PROVABLE SPEECH ATTRIBUTE CONTROL VIA LATENT INDEPENDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Conditional generation and disentangled representation learning are central to controlled generation across audio, vision, and multimodal domains. However, despite strong empirical progress, particularly in speech style transfer, most existing approaches rely on heuristic objectives and architectural choices, offering limited theoretical understanding of when and why reliable attribute control is achievable. In this work, we develop a formal framework for speech attribute conversion and provide a theoretical analysis of sufficient conditions for exact and consistent transfer. Our analysis focuses on a deterministic autoencoder setting augmented with an independence constraint between the learned latent representation and the controllable attribute. Under explicit population-level assumptions about the data-generating process, we establish guarantees linking reconstruction, independence, and the feasibility of attribute manipulation while preserving task-relevant content. We further show how the theoretical framework translates into practice by proposing a practical voice conversion method that directly implements its core principles. Experimental evaluations on voice and pitch conversion tasks demonstrate the applicability of the theoretical analysis to real-world speech conversion settings and show that the resulting method achieves competitive performance against existing approaches.

1 INTRODUCTION

Understanding and controlling structured variability in complex data, such as speech, is a fundamental goal in machine learning. In many applications, observed signals are governed by multiple underlying factors (e.g., linguistic content, speaker identity, emotional tone), and isolating and controlling these components is crucial for tasks such as personalized speech synthesis, cross-lingual voice cloning, and emotion-aware dialogue systems. For instance, cross-lingual voice conversion systems aim to generate speech in a new language while preserving speaker identity (Yang et al., 2021; Zhang et al., 2019), while emotion transfer models seek to modify the affective content of speech without altering who is speaking (Gao et al., 2018; Zhou et al., 2023). These applications assume that meaningful latent representations, such as content and style, can be reliably recovered and manipulated in a disentangled and stable manner. However, ensuring that these latent variables are both identifiable and robustly recoverable remains a fundamental challenge, especially without direct supervision.

Recent advancements in deep learning have spurred significant interest in autoencoder-based (AE) approaches for transforming specific attributes of speech signals, such as speaker identity, emotion, or linguistic content. Techniques such as voice conversion, where the identity of a speaker is altered while preserving the linguistic content (Qian et al., 2019; Polyak & Wolf, 2019), and emotion conversion, which transforms emotional expression without affecting speaker identity (van Niekerk et al., 2022; Gao et al., 2018; Zhou et al., 2021), exemplify the potential of autoencoders for disentangling and manipulating distinct speech attributes.

Despite these empirical successes, a gap remains in the theoretical understanding of whether the true underlying latent variables can be accurately recovered from the observed data and auxiliary inputs. Specifically, it is unclear under what conditions a trained model ensures that the latent representation inferred by the encoder corresponds to the original unobserved variable that generated the data.

We propose an AE framework for structured variable conversion and provide theoretical guarantees for recovering the task-relevant latent variable under explicit assumptions on the generative process.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Our setting is a special case of nonlinear Independent Component Analysis (ICA), where we recover only a single latent component, while the remaining components are known and provided as auxiliary information. Although the recovered component may itself be a nonlinear mixture of multiple real-world factors, we show that it is sufficient for accurately converting the input variable. This relaxation of the complete identifiability requirement, which recovers only the relevant component, enables a more tractable and practical approach, especially when disentangling all underlying factors is unnecessary. Our results thus extend the applicability of nonlinear ICA theory (e.g., Hyvarinen et al. (2019); Khemakhem et al. (2020)) to focused representation learning in conditional generation and variable conversion tasks, such as those found in speech processing. elsec:problem

Let $\mathbf{S} \in \mathbb{S}$ be a latent random variable representing speech content, and let $\mathbf{C}_i \in \mathbb{C}_i$ be a collection of observed random variables representing different speech characteristics such as speaker identity, pitch, emotion, and others. We use $\mathbf{C}$ to denote the random vector $(\mathbf{C}_1, ..., \mathbf{C}_k)^T$. Let $\mathcal{P}$ be a probability distribution over $\mathbb{S} \times \mathbb{C}$ having density p. We assume that $\mathbf{S}$ and $\mathbf{C}$ are independent, i.e., p is an outer product of the marginal densities $p = p_S \times p_C$.

Let $\mathbf{X}$ be an observed variable generated as an invertible function f of $\mathbf{S}, \mathbf{C}$, i.e., $\mathbf{X} = f(\mathbf{S}, \mathbf{C})$. We denote by $\mathcal{P}_X$ the pushforward distribution induced by f, with the corresponding density p_X. Our statistical task is for any (new) realizations $\mathbf{s}, \mathbf{c}$ to synthesize samples $\mathbf{x} = f(\mathbf{s}, \mathbf{c})$.

The independence assumption above is a modeling assumption on the underlying content variable $\mathbf{S}$ and the chosen controllable attributes $\mathbf{C}$, not a claim that arbitrary speech factors are independent in natural corpora. Real speech contains correlations between timbre, pitch range, accent, emotion, and rhythm. Our guarantee therefore applies to settings in which the user-specified condition captures the attribute to be changed, and the remaining content variable is approximately invariant to it. Strongly coupled attributes require either redefining the condition set or accepting that exact conversion is outside the scope of the theorem.

2 Related Work

Recent efforts in voice conversion and expressive speech modeling have explored diverse directions, including emotional control, disentangled representations, and lightweight architectures. A prominent line of work relies on conditional flow-matching or prompt-based architectures. For example, Pan et al. (2025) introduces an encoder–decoder framework for emotional speech conversion in which content is extracted using a pretrained ASR HybridFormer (Yang et al., 2023), while emotion is represented via embeddings from a pretrained contrastive language–audio model. In this setting, the choice of external representations enforces independence between content and emotion by design, while reconstruction is learned with a conditional flow-matching decoder. Similarly, Zuo et al. (2025) employs discrete pitch tokens and target-speaker prompts for expressive conversion within a flow-matching framework, and Wang et al. (2025a) adopts a token-based in-context learning strategy for controllable voice conversion. In all these approaches, independence between attributes is implicitly assumed through architectural choices rather than explicitly enforced or analyzed, which limits theoretical insight into when and why conversion is guaranteed to succeed.

Another class of methods builds on vector-quantized autoencoders. For instance, Zhang et al. (2025) proposes a VQ-VAE-based conversion model that progressively disentangles timbre, style, and linguistic content via discrete latent representations, while Yao et al. (2025) demonstrates controllable zero-shot conversion using conditional flow matching. Although these approaches align naturally with the encoder–decoder paradigm, discreteness alone does not guarantee statistical independence between the learned content representation and the conditioning attributes, nor does it establish conversion correctness. Moreover, codebook learning introduces additional inductive biases and optimization complexity that are difficult to characterize theoretically.

Another line of work focuses on self-supervised disentanglement without relying on external pre-trained models. Cai et al. (2025) separates linguistic content from speaker style using large-scale self-supervised training on unlabeled corpora, while Liu et al. (2025) develops a streaming voice conversion system based on differentiable digital signal processing. While these approaches show strong empirical performance, they rely on large models and extensive data and do not provide formal guarantees linking disentanglement to the correctness of synthesis or conversion.

