driven losses:

1. **Mean Opinion Score (MOS) Loss:** The MOS is a widely used subjective metric for assessing the quality or naturalness of speech [32]. However, incorporating human annotators to rate speech conversion during the training process is impractical. To address this, we use a neural network, Net_{mos}, as a proxy for human evaluation, which is trained to predict the MOS score of a speech audio signal. Specifically, we employ the model proposed in [33], which consists of a fine-tuned Wav2Vec2.0 model [34] with a regression head added to the encoded features, resulting in a total of $|J| = 4$ layer activations. The corresponding loss function $L_{\text{DL=mos}}$ is defined in Equation 3, where the activations α are produced by the Net_{mos} model.
2. **WavLM Loss:** WavLM [35] is a state-of-the-art model for comprehensive speech processing tasks, demonstrating leading performance on SUPERB benchmarks [36] in areas such as speaker verification and diarization. Studies like [37, 38] highlight WavLM's capability to extract meaningful phoneme embeddings, with similar-sounding phonemes clustering in its latent space. The later layers of WavLM show reduced predictive power for pitch and prosody [39], while the embeddings from layer J=6 are highly correlated with phoneme identification [37]. Therefore, we compute the WavLM loss $L_{\text{DL=wavlm}}$ using the activations from the 6th layer, resulting in$|J| = 1$ in Equation 3.

Training Objectives: Put together, the full objective function of our proposed approach consists of the following terms that are weighted by λ_i, where $i \in \{\text{recon,code,comm,mos,wavlm,formant}\}$:

$$L = \lambda_{\text{recon}} L_{\text{recon}} + \lambda_{\text{code}} L_{\text{code}} + \lambda_{\text{com}} L_{\text{com}} + \lambda_{\text{mos}} L_{\text{DL=mos}} + \lambda_{\text{wavlm}} L_{\text{DL=wavlm}} + \lambda_{\text{formant}} L_{\text{formant}} \tag{4}$$

4 Experiment Details

We use three datasets: VCTK [40] and LibriSpeech [41] for English utterances, and mlsGerman [42] for German. Utterances are re-sampled to 16 kHz. The vanilla hierarchical VQVAE without perception-informed losses serves as our baseline model (M_{base}), while our proposed model (M_{PL}) incorporates these losses. All models are trained on the same splits and evaluated on the same test set. Log mel-spectrograms are used as input to the models. The models are optimized using the Adam optimizer with a cyclic learning rate, ranging from 5×10^{-4} to 2×10^{-3}. The models are trained from scratch, employing early stopping with predicted mean opinion score (pMOS) [33] on the validation set as the stopping criterion. A pre-trained HiFiGAN vocoder [12] is used to generate the waveform from the model's output. Additional details are provided in the appendix.

4.1 Evaluation

We evaluate the baseline model (M_{base}) and our approach (M_{PL}) across three scenarios using both objective and subjective measures:

1. **English→English**: Both source and target speakers are English-speaking, evaluated within the same corpus (VCTK→VCTK) and across different corpora (LibriSpeech→VCTK). We also assess inter-accent conversion (Canadian, American, British).
2. **English→ German**: German utterances are converted using English VCTK target speakers.
3. **German→ German**: Both source and target speakers are German, using the mlsGerman dataset.

In each scenario, we have 10 source speakers, each providing 10 utterances, and 10 target speakers. The speakers are selected randomly, ensuring a disjoint set and balanced gender distribution, leading to 1000 total conversions.

Objective Measures: Intelligibility is measured by character error rate (CER) using the transcriptions from Whisper [43] *medium-english* model for the English conversions and *medium* model for German conversions. For anonymization, we measure the equal error rate (EER) using the speaker verification model ECAPA-TDNN [44], as in [12]. We avoid using predicted MOS (pMOS) for quality evaluation, as in [12], as it is used as a loss function in our model and could lead to overfitting, as discussed in [45]. Instead, we rely on subjective testing for quality assessment.