In contrast, our approach emphasizes reproducibility, accessibility, and theoretical clarity. We design a lightweight architecture trained on moderate-scale public datasets, providing a practical and analyzable baseline for expressive voice conversion. Our design is inspired by Polyak et al. (2021), where conversion uses discrete units extracted by pretrained content, pitch, and speaker encoders. We generalize this framework by introducing an end-to-end trainable autoencoder that operates directly on speech units containing both linguistic and acoustic information. The encoder is optimized to predict speech units independent of the conditioning variables, and training is performed jointly with the decoder.

A complementary research direction relies on nearest-neighbor search in the embedding space. Baas et al. (2023) propose KNN-VC, a few-shot voice conversion method in which embeddings from a target speaker's reference set guide the conversion process. While effective in low-data regimes, inference complexity grows with reference-set size, and correctness depends implicitly on the structure of the embedding space. Our method can be viewed as distilling the implicit invariances exploited by KNN-VC into an explicit latent variable learned by an autoencoder, thereby eliminating the need for nearest-neighbor search and enabling efficient one-shot or short-utterance conversion. Moreover, we generalize autoencoder-based methods by introducing an independence objective that applies to arbitrary conditioning attributes, such as speaker identity or emotion.

Following this direction, Shan et al. (2024) introduces a Phoneme Hallucinator that generates diversified phonemes from short target-speaker references, improving sample quality but inheriting the reliance on synthesized reference samples and nearest-neighbor search, which increases inference latency. In contrast, our method reduces runtime complexity from quadratic to linear with respect to the number of speech samples. Although KNN-VC avoids training a conversion module, our training procedure remains simple and involves two loss terms: independence and reconstruction. Finally, Wang et al. (2025b) extends the KNN-VC framework by clustering semantically similar representations using 2D structural entropy (Huang et al., 2025) and organizing embeddings into a graph. Unlike these approaches, our framework explicitly links latent structure, independence, and conversion correctness through a unified theoretical analysis.

3 General Framework

We introduce the Independence Conditional Autoencoder (ICAE), an effective non-probabilistic framework that avoids priors and posterior inference. ICAE is trained by jointly optimizing two complementary objectives: (1) accurate reconstruction of the input signal, and (2) encouraging statistical independence between the learned latent representation and a given set of conditioning variables. Formally, the objective is defined as:

$$\min_{\boldsymbol{\theta}_e \cup \boldsymbol{\theta}_d} \frac{1}{N} \sum_{\mathbf{x},\mathbf{c}} \left[\mathcal{R}(\mathbf{x}, d_{\boldsymbol{\theta}_d}(e_{\boldsymbol{\theta}_e}(\mathbf{x}), \mathbf{c})) + \lambda \mathcal{I}(\mathbf{c}, e_{\boldsymbol{\theta}_e}(\mathbf{x})) \right], \tag{1}$$

where $e(\mathbf{x}) = \hat{\mathbf{s}}$ is the latent representation produced by the encoder e, and $d(e(\mathbf{x}), \mathbf{c}) = d(\hat{\mathbf{s}}, \mathbf{c}) = \hat{\mathbf{x}}$ is the reconstructed sample generated by the decoder d. The model is parametrized by a learnable set of parameters, i.e., $\boldsymbol{\theta}_e$ for the encoder and $\boldsymbol{\theta}_d$ for the decoder, such that $\boldsymbol{\theta} = \boldsymbol{\theta}_e \cup \boldsymbol{\theta}_d$. The term $\mathcal{R}(\mathbf{x}, \hat{\mathbf{x}})$ captures the reconstruction discrepancy, while $\mathcal{I}(\mathbf{c}, \hat{\mathbf{s}})$ quantifies the dependence between the latent representation and the conditioning variables. Minimizing this dependence term encourages, but does not by itself certify, independence in finite data. The trade-off between the two objectives is controlled by a scalar $\lambda > 0$.

Our task can be seen as a simplified variant of nonlinear ICA (Hyvarinen et al., 2019). In this setting, only a single latent component is unknown and must be recovered, while the remaining components are observed. Recent work by Shaham et al. (2022) provides theoretical guarantees for latent component analysis; we extend these results to synthesis. Specifically, we show that exact identification of the latent variable is unnecessary for conversion; it is sufficient to recover it up to an arbitrary invertible transformation. We formally show that this level of recovery meets the requirements of the conversion task.

Our primary objective is to learn a decoder d that approximates the generative function f, thereby enabling the synthesis of novel samples $\mathbf{X}' \sim p_X$. Complementing this, we also train an encoder e that recovers a transformed version of $\mathbf{S}$, denoted as $\hat{\mathbf{S}} = e(\mathbf{X})$. Together, these mappings unify

synthesis and analysis: the decoder generates new data consistent with the underlying distribution, while the encoder ensures latent recovery, guiding and validating the conversion task.

In the following sections, we present a theoretical analysis of the proposed framework, defined by the objective function in (1), and show that it provides guarantees for style conversion. We also introduce a specific method that instantiates this framework and demonstrates, both theoretically and empirically, that it effectively performs speaker-identity style conversion. These results strengthen the practical applicability of the proposed theoretical framework for real-world style conversion tasks.

4 Theoretical Analysis

4.1 Ideal Variable Conversion

For the following claims, we assume we have an ICAE model that achieves perfect reconstruction and zero dependence.

Assumption 1 (**Perfect Model Convergence**). *For a given trained model that is represented by the map $d(e(\mathbf{x}), \mathbf{c})$:*

$$\mathcal{R}(\mathbf{x}, \hat{\mathbf{x}}) = 0, \mathcal{I}(\mathbf{c}, \hat{\mathbf{s}}) = 0, \forall \mathbf{x} \in \mathbb{X}, \textit{ s.t. } \mathbf{x} = f(\mathbf{s}, \mathbf{c})$$

Assumption 1 should be understood as an idealized population-level condition, corresponding to the limit of infinite data, sufficient model capacity, and exact optimization. It does not assert that such conditions are achieved during finite-sample training. This perspective is standard in identifiability analyses of nonlinear ICA, where recovery guarantees are derived under population-level objectives (e.g., exact likelihood or contrastive loss minimization in the infinite-data regime; see Hyvarinen et al. (2019)). Thus, the theory should be read as a set of sufficient conditions and a design principle for conversion models, rather than as a claim that every real speech pipeline exactly satisfies the assumptions. The reconstruction assumption is relaxed in Section 4.2 to derive an error bound.

Assumption 2 (**Finite Discrete Latent Variable**). *The latent variable $\mathbf{S}$ is discrete and admits a finite set of possible realizations,* $\mathrm{supp}(\mathbf{S}) = \{\mathbf{s}_1, \ldots, \mathbf{s}_M\}, M \geq 2$, $P_{\mathbf{S}}(\mathbf{s}_m) > 0$.

This assumption aligns with common preprocessing steps in speech representation learning, such as clustering or quantizing continuous features. The identifiability result below applies to this finite discrete setting; the continuous latent variant used in experiments is included as an empirical comparison rather than as a case covered by the discrete proof.

Assumption 3 (**Asymmetric Latent Prior**). *For the latent variable $\mathbf{S}$ in the generative model $\mathbf{X} = f(\mathbf{S}, \mathbf{C})$, the symmetry group $\{g : \mathbb{S} \to \mathbb{S} \mid g \text{ is invertible}, \ g_{\#}P_{\mathbf{S}} = P_{\mathbf{S}}\}$ contains only the identity map. Equivalently, in the finite discrete setting of Assumption 2, all point masses are distinct. We define the corresponding* identifiability margin

$$\Delta_p := \min_{i \neq j} |p_i - p_j| > 0, \qquad p_{\max} := \max_i p_i.$$

Since $\mathbf{S}$ is unobserved, Appendix J (Figure 2) examines the empirical distribution of a proxy variable $\tilde{\mathbf{S}} \approx \mathbf{S}$ derived from real-world speech data. This analysis suggests that symmetry is unlikely in practice, providing empirical support for the approximate satisfaction of Assumption 3.