Subjective Evaluation Setup: We evaluated quality, prosody preservation, intelligibility and anonymization by user studies via the Crowdee platform[1]. We evaluated a random selection of 100 conversions for each of the three scenarios, as assessing all conversions would be both time-consuming and costly. 72 online participants had taken part in the studies. For English→English and German→German scenarios, only native speakers of English and German, respectively, were allowed to participate. In the German→VCTK (English) scenario, native German speakers who were proficient in English were considered. Participants rated quality (naturalness) on a scale from 1 (poor) to 5 (excellent). They compared intonation and stress patterns between the original and converted samples for prosody preservation. Intelligibility was assessed by selecting the most intelligible sample between the two conversions, for the same source utterance. Anonymization was evaluated by rating the similarity on a scale ranging from 1 (different) to 5 (similar), between a converted sample and another utterance from the same speaker. Each task was rated by at least 3 subjects, who were unaware of whether the samples were original or converted. Trap questions and anchoring examples were used to ensure accuracy, and raters who failed trap questions twice were excluded.

5 Results and Discussion

Overall, our proposed method M_{PL}, significantly enhances intelligibility compared to the baseline M_{base}, as demonstrated in Table 1. This improvement is also accompanied by improvement in naturalness corroborated by the MOS ratings from user studies, as shown in Figure 1. Additionally, 83% of the conversions using the proposed model were rated as more intelligible than those from the baseline. In terms of speaker anonymization, there is a modest increase in EER from 41.07% to 43.21% across all scenarios, which is similarly reflected in the speaker similarity scores from user studies. For prosody preservation, our approach significantly outperforms the baseline, with 83.2% of participants favouring the proposed model having perception-informed losses, as seen in Figure 1. Similar trends are observed for within-corpus scenario VCTK→VCTK, where the mean CER showed a significant improvement from 73.32% to 45.49% with the incorporation of the proposed losses, as detailed in Table 1. These improvements are also observed in cross-gender (refer to Appendix) and cross-accent conversions within the corpus. For prosody preservation and intelligibility, our model received significantly higher support with 83% and 85% of the votes, respectively. Furthermore, in inter-accent conversions, we observe a change in accent after the VC, where the converted sample adopted the accent of the target speaker, potentially leading to better anonymization. In the cross-corpus scenario LibriSpeech→VCTK, similar trends are observed for all metrics.

In the German to English (VCTK) conversion, intelligibility did not improve much compared to intra-lingual conversions, as shown in Table 1. Listening to samples[2] reveals that using English target

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

[2] Audio samples available at: `https://shorturl.at/mqSrs`

Table 1: Objective evaluation results are presented with 95% confidence intervals.

Source	**All conversions (All) /Accent-wise**	**CER** [%] ↓		**EER** [%] ↑	
		M_{base}	M_{PL}	M_{base}	M_{PL}
All Conversions	All	71.33±0.40	**53.09**± 0.67	41.07	**43.21**
VCTK → VCTK	All	73.32±0.79	**45.49**±1.13	37.88	**38.17**
	American → British	73. 31±1.49	**45.32**±2.24	36.89	**39.10**
	Canadian → British	72.02±1.39	**48.89**±1.97	37.10	**42.23**
	British → British	74.30±1.26	**43.07**±1.68	36.07	**38.02**
LibriSpeech → VCTK	All	70.16±0.64	**53.50**±1.12	**38.26**	38.00
German → VCTK	All	77.44±0.56	**75.21**±0.75	42.35	**44.84**
German → German	All	64.41±0.91	**38.15**±1.02	49.63	**51.46**

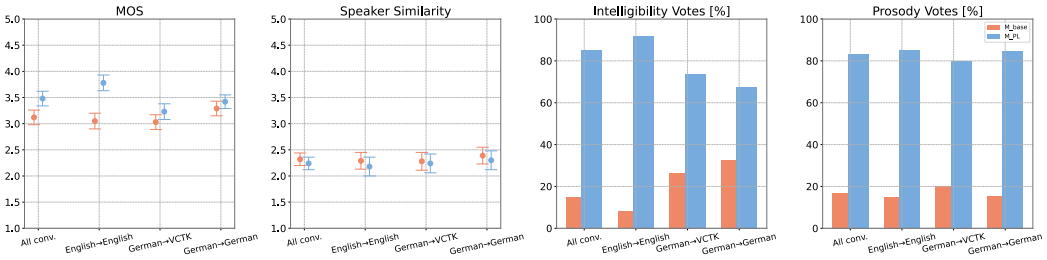

Figure 1: User study results for different scenarios and all conversions (All conv.). The Speaker Similarity plot indicates the similarity between the source and converted utterances (lower is better). The MOS plot shows the naturalness ratings from the user study (higher is better). The Prosody and Intelligibility Votes plots show the percentage of votes each model received. The mean MOS of the original files is 3.54.

speakers introduced an English accent in the conversions, failing to preserve the original intonation. This occurred because German phonemes that do not exist in English were replaced by similar English sounds. For example, the German uvular fricative [ʁ] in "Rad" became the alveolar [ɹ] as in "run". The German fricatives [ç] ("ich") and [x] ("ach") were replaced by [ʃ] ("sh") and [k] ("cat"). This likely occurs because the VQVAE model is trained on English data, substituting German sounds with the closest English equivalents. However, this accent shift aids anonymization, potentially leading to a higher EER compared to the VCTK→VCTK scenario. For German→German conversions, a similar improvement is observed as in the English→English scenario. However, regarding naturalness, M_{PL} shows less improvement compared to the English→English scenario, as indicated in the user study results (MOS score in Figure 1). This might be attributed to the formant prediction network Net_{formant} being trained solely on English data. Consequently, the network may not accurately capture German vowel nuances, leading to a mismatch in vowel prediction that results in converted German speech sounding less authentic, as reflected in subjective evaluations.

6 Conclusion

We present model-agnostic perception-informed losses as an innovative approach to enhance the quality of voice conversion (VC) for speech anonymization without increasing model complexity. By integrating quality-related knowledge into the training process through handcrafted acoustic features and deep learning representations, our framework significantly improves the performance of a vanilla hierarchical VQVAE-based model. Augmented solely by our proposed loss functions, the model shows notable enhancements in naturalness, intelligibility, and prosody preservation across diverse conversion scenarios, including cross-corpus conversions, varying genders, accents, and languages. Objective and subjective evaluations validate these results, highlighting the importance of incorporating speech-specific features within the loss function, rather than increasing model complexity. Looking ahead, we plan to develop loss functions to specifically target and reduce the graininess observed in some conversions.

7 Acknowledgements

This research has been supported by the Federal Ministry of Education and Research of Germany through project Emonymous (project number S21060A) and Medinym (focused on AI-based anonymization of personal patient data in clinical text and voice datasets). We would also like to acknowledge Frank Dreyer for developing the baseline used in this study, which provided a valuable foundation for our work.

References

[1] D. S. Zwakman, D. Pal, and C. Arpnikanondt, "Usability evaluation of artificial intelligence-based voice assistants: The case of Amazon Alexa," *SN Computer Science*, vol. 2, pp. 1–16, 2021.

[2] 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, 2021.

[3] M. Haase, J. Krüger, and I. Siegert, "User perspective on anonymity in voice assistants," in *International Conference on Human-Computer Interaction*. Springer, 2023, pp. 156–166.

[4] 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.

[5] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5660–5664.

[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] 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.

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

[9] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. S. Kudinov, and J. Wei, "Diffusion-based voice conversion with fast maximum likelihood sampling scheme," in *ICLR*, 2024.

[10] 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.

[11] A. Das, S. Ghosh, T. Polzehl, and S. Stober, "StarGAN-VC++: Towards emotion preserving voice conversion using deep embeddings," in *Speech Synthesis Workshop (SSW)*. ISCA, 2023.

[12] 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 *Proceedings of the Conference of the International Speech Communication Association Interspeech 2023*. ISCA, 2023.