The next lemma shows that the learned latent representation $\hat{\mathbf{S}}$ preserves all information about the true content $\mathbf{S}$ up to an invertible transformation, ensuring that no information relevant for conversion is lost.

Lemma 1. *For a given encoder e, there is an invertible map $T : \mathbb{S} \to \hat{\mathbb{S}}$ that such that $T(\mathbf{S}) = \hat{\mathbf{S}}$.*

The proof can be found in Appendix D. The next proposition shows that the decoder mimics the unknown variable generation function f composed on top of some unknown invertible map $T : \mathbb{S} \to \hat{\mathbb{S}}$.

Proposition 1. *Let T be the transformation from Lemma 1. Then, for every admissible condition $\mathbf{c}$ and every $\hat{\mathbf{s}} \in T(\mathbb{S}) = \mathrm{supp}(\hat{\mathbf{S}})$,*

$$d(\hat{\mathbf{s}}, \mathbf{c}) = f(T^{-1}(\hat{\mathbf{s}}), \mathbf{c}).$$

Thus, on the learned latent support, the decoder equals the true generative map composed with the inverse latent reparametrization T^{-1}. We provide the proof in Appendix E. The following corollary is the main consequence of the above results, showing that the model performs exact conditional conversion.

Corollary 1. *Under Assumptions 1, 2, and 3, the conversion of random variable* $\mathbf{X}$ *is guaranteed, i.e. for a given set of samples* $\{(\mathbf{x}, \mathbf{c}), (\mathbf{x}', \mathbf{c}')\}$ *such that* $\mathbf{x} = f(\mathbf{s}, \mathbf{c})$ *and* $\mathbf{x}' = f(\mathbf{s}, \mathbf{c}')$ *it holds that:* $d(e(\mathbf{x}), \mathbf{c}') = \mathbf{x}'$.

The corollary holds for both seen and unseen realizations of $\mathbf{X}$ within the support covered by the assumptions. It guarantees that preserving the condition-independent variable $\mathbf{S}$ up to an invertible transformation T, while replacing the specified condition $\mathbf{C}$, allows us to manipulate samples $\mathbf{X}$: we change the attributes represented in $\mathbf{C}$ while preserving the remaining information represented by $\mathbf{S}$.

4.2 Model Convergence Implies low Conversion Error

Section 4.1 establishes exact conversion under perfect reconstruction and exact independence. We now show that this guarantee degrades gracefully when both conditions are only approximately satisfied, and derive an error bound for the conversion task.

Assumption 4 (Uniform Reconstruction Error). *There exists* $\epsilon \geq 0$ *such that*

$$\|d(e(\mathbf{x}), \mathbf{c}) - \mathbf{x}\|_2^2 \leq \epsilon, \qquad \forall\, \mathbf{x} = f(\mathbf{s}, \mathbf{c}) \in \mathbb{X}.$$

Assumption 5 (L-Lipschitz Decoder). *There exists* $L \geq 0$ *such that*

$$\|d(\hat{\mathbf{s}}_1, \mathbf{c}) - d(\hat{\mathbf{s}}_2, \mathbf{c})\|_2 \leq L\|\hat{\mathbf{s}}_1 - \hat{\mathbf{s}}_2\|_2,$$

for all admissible latent values $\hat{\mathbf{s}}_1, \hat{\mathbf{s}}_2$ *and* $\mathbf{c} \in \mathrm{supp}(\mathbf{C})$.

For each condition value $\mathbf{c} \in \mathrm{supp}(\mathbf{C})$, let $Q_{\mathbf{c}} := P_{\hat{\mathbf{S}}|\mathbf{C}=\mathbf{c}}, Q := P_{\hat{\mathbf{S}}}$, denote the conditional and marginal distributions of the learned latent representation. The next assumption requires four intuitive properties. First, different realizations of the latent variable $\mathbf{S}$ must remain distinguishable in the observation space under the same condition, as expressed by $\gamma_X > 0$. Second, the reconstruction error must be sufficiently small relative to this separation, so that two distinct latent realizations cannot be mapped to the same learned representation while still being reconstructed accurately. Third, the learned representation should be approximately independent of the conditioning variable: for every condition value $\mathbf{c}$, the conditional distribution $Q_{\mathbf{c}}$ should remain close to the marginal latent distribution Q, with δ measuring the largest such deviation. In particular, $\delta = 0$ recovers exact independence, $\hat{\mathbf{S}} \perp\!\!\!\perp \mathbf{C}$. Finally, changing the condition while keeping the same underlying latent realization $\mathbf{s}$ should not move its learned representation by an arbitrarily large amount; $D_{\hat{S}}$ provides a finite bound on this cross-condition variation. Together, these conditions ensure that the encoder preserves distinctions between different values of $\mathbf{S}$ while remaining approximately insensitive to changes in $\mathbf{C}$, which is the key property required for robust conversion.

Assumption 6 (Approximate Independence and Regularity). *There exists* $\delta \geq 0$ *such that* ***(1)*** $\gamma_X := \inf_{\substack{\mathbf{c}\in\mathrm{supp}(\mathbf{C}) \\ i\neq j}} \|f(\mathbf{s}_i, \mathbf{c}) - f(\mathbf{s}_j, \mathbf{c})\|_2 > 0$, ***(2)*** $\epsilon < \frac{\gamma_X^2}{4}$, ***(3)*** $\mathrm{TV}(Q_{\mathbf{c}}, Q) := \frac{1}{2}\sum_{\hat{\mathbf{s}}\in\hat{\mathbb{S}}} |Q_{\mathbf{c}}(\hat{\mathbf{s}}) - Q(\hat{\mathbf{s}})|$ *and* $\sup_{\mathbf{c}\in\mathrm{supp}(\mathbf{C})} \mathrm{TV}(Q_{\mathbf{c}}, Q) \leq \delta$, ***(4)*** $D_{\hat{S}} := \sup_{\substack{\mathbf{s}\in\mathrm{supp}(\mathbf{S}) \\ \mathbf{c},\mathbf{c}'\in\mathrm{supp}(\mathbf{C})}} \|e(f(\mathbf{s}, \mathbf{c})) - e(f(\mathbf{s}, \mathbf{c}'))\|_2$ *is finite.*

Proposition 2 (Robust Conversion Error Bound). *Under Assumptions 2, 3, 4, 5, and 6, for every* $\mathbf{c}, \mathbf{c}' \in \mathrm{supp}(\mathbf{C})$, *let* $\mathbf{x}_0 = f(\mathbf{S}, \mathbf{c})$, $\mathbf{x}' = f(\mathbf{S}, \mathbf{c}')$, $\hat{\mathbf{x}}' = d(e(\mathbf{x}_0), \mathbf{c}')$, *then the conversion error* $\epsilon_{\mathrm{conv}}(\mathbf{S}, \mathbf{c}, \mathbf{c}') := \|\mathbf{x}' - \hat{\mathbf{x}}'\|_2^2$ *is bounded:*

$$\mathbb{E}_{\mathbf{S}}\left[\epsilon_{\mathrm{conv}}\right] \leq 2\epsilon + \frac{8L^2 D_{\hat{S}}^2 p_{\max}}{\Delta_p}\delta. \tag{2}$$

Corollary 2 (Robust Conversion Error Bound for Discrete Latent $\hat{\mathbf{S}}$). *Assuming that each* $\hat{\mathbf{s}} \in \mathbb{R}^K$ *is a probability vector such that:* $\sum_{k=1}^K \hat{s}_k = 1$, *the conversion error from Proposition 2 becomes:*

$$\mathbb{E}_{\mathbf{S}}\left[\epsilon_{\mathrm{conv}}\right] \leq 2\epsilon + \frac{16L^2 p_{\max}}{\Delta_p}\delta. \tag{3}$$

Thus, reconstruction error ϵ and residual dependence δ are the two quantities that must be driven toward zero, while L, $D_{\hat{S}}$, and Δ_p determine how strongly these errors are amplified.