[13] D. Saxena and J. Cao, "Generative adversarial networks (GANs) challenges, solutions, and future directions," *ACM Computing Surveys (CSUR)*, vol. 54, no. 3, pp. 1–42, 2021.

[14] J. Williams, Y. Zhao, E. Cooper, and J. Yamagishi, "Learning disentangled phone and speaker representations in a semi-supervised VQ-VAE paradigm," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7053–7057.

[15] M.-A. Georges, J.-L. Schwartz, and T. Hueber, "Self-supervised speech unit discovery from articulatory and acoustic features using VQ-VAE," in *Proceedings of the Conference of the International Speech Communication Association Interspeech 2022*. ISCA, 2022.

[16] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[17] S. Ding and R. Gutierrez-Osuna, "Group latent embedding for vector quantized variational autoencoder in non-parallel voice conversion," in *Proceedings of the Conference of the International Speech Communication Association Interspeech 2019*. ISCA, 2019.

[18] T. V. Ho and M. Akagi, "Non-parallel voice conversion based on hierarchical latent embedding vector quantized variational autoencoder," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*. ISCA, 2020, pp. 140–144.

[19] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[20] I. Ananthabhotla, S. Ewert, and J. A. Paradiso, "Towards a perceptual loss: Using a neural network codec approximation as a loss for generative audio models," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1518–1525.

[21] F. G. Germain, Q. Chen, and V. Koltun, "Speech denoising with deep feature losses," in *Proceedings of the Conference of the International Speech Communication Association Interspeech 2019*. ISCA, 2019.

[22] J. Benesty, M. M. Sondhi, and Y. A. Huang, Eds., *Springer Handbook of Speech Processing*, 1st ed., ser. Springer Handbooks. Springer Berlin, Heidelberg, 2008.

[23] I. Titze, *Principles of Voice Production*. Prentice Hall, 1994.

[24] Y. Mo, J. Cole, and M. Hasegawa-Johnson, "Prosodic effects on vowel production: evidence from formant structure," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[25] M. Gordon, "Erik r. thomas. 2011. Sociophonetics. An Introduction," *English World-Wide*, vol. 34, 10 2013.

[26] T. Smit, F. Türckheim, A. Jakob, and R. Mores, "Deviation of perceived vowel quality as a result of f3 manipulation," in *Proc. 20th Int. Congress of Acoustics, ICA*, 2010, pp. 23–27.

[27] N. Almaadeed, A. Aggoun, and A. Amira, "Text-independent speaker identification using vowel formants," *Journal of Signal Processing Systems*, vol. 82, pp. 345–356, 2016.

[28] A. Amano-Kusumoto, J.-P. Hosom, A. Kain, and J. M. Aronoff, "Determining the relevance of different aspects of formant contours to intelligibility," *Speech Communication*, vol. 59, pp. 1–9, 2014.

[29] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 914–921.

[30] T. Nagamine, M. L. Seltzer, and N. Mesgarani, "Exploring how deep neural networks form phonemic categories." in *Proceedings of the Conference of the International Speech Communication Association Interspeech 2015*, 2015, pp. 1912–1916.

[31] S. Ghosh, A. Krug, G. Rose, and S. Stober, "Perception-aware losses facilitate ct denoising and artifact removal," in *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)*. IEEE, 2021, pp. 1–6.

[32] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives," *Multimedia Systems*, vol. 22, pp. 213–227, 2016.

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

[34] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12 449–12 460.

[35] 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, Oct. 2022.

[36] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin *et al.*, "SUPERB: Speech processing universal performance benchmark,"

in *Proceedings of the Conference of the International Speech Communication Association Interspeech 2021*. ISCA, 2021.

[37] M. Baas, B. van Niekerk, and H. Kamper, "Voice conversion with just nearest neighbors," in *Proceedings of the Conference of the International Speech Communication Association Interspeech 2023*, 2023, pp. 2053–2057.

[38] S. Ghosh, M. Jouaiti, A. Das, Y. Sinha, T. Polzehl, I. Siegert, and S. Stober, "Anonymising elderly and pathological speech: Voice conversion using DDSP and query-by-example," in *Proceedings of the Conference of the International Speech Communication Association Interspeech 2024*, June 2024.

[39] G.-T. Lin, C.-L. Feng, W.-P. Huang, Y. Tseng, T.-H. Lin, C.-A. Li, H.-y. Lee, and N. G. Ward, "On the utility of self-supervised models for prosody-related tasks," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 1104–1111.

[40] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit," [sound], 2017.

[41] 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 (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[42] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A large-scale multilingual dataset for speech research," *ArXiv*, vol. abs/2012.03411, 2020.

[43] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *ICML*. PMLR, 2023, pp. 28 492–28 518.

[44] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proceedings of the Conference of the International Speech Communication Association Interspeech 2020*, 2020, pp. 3830–3834.

[45] D. de Oliveira, S. Welker, J. Richter, and T. Gerkmann, "The PESQetarian: On the relevance of Goodhart's law for speech enhancement," in *Proceedings of the Conference of the International Speech Communication Association Interspeech 2024*, June 2024.

[46] 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.

[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[48] L. Deng, X. Cui, R. Pruvenok, Y. Chen, S. Momen, and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 06 2006, pp. I – I.

8 Appendix

8.1 Detailed Objective Evaluation

Table 2 shows a more detailed version of the objective evaluation portrayed in Table 1, where gender-wise information is also mentioned.

Table 2: Objective evaluation results are presented with 95% confidence intervals.

Source	**All conversions (All) /Accent-wise**	**CER** [%] ↓		**EER** [%] ↑	
		$M_{\textbf{base}}$	$M_{\textbf{PL}}$	$M_{\textbf{base}}$	$M_{\textbf{PL}}$
All Conversions	All	71.33±0.40	**53.09**± 0.67	41.07	**43.21**
VCTK → VCTK	All	73.32±0.79	**45.49**±1.13	37.88	**38.17**
	American → British	73. 31±1.49	**45.32**±2.24	36.89	**39.10**
	Canadian → British	72.02±1.39	**48.89**±1.97	37.10	**42.23**
	British → British	74.30±1.26	**43.07**±1.68	36.07	**38.02**
	Different gender	73.05±1.12	**46.24**±1.58	-	-
	Same gender	73.60±1.12	**44.72**±1.61	-	-
LibriSpeech → VCTK	All	70.16±0.64	**53.50**±1.12	**38.26**	38.00
	Different gender	69.71±0.92	**53.73**±1.61	-	-
	Same gender	70.63±0.88	**53.27**±1.58	-	-
German → VCTK	All	77.44±0.56	**75.21**±0.75	42.35	**44.84**
	Different gender	76.93±0.79	**75.71**±1.03	-	-
	Same gender	77.96±0.80	**74.70**±1.08	-	-
German → German	All	64.41±0.91	**38.15**±1.02	49.63	**51.46**
	Different gender	65.87±1.27	**39.17**±1.44	-	-
	Same gender	63.00±1.31	**37.17**±1.45	-	-

8.2 Ablation Study

We perform ablation studies to assess the contribution of each loss component. Table 3 demonstrates that formant L_{formant} individually contributes the most to naturalness and intelligibility. This suggests that calculating loss on specific frequency components effectively enhances the overall quality of VC. These components correspond to the resonant frequencies of the vocal tract, which are essential for perceiving vowel sounds and overall intelligibility. Further, listening to the samples reveals that the model not trained with L_{formant} has the worst prosody preservation. Removing L_{formant} loss (when using $L_{\text{DL=wavlm}}$ + $L_{\text{DL=mos}}$) significantly increases the CER from 50.02% to 66.85%, highlighting the critical role of formants in speech intelligibility.

Table 3: Ablation study results with 95% confidence intervals shown on the VCTK → VCTK conversion setup. pMOS is the MOS score predicted by Net_{mos}.