The quantity Δ_p plays a particularly important role as an *identifiability margin*. It measures the minimum separation between the prior probabilities of distinct latent realizations. When Δ_p is large, assigning a latent value to a code associated with another value produces a detectable change in the conditional distribution of $\hat{\mathbf{S}}$. Approximate distributional independence therefore restricts how often such condition-dependent reassignments can occur, yielding approximate paired invariance. Decoder smoothness then propagates this latent agreement to the output space, producing the conversion bound above. Conversely, as $\Delta_p \to 0$, the latent prior becomes increasingly symmetric and the guarantee weakens. Consequently, for fixed L, $D_{\hat{S}}$, $p_{\max}$, and $\Delta_p > 0$, training the model towards $\epsilon \to 0, \delta \to 0$ implies $\mathbb{E}_{\mathbf{S}}[\epsilon_{\text{conv}}] \to 0$, recovering the exact-conversion regime of Section 4.1. The detailed proof is provided in Appendix G.

Having established theoretical guarantees under our assumptions, we now present a practical method that approximates the independence objective in real-world speech conversion tasks using a tractable empirical dependence penalty.

5 Practical Instantiation of the Framework

5.1 Model Architecture

We next demonstrate how the theoretical framework developed above can be instantiated in a practical speech attribute conversion system. To this end, we introduce the **Independence-based Voice Conversion (IVC)** method, illustrated in Figure 1. The encoder e maps the input sequence to a shallow latent representation of dimension $D \times T$. The latent representation is driven by a reconstruction loss and a dependence penalty (HSIC) that encourages statistical independence from the conditioning variable. The decoder d reconstructs the input embeddings.

We focus on the discrete latent representation $\hat{\mathbf{S}}$, which directly matches the assumptions used in our main identifiability result. The encoder outputs logits $e(\mathbf{x}) = \hat{\mathbf{s}} \in \mathbb{R}^{H \times T}$, which are discretized into one-hot vectors using the Gumbel-Softmax relaxation (Jang et al., 2017). The decoder learns an embedding matrix $\mathbf{W} \in \mathbb{R}^{K \times H}$, where K is the dictionary size and H is the dimensionality of each dictionary entry, and reconstructs the input from the embedded discrete representation as $\hat{\mathbf{X}} = d(\mathbf{W}^\top \hat{\mathbf{S}}, \mathbf{C})$. For completeness, Appendix I describes a continuous latent variant and reports its corresponding experimental results.

Both the encoder and decoder are built using non-causal WaveNet residual blocks, as employed in WaveGlow (Prenger et al., 2019), Glow-TTS (Kim et al., 2020), and VITS (Kim et al., 2021). Each WaveNet residual block consists of layers of dilated convolutions, a gated activation unit, and a skip connection. A linear projection layer on top of the residual blocks produces the final sequence of latent embeddings of dimension H for the encoder and reconstructed input representation $\hat{\mathbf{X}}$ for the decoder. The decoder receives both the encoder outputs and an additional conditioning tensor. This tensor is first passed through a single convolutional layer and then added before the gated tanh nonlinearities in each residual block (Kim et al., 2020). It is formed by concatenating the provided features along the channel dimension, while time-invariant embeddings (e.g., speaker identity) are broadcast along the time axis.

5.2 Dataset Preparation

Building on the success of prior works (Polyak et al., 2021; Baas et al., 2023; van Niekerk et al., 2022), we construct our dataset $\mathbb{X}$ from embeddings extracted with a self-supervised learned (SSL) model. Specifically, we employ the WavLM model (Chen et al., 2022) to convert waveforms into embeddings taken from its sixth layer, which is known to preserve both semantic and prosodic information (Wang et al., 2025b; Martín-Cortinas et al., 2024; Shan et al., 2024; Baas et al., 2023).

The observable variable $\mathbf{C}$ used in this voice-conversion instantiation is an embedding vector for the speaker condition. To extract speaker embeddings, we use a pre-trained speaker encoder, RedimNet (Yakovlev et al., 2024). In addition we experiment with pitch condition which is extracted from Swift-F0 (Nieradzik, 2025) as a sequence of continuous pitch values predicted for each time step.

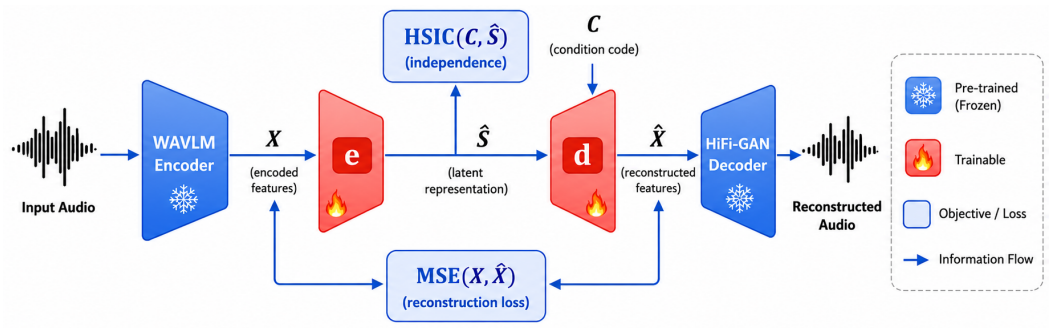


Figure 1: The proposed implementation of the **IVC** framework for speech attribute conversion. Speech features $\mathbf{X}$ are first extracted using a frozen pretrained SSL encoder. The encoder–decoder model (e, d) is then trained to learn a latent representation $\hat{\mathbf{S}}$ that is approximately independent of the observed variable $\mathbf{C}$, while reconstructing the input features, $\mathbf{X} \approx \hat{\mathbf{X}}$. Finally, the converted speech waveform is generated using a pretrained acoustic vocoder.

Finally, to generate waveforms from the converted features, we use a pre-trained acoustic vocoder based on the HiFi-GAN model (Kong et al., 2020), pre-trained by Baas et al. (2023).

5.3 Training Objectives

In this work, we propose a simple and effective objective for encouraging independence in learned representations, inspired by recent advances in discrete speech representation learning (Hsu et al., 2021; Polyak et al., 2021).

We measure statistical dependence between the latent representation $\hat{\mathbf{S}}$ and the conditioning variable $\mathbf{C}$ using the empirical Hilbert–Schmidt Independence Criterion (HSIC). Given a batch of latent sequences $e(\mathbf{x}) = \hat{\mathbf{s}} \in \mathbb{R}^{B\times D\times T}$ and corresponding conditions $\mathbf{c} \in \mathbb{R}^{B\times D_c\times T}$, we first reshape them into a set of $n = B \cdot T$ paired samples: $\{(\hat{\mathbf{s}}_i, \mathbf{c}_i)\}_{i=1}^{n}$, $\hat{\mathbf{s}}_i \in \mathbb{R}^{D}$, $\mathbf{c}_i \in \mathbb{R}^{D_c}$. For time-invariant conditions such as speaker identity, the same condition vector is broadcast across frames before this reshape. We therefore enforce independence at the frame level by applying the HSIC objective between the framewise latent representations and the corresponding condition. This choice is motivated by the local structure of self-supervised speech representations, where individual frames encode substantial phonetic and acoustic information; indeed, framewise representations have been successfully used for content-preserving voice conversion in prior work such as KNN-VC.

To capture nonlinear dependence, we compute a biased centered HSIC estimator using radial basis function (RBF) kernels [1]. Let $\mathbf{K}_{\hat{\mathbf{S}}} = k(\hat{\mathbf{s}}), \mathbf{K}_{\mathbf{C}} = k(\hat{\mathbf{c}}) \in \mathbb{R}^{n\times n}$ be kernel matrices with zero diagonal, then the dependence loss term is given by:

$$\mathcal{L}_{\text{HSIC}}(\hat{\mathbf{s}}, \mathbf{c}) = (n-1)^{-2}\text{tr}(\mathbf{K}_{\hat{\mathbf{S}}}\mathbf{H}\mathbf{K}_{\mathbf{C}}\mathbf{H}), \tag{4}$$

where $\mathbf{H} \in \mathbb{R}^{n\times n}$ is the centering matrix defined by $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ to ensure that the kernels are computed on zero-mean features. Intuitively, HSIC measures the covariance between features in two Reproducing Kernel Hilbert Spaces induced by the kernels on $\hat{\mathbf{S}}$ and $\mathbf{C}$. If the variables are statistically independent, the cross-covariance operator vanishes, and thus the objective in Eq. 4 approaches zero. Finally, the reconstruction is enforced by the mean-squared error loss term:

$$\mathcal{L}_{\text{recon}}(\mathbf{x}, \mathbf{c}) = \|\mathbf{x} - d(e(\mathbf{x}), \mathbf{c})\|_2^2$$

The model is then trained with the following objective:

$$\mathcal{L}_{\mathbf{IVC}}(\mathbf{x}, \mathbf{c}) = \frac{1}{N}\sum_{(\mathbf{x},\mathbf{c})}[\underbrace{\mathcal{L}_{\text{recon}}}_{\mathcal{R}(\cdot)}(\mathbf{x}, \mathbf{c}) + \lambda\underbrace{\mathcal{L}_{\text{HSIC}}(\hat{\mathbf{s}}, \mathbf{c})}_{\mathcal{I}(\cdot)}],$$

[1] We use a multi-scale RBF kernel composed of three kernels with different bandwidths, where each bandwidth is set to the maximum squared distance to the m-th nearest neighbor across all samples $(m = 5, 10, 20)$.

where λ is a hyperparameter. A key advantage of this formulation is its simplicity. Unlike other approaches, such as vCLUB (Cheng et al., 2020) or discriminator-based methods (Shaham et al., 2022), our optimization avoids adversarial training, which is notoriously unstable. Moreover, our method supports one-shot inference, requiring only a single reference speaker example at test time. It also achieves a linear inference-time complexity with respect to the number of input samples, since it bypasses pairwise similarity computations and neighbor-search steps used in methods such as KNN-VC.

6 EXPERIMENTS

6.1 DATASETS

We train our model with the VCTK dataset (110 speakers), LibriTTS (Zen et al., 2019) train-clean-360 (904 speakers), LibriSpeech train-clean-100 (251 speakers) (Panayotov et al., 2015), and ESD English (10 speakers) (Zhou et al., 2021). This dataset yields 1,265 speakers. To evaluate any-to-any voice conversion, we conduct our experiments on the LibriSpeech test-clean dataset (Panayotov et al., 2015), which comprises 40 speakers. To perform the voice conversion, we constructed 7800 content-reference pairs from the test-clean set of LibriSpeech by taking 5 utterances for each speaker and converting them into the remaining 39 speakers, which gives a total of $40 \times 5 \times 39 = 7800$ samples. During inference, the reference speaker is provided using a single 3-second center-cropped utterance. We provide audio samples on GitHub[2].

6.2 EVALUATION

We evaluate our method on the voice conversion task by measuring word error rate (WER) and character error rate (CER) for speech intelligibility. For speaker similarity, we report the target-vs-converted equal error rate (Target EER): higher values indicate that the verification model finds converted samples harder to distinguish from target-speaker samples. This is therefore a similarity score derived from an EER protocol, not a conventional same/different speaker verification error where lower is better. For intelligibility, we utilize the Wav2Vec2-Base model (Baevski et al., 2020), and for speaker similarity, we employ the speaker verification system developed by Snyder et al. (2018) and implemented by Ravanelli et al. (2021). For assessing the naturalness of the converted speech, we adopt the experimental protocol introduced in the recent GenVC study (Cai et al., 2025) and use the UTMOSv2 model to predict perceptual scores that have been shown to correlate strongly with human mean opinion score (MOS) judgments (Baba et al., 2024). Objective evaluation results are reported as UTMOSv2, with all values accompanied by 95% confidence intervals (higher is better). Evaluation is conducted on the EMIME dataset, which, consistent with (Cai et al., 2025), serves as a test set containing speakers unseen during training for all evaluated voice conversion models. We treat UTMOSv2 as an automatic proxy for naturalness rather than a substitute for a human listening study.

In addition, we evaluate our method on the pitch transfer task. We use the Pearson correlation coefficient (PCC) between the target and generated pitch trajectories. The trajectories are first aligned using dynamic time warping (DTW), after which PCC is computed on the aligned sequences. Emotion labels are used only to define source–target utterance pairs and are not provided as conditioning variables. Additiona details are provided in Appendix A.

6.3 RESULTS

Table 1 presents the speaker conversion results of the proposed method. All baselines were run by us with the same inference setup as our method. Table 1 shows that our method achieves the strongest Target EER speaker-similarity score under this protocol when a short reference utterance is provided for voice conversion. The proposed method also improves over KNN-VC in intelligibility (WER, CER metrics) and speaker similarity, and it achieves intelligibility results comparable to the MKL-VC (Lobashev et al., 2025) and Free-VC (Li et al., 2023) baselines. The discrete and continuous variants perform similarly on intelligibility, while the discrete variant obtains the strongest Target EER among the compared conversion methods, consistent with the theory's emphasis on

[2] https://ivc2027.github.io/ivc/

Table 1: Comparison of different methods on objective evaluation metrics. The reference speaker is given by a single **3-second** utterance.

Method	WER (%) ↓	CER (%) ↓	EER (%) ↑
Target	3.39	0.84	47.83
KNN-VC (Baas et al., 2023)	43.84	21.38	12.12
FreeVC (Li et al., 2023)	4.98	1.51	9.86
MKL-VC (Lobashev et al., 2025)	**3.76**	**1.02**	9.12
GENVC-Small (Cai et al., 2025)	10.20	3.81	12.97
MeanVC (Ma et al., 2026)	29.26	13.25	6.14
IVC	5.29	1.59	**23.00**

Table 2: Objective naturalness evaluation using UTMOSv2 (higher is better) on the EMIME dataset. All results are reported with a 95% confidence interval.

Model	UTMOSv2 ↑
YourTTS (Casanova et al., 2022)	2.43 ± 0.018
FreeVC (Li et al., 2023)	2.81 ± 0.014
NEURAL VC (Cao et al., 2024)	2.60 ± 0.015
GENVC-SMALL (Cai et al., 2025)	2.66 ± 0.021
GENVC-LARGE (Cai et al., 2025)	2.65 ± 0.019
IVC	**3.14** ± 0.015

discrete latent content. Table 2 reports the UTMOSv2 results for our method alongside the models evaluated under the same experimental conditions in (Cai et al., 2025). These automatic metrics suggest that the proposed design is a useful and reproducible baseline, while human listening tests remain necessary for a complete perceptual evaluation. Finally, our method achieves the second-best performance on the pitch conversion task, as shown in Table 3, while remaining substantially more parameter-efficient. In particular, GENVC uses more than 500M parameters, whereas our full pipeline contains approximately 122M parameters.

Table 3: Average pitch correlation (PCC) across all source–target emotion directions.

	MKL-VC	**FreeVC**	**KNN-VC**	**MeanVC**	**GENVC**	**IVC**
PPC across 20 directions	0.668	0.729	0.726	0.615	**0.754**	<u>0.736</u>

7 Conclusion

We introduced a principled framework for speech attribute conversion with formal guarantees. Our analysis shows that recovering the latent variable $\mathbf{S}$ up to an invertible transformation is sufficient for correct synthesis and attribute manipulation, and we further characterize the effect of imperfect training through explicit conversion-error bounds.

Our method is simple and reproducible: it relies only on reconstruction and independence losses, avoids adversarial training, and uses a lightweight architecture for efficient one-shot any-to-any voice conversion. Despite its simplicity, the proposed approach achieves competitive intelligibility and strong performance under the EER-based speaker-similarity protocol, and outperforms all baselines in the pitch conversion task. The guarantees are intentionally stated under explicit idealized assumptions, and the experiments should be read as evidence that the resulting design principle is useful rather than as a claim that the assumptions hold exactly in real speech data. Important remaining limitations include the absence of a large-scale human perceptual study and the reliance on automatic proxies such as UTMOSv2. Overall, this work bridges theory and practice in controllable speech generation, establishing IVC as a reliable baseline and a foundation for future theory-grounded speech conversion research.

AI USE STATEMENT

In this work, we used generative AI tools only to improve the writing style, clarity, and grammar of the manuscript. We did not use generative AI tools to formulate scientific claims, derive theoretical results, construct proofs, analyze experimental results, or draw conclusions. All scientific claims, mathematical derivations, proofs, experimental design, analyses, and conclusions are the authors' own work.

We have reviewed all AI-assisted edits and verified that they preserve the intended technical meaning and do not introduce unsupported statements or modifications to the scientific content. We take full responsibility for the final content of this work, including any text revised with the aid of generative AI.

REFERENCES

Matthew Baas, Benjamin van Niekerk, and Herman Kamper. Voice conversion with just nearest neighbors. *arXiv preprint arXiv:2305.18975*, 2023.

Kaito Baba, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. The t05 system for the voicemos challenge 2024: Transfer learning from deep image classifier to naturalness mos prediction of high-quality synthetic speech. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 818–824, 2024. doi: 10.1109/SLT61566.2024.10832315.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12449–12460. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf.

Zexin Cai, Henry Li Xinyuan, Ashi Garg, Leibny Paola García-Perera, Kevin Duh, Sanjeev Khudanpur, Matthew Wiesner, and Nicholas Andrews. Genvc: Self-supervised zero-shot voice conversion. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8, 2025. doi: 10.1109/ASRU65441.2025.11434735.

Danyang Cao, Zeyi Zhang, and Jinyuan Zhang. Neuralvc: Any-to-any voice conversion using neural networks decoder for real-time voice conversion. *IEEE Signal Processing Letters*, 31:2070–2074, 2024. doi: 10.1109/LSP.2024.3439469.

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International conference on machine learning*, pp. 2709–2720. PMLR, 2022.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6): 1505–1518, 2022.

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pp. 1779–1788. PMLR, 2020.

Yang Gao, Rita Singh, and Bhiksha Raj. Voice impersonation using generative adversarial networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2506–2510, 2018. doi: 10.1109/ICASSP.2018.8462018.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pp. 63–77. Springer, 2005.

Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. In J. Platt, D. Koller, Y. Singer, and S. Roweis (eds.), *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/file/d5cfead94f5350c12c322b5b664544c1-Paper.pdf.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. doi: 10.1109/TASLP.2021.3122291.

Xiang Huang, Hao Peng, Li Sun, Hui Lin, Chunyang Liu, Jiang Cao, and Philip S Yu. Structural entropy guided probabilistic coding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17467–17475, 2025.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd international conference on artificial intelligence and statistics*, pp. 859–868. Pmlr, 2019.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=rkE3y85ee.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/khemakhem20a.html.

Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 8067–8077. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/5c3b99e8f92532e5ad1556e53ceea00c-Paper.pdf.

Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5530–5540. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/kim21f.html.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033, 2020.

Jingyi Li, Weiping Tu, and Li Xiao. Freevc: Towards high-quality text-free one-shot voice conversion. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Yisi Liu, Chenyang Wang, Hanjo Kim, Raniya Khan, and Gopala Anumanchipalli. RT-VC: Real-time zero-shot voice conversion with speech articulatory coding. In Pushkar Mishra, Smaranda Muresan, and Tao Yu (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 385–393, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-253-4. doi: 10.18653/v1/2025.acl-demo.37. URL https://aclanthology.org/2025.acl-demo.37/.

Alexander Lobashev, Assel Yermekova, and Maria Larchenko. Training-Free Voice Conversion with Factorized Optimal Transport. In *Interspeech 2025*, pp. 1373–1377, 2025. doi: 10.21437/Interspeech.2025-2043.

Guobin Ma, Jixun Yao, Ziqian Ning, Yuepeng Jiang, Lingxin Xiong, Lei Xie, and Pengcheng Zhu. Meanvc: Lightweight and streaming zero-shot voice conversion via mean flows. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 17347–17351, 2026. doi: 10.1109/ICASSP55912.2026.11462227.

Wan-Duo Kurt Ma, JP Lewis, and W Bastiaan Kleijn. The hsic bottleneck: Deep learning without back-propagation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 5085–5092, 2020.

Álvaro Martín-Cortinas, Daniel Sáez-Trigueros, Grzegorz Beringer, Iván Vallés-Pérez, Roberto Barra-Chicote, Biel Tura-Vecino, Adam Gabryś, Thomas Merritt, Piotr Biliński, and Jaime Lorenzo-Trueba. Investigating self-supervised features for expressive, multilingual voice conversion. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pp. 341–345, 2024. doi: 10.1109/ICASSPW62465.2024.10627128.

Lars Nieradzik. Swiftf0: Fast and accurate monophonic pitch detection, 2025. URL https://arxiv.org/abs/2508.18440.

Yu Pan, Yanni Hu, Yuguang Yang, Jixun Yao, Jianhao Ye, Hongbin Zhou, Lei Ma, and Jianjun Zhao. ClapFM-EVC: High-Fidelity and Flexible Emotional Voice Conversion with Dual Control from Natural Language and Speech. In *Interspeech 2025*, pp. 4583–4587, 2025. doi: 10.21437/Interspeech.2025-203.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

Adam Polyak and Lior Wolf. Attention-based wavenet autoencoder for universal voice conversion. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6800–6804, 2019. doi: 10.1109/ICASSP.2019.8682589.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. Speech Resynthesis from Discrete Disentangled Self-Supervised Representations. In *Interspeech 2021*, pp. 3615–3619, 2021. doi: 10.21437/Interspeech.2021-475.

Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3617–3621, 2019. doi: 10.1109/ICASSP.2019.8683143.

Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. Autovc: Zero-shot voice style transfer with only autoencoder loss. In *International Conference on Machine Learning*, pp. 5210–5219. PMLR, 2019.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, et al. Speechbrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*, 2021.

Uri Shaham, Jonathan Svirsky, Ori Katz, and Ronen Talmon. Discovery of single independent latent variable. *Advances in Neural Information Processing Systems*, 35:25251–25263, 2022.

Siyuan Shan, Yang Li, Amartya Banerjee, and Junier B Oliva. Phoneme hallucinator: One-shot voice conversion via set expansion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14910–14918, 2024.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333, 2018. doi: 10.1109/ICASSP.2018.8461375.

Benjamin van Niekerk, Marc-André Carbonneau, Julian Zaïdi, Matthew Baas, Hugo Seuté, and Herman Kamper. A comparison of discrete and soft speech units for improved voice conversion. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6562–6566, 2022. doi: 10.1109/ICASSP43922.2022.9746484.

Kaidi Wang, Wenhao Guan, Ziyue Jiang, Hukai Huang, Peijie Chen, Weijie Wu, Qingyang Hong, and Lin Li. Discl-VC: Disentangled Discrete Tokens and In-Context Learning for Controllable Zero-Shot Voice Conversion. In *Interspeech 2025*, pp. 1383–1387, 2025a. doi: 10.21437/Interspeech.2025-2684.

Linqin Wang, Zhengtao Yu, Shengxiang Gao, Cunli Mao, Ling Dong, and Yuxin Huang. Voice conversion via structural entropy. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025b. doi: 10.1109/ICASSP49660.2025.10889652.

Ivan Yakovlev, Rostislav Makarov, Andrei Balykin, Pavel Malov, Anton Okhotnikov, and Nikita Torgashov. Reshape Dimensions Network for Speaker Recognition. In *Interspeech 2024*, pp. 3235–3239, 2024. doi: 10.21437/Interspeech.2024-2116.

Yuguang Yang, Yu Pan, Jingjing Yin, Jiangyu Han, Lei Ma, and Heng Lu. Hybridformer: Improving squeezeformer with hybrid attention and nsr mechanism. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10096467.

Zhenchuan Yang, Weibin Zhang, Yufei Liu, and Xiaofen Xing. Cross-Lingual Voice Conversion with Disentangled Universal Linguistic Representations. In *Interspeech 2021*, pp. 1604–1608, 2021. doi: 10.21437/Interspeech.2021-552.

Jixun Yao, Yuguang Yang, Yu Pan, Ziqian Ning, Jianhao Ye, Hongbin Zhou, and Lei Xie. Stablevc: Style controllable zero-shot voice conversion with conditional flow matching. In *AAAI*, pp. 25669–25677, 2025. URL https://doi.org/10.1609/aaai.v39i24.34758.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. In *Interspeech 2019*, pp. 1526–1530, 2019. doi: 10.21437/Interspeech.2019-2441.

Xueyao Zhang, Xiaohui Zhang, Kainan Peng, Zhenyu Tang, Vimal Manohar, Yingru Liu, Jeff Hwang, Dangna Li, Yuhao Wang, Julian Chan, Yuan Huang, Zhizheng Wu, and Mingbo Ma. Vevo: Controllable zero-shot voice imitation with self-supervised disentanglement. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 58436–58459, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/9328208f88ec69420031647e6ff97727-Paper-Conference.pdf.

Yu Zhang, Ron J. Weiss, Heiga Zen, Yonghui Wu, Zhifeng Chen, R.J. Skerry-Ryan, Ye Jia, Andrew Rosenberg, and Bhuvana Ramabhadran. Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning. In *Interspeech 2019*, pp. 2080–2084, 2019. doi: 10.21437/Interspeech.2019-2668.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 920–924. IEEE, 2021.

Kun Zhou, Berrak Sisman, Rajib Rana, Björn W. Schuller, and Haizhou Li. Emotion intensity and its control for emotional voice conversion. *IEEE Transactions on Affective Computing*, 14(1):31–48, 2023. doi: 10.1109/TAFFC.2022.3175578.

Jialong Zuo, Shengpeng Ji, Minghui Fang, Ziyue Jiang, Xize Cheng, Qian Yang, Wenrui Liu, Guangyan Zhang, Zehai Tu, Yiwen Guo, and Zhou Zhao. Enhancing expressive voice conversion with discrete pitch-conditioned flow matching model. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10889108.

A Pitch Conversion Results

This experiment evaluates pitch-sequence transfer across emotion-labeled utterance pairs, rather than full emotion conversion. The experimental setup is as follows:

- **Condition C:** a time-varying pitch trajectory extracted using Swift-F0
- **Source signal:** an utterance wtih acoustic-content representation
- **Target condition:** the pitch trajectory extracted from an utterance belonging to the target emotion category
- **Output:** speech reconstructed from the source latent representation and the target pitch sequence
- **Evaluation metric:** Pearson correlation coefficient (PCC) between the generated and target pitch trajectories after dynamic-time-warping (DTW) alignment.

The emotion labels are used only to define source–target groups, such as Angry → Happy. Emotion itself is not provided as the conditioning variable, and therefore this experiment does not establish full emotion conversion. We accordingly refer to this setting as *pitch transfer across emotion-labeled pairs*.

Table 4: Pitch correlation for emotion conversion across source-target emotion directions.

Conversion Direction	**MKL-VC**	**FreeVC**	**KNN-VC**	**MeanVC**	**GENVC**	**IVC**
Angry → Happy	0.738	0.671	0.796	0.733	0.793	0.823
Angry → Neutral	0.789	0.899	0.700	0.635	0.790	0.875
Angry → Sad	0.776	0.859	0.755	0.656	0.786	0.821
Angry → Surprise	0.600	0.564	0.702	0.726	0.718	0.730
Happy → Angry	0.710	0.743	0.757	0.582	0.641	0.570
Happy → Neutral	0.763	0.916	0.732	0.676	0.814	0.782
Happy → Sad	0.682	0.830	0.674	0.377	0.773	0.752
Happy → Surprise	0.665	0.645	0.692	0.686	0.637	0.662
Neutral → Angry	0.638	0.724	0.701	0.758	0.781	0.782
Neutral → Happy	0.653	0.695	0.776	0.776	0.769	0.767
Neutral → Sad	0.628	0.831	0.686	0.833	0.803	0.808
Neutral → Surprise	0.556	0.572	0.648	0.689	0.647	0.698
Sad → Angry	0.584	0.659	0.768	0.570	0.695	0.617
Sad → Happy	0.636	0.671	0.823	0.615	0.758	0.782
Sad → Neutral	0.684	0.873	0.722	0.730	0.822	0.686
Sad → Surprise	0.545	0.527	0.620	0.506	0.714	0.611
Surprise → Angry	0.604	0.692	0.744	0.466	0.669	0.793
Surprise → Happy	0.788	0.820	0.822	0.708	0.794	0.799
Surprise → Neutral	0.718	0.755	0.760	0.414	0.844	0.677
Surprise → Sad	0.597	0.638	0.649	0.158	0.824	0.683
Average	0.668	0.729	0.726	0.615	**0.754**	<u>0.736</u>

B Ablation Study

Table 5: Evaluation of different implementations for our method. $\mathbf{IVC}^{\mathcal{R}}$ is trained without HSIC loss term.

Method	WER (%) ↓	CER (%) ↓	EER (%) ↑
IVC	5.18	1.58	23.00
$\mathbf{IVC}^{\text{Cont}}$(continuous $\hat{\mathbf{S}}$)	5.30	1.60	20.50
$\mathbf{IVC}^{\mathcal{R}}$(without HSIC)	5.29	1.59	22.67

C Additional Lemmas

For a fixed condition value $\mathbf{c} \in \mathbb{C}$, we define $\mathbb{X}^c$ as the support of the conditional distribution of $\mathbf{X}$ given $\mathbf{C} = \mathbf{c}$, namely

$$\mathbb{X}^c := \left\{ \mathbf{x} \in \mathbb{X} \ : \ \mathbf{x} = f(\mathbf{s}, \mathbf{c}),\ \mathbf{s} \in \mathrm{supp}(\mathbf{S}) \right\} \ \subseteq \ \mathbb{X}.$$

We further define the encoder map $e^c : \mathbb{X}^c \to \hat{\mathbb{S}}$ by $e^c(\mathbf{x}) := e(\mathbf{x}), \forall \mathbf{x} \in \mathbb{X}^c$, and similarly define $d^c : \hat{\mathbb{S}} \to \hat{\mathbb{X}}$ by $d^c(\hat{\mathbf{s}}) := d(\hat{\mathbf{s}}, \mathbf{c})$

Lemma 2. *Under Assumption 1, for every admissible condition* $\mathbf{c}$*, the restricted encoder map* $e^c : \mathbb{X}^c \to \hat{\mathbb{S}}$ *is injective.*

Proof. Fix a condition value $\mathbf{c}$ and take any $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{X}^c$. Suppose that $e^c(\mathbf{x}_1) = e^c(\mathbf{x}_2) = \hat{\mathbf{s}}$. By perfect reconstruction in Assumption 1,

$$\mathbf{x}_1 = d(\hat{\mathbf{s}}, \mathbf{c}) = \mathbf{x}_2.$$

Therefore equality of encoder outputs implies equality of inputs, which is exactly injectivity of e^c on $\mathbb{X}^c$. □

D Proof of Lemma 1

Proof. For every admissible condition $\mathbf{c}$, define

$$T^c : \mathbb{S} \to \hat{\mathbb{S}}, \qquad T^c(\mathbf{s}) := e(f(\mathbf{s}, \mathbf{c})).$$

Because $f(\cdot, \mathbf{c})$ is injective in $\mathbf{s}$ by the assumed invertibility of the generative map and e^c is injective by Lemma 2, their composition $T^c = e^c \circ f(\cdot, \mathbf{c})$ is injective.

We next show that T^c cannot depend on $\mathbf{c}$. Fix two admissible condition values $\mathbf{c}_1, \mathbf{c}_2$ and suppose, for contradiction, that there is an index i such that

$$T^{c_1}(\mathbf{s}_i) \neq T^{c_2}(\mathbf{s}_i).$$

Set $\hat{\mathbf{s}}_i := T^{c_1}(\mathbf{s}_i)$. Since T^{c_1} is injective and $P_{\mathbf{S}}(\mathbf{s}_i) = p_i > 0$, the conditional distribution of $\hat{\mathbf{S}}$ given $\mathbf{C} = \mathbf{c}_1$ assigns mass

$$P_{\hat{\mathbf{S}}|\mathbf{C}=\mathbf{c}_1}(\hat{\mathbf{s}}_i) = p_i.$$

Now consider the same latent point $\hat{\mathbf{s}}_i$ under condition $\mathbf{c}_2$. There are two exhaustive cases.

Case 1: there is no j such that $T^{c_2}(\mathbf{s}_j) = \hat{\mathbf{s}}_i$. Then

$$P_{\hat{\mathbf{S}}|\mathbf{C}=\mathbf{c}_2}(\hat{\mathbf{s}}_i) = 0,$$

which differs from $p_i > 0$.

Case 2: there exists j such that $T^{c_2}(\mathbf{s}_j) = \hat{\mathbf{s}}_i$. Injectivity of T^{c_2} makes this j unique. Moreover $j \neq i$, because by assumption $T^{c_2}(\mathbf{s}_i) \neq \hat{\mathbf{s}}_i$. Hence

$$P_{\hat{\mathbf{S}}|\mathbf{C}=\mathbf{c}_2}(\hat{\mathbf{s}}_i) = p_j.$$

Assumption 3 gives $p_j \neq p_i$.

Thus in either case the conditional law of $\hat{\mathbf{S}}$ differs between $\mathbf{c}_1$ and $\mathbf{c}_2$. This contradicts the zero-dependence part of Assumption 1, which implies $\hat{\mathbf{S}} \perp\!\!\!\perp \mathbf{C}$ and therefore

$$P_{\hat{\mathbf{S}}|\mathbf{C}=\mathbf{c}_1} = P_{\hat{\mathbf{S}}|\mathbf{C}=\mathbf{c}_2}$$

for admissible condition values (equivalently, for $P_{\mathbf{C}}$-almost every pair in the general measure-theoretic formulation).

Consequently $T^{c_1}(\mathbf{s}_i) = T^{c_2}(\mathbf{s}_i)$ for every i and every pair of admissible conditions. Hence all condition-specific maps coincide with a single map T, and

$$e(f(\mathbf{s}, \mathbf{c})) = T(\mathbf{s}) \qquad \forall\, \mathbf{s} \in \mathbb{S},\ \mathbf{c} \in \mathrm{supp}(\mathbf{C}).$$

Since each T^c is injective, T is injective. Because $\mathbb{S}$ is finite, T is a bijection from $\mathbb{S}$ onto its image $T(\mathbb{S})$ and therefore has a well-defined inverse on that image. Finally, $\hat{\mathbf{S}} = e(\mathbf{X}) = e(f(\mathbf{S}, \mathbf{C})) = T(\mathbf{S})$ almost surely, completing the proof. □

E Proof of Proposition 1

Proof. Take any $\hat{\mathbf{s}} \in T(\mathbb{S})$. Since T is bijective from $\mathbb{S}$ onto $T(\mathbb{S})$ by Lemma 1, there exists a unique $\mathbf{s} = T^{-1}(\hat{\mathbf{s}})$. For any admissible condition $\mathbf{c}$, Lemma 1 gives

$$e(f(\mathbf{s}, \mathbf{c})) = T(\mathbf{s}) = \hat{\mathbf{s}}.$$

Applying perfect reconstruction from Assumption 1 to the sample $\mathbf{x} = f(\mathbf{s}, \mathbf{c})$ yields

$$d(e(f(\mathbf{s}, \mathbf{c})), \mathbf{c}) = f(\mathbf{s}, \mathbf{c}).$$

Substituting $e(f(\mathbf{s}, \mathbf{c})) = \hat{\mathbf{s}}$ and $\mathbf{s} = T^{-1}(\hat{\mathbf{s}})$ gives

$$d(\hat{\mathbf{s}}, \mathbf{c}) = f(\mathbf{s}, \mathbf{c}) = f(T^{-1}(\hat{\mathbf{s}}), \mathbf