Method	**pMOS** ↑	**CER** [%] ↓	**EER** [%] ↑
M_{base}	1.59 ±0.04	73.32±0.79	37.88
M_{PL}	**3.56** ±0.02	**45.49** ±1.13	**38.17**
L_{formant}	**3.13** ±0.01	**50.67** ±1.02	37.07
$L_{\text{DL=wavlm}}$	3.02 ±0.02	51.74 ±3.02	37.89
$L_{\text{DL=mos}}$	2.33 ±0.04	68.17 ±2.02	**37.97**
L_{formant} + $L_{\text{DL=mos}}$	2.71 ±0.02	50.02 ±1.02	**38.16**
L_{formant} + $L_{\text{DL=wavlm}}$	**3.47** ±0.01	**49.28** ±1.08	38.12
$L_{\text{DL=wavlm}}$ + $L_{\text{DL=mos}}$	2.34 ±0.03	66.85 ±1.02	37.91

Interestingly, individually L_{formant} and $L_{\text{DL=wavlm}}$ have a greater positive impact on MOS scores compared to $L_{\text{DL=mos}}$, indicating that these losses better capture the aspects of speech that influence perceived quality. One reason $L_{\text{DL=mos}}$ underperforms compared to $L_{\text{DL=wavlm}}$ is that WavLM was trained on a much larger corpus to capture more generic speech representations, encompassing noise, distortion, natural variations in pitch, loudness, and other factors. In contrast, the MOS network is specifically trained to predict MOS scores, focusing solely on naturalness.

The combination of L_{formant} and $L_{\text{DL=wavlm}}$ significantly improves the pMOS and CER compared to the baseline, nearly reaching the performance of the model incorporating all losses M_{PL}. We also note that the anonymization capability of the model is not significantly affected by the removal of the loss components individually or in combination. This indicates that the mechanisms responsible for anonymization are robust and independent of the specific losses used to enhance naturalness and intelligibility.

8.3 Training Details

We trained all models on log mel-spectrograms with 80 mel bands, generated from 2-second audio clips. For STFT parameters, we used a hop length of 320 and a window length of 1024.

For scenarios involving English-speaking target speakers, our models were trained on approximately 5 hours of English utterances from 20 speakers in the VCTK dataset, with the data divided into a 90:10 split for training and validation. In cases requiring German-speaking targets, we utilized around 10 hours of German utterances from 20 speakers in the mlsGerman dataset, allocating 80% for training and 20% for validation.

The number of trainable parameters in all the voice conversion models is the same, as we only augment the vanilla model with our proposed losses. Training with all three perception-aware losses required approximately 2 days on average to complete on a 80GB A100 GPU. We set $\lambda_{\text{recon}} = 1$, $\lambda_{\text{code}} = 1$, $\lambda_{\text{comm}} = 3$, $\lambda_{\text{mos}} = 0.1$, $\lambda_{\text{wavlm}} = 0.1$, and $\lambda_{\text{formant}} = 10^6$, ensuring that all loss terms were within the same order of magnitude. We incorporated $L_{\text{DL=mos}}$ from epoch 0 and $L_{\text{DL=wavlm}}$, L_{formant} from epoch 45 into the training based on empirical observations obtained during the development phase.

We used a pre-trained HiFiGAN [46] vocoder from [12] to generate the waveform from the mel-spectrogram, which produced a one-minute long waveform from the converted mel-spectrogram in 0.1 seconds on the A100.

We trained the Net_{formant} model to derive the F1, F2, and F3 values needed to compute the L_{formant} loss. The formant network consists of a transformer encoder architecture as proposed in [47], additionally featuring a regression head with three output neurons that predict based on the encodings of the input for each time frame. As training data, we used the VTR dataset [48], comprising 538 manually formant-annotated utterances from domain experts who ensured balance across phonetic contexts, speakers, genders, and dialects in the English language. We used the default VTR parameters for pre-processing and achieved a final MSE of 3.16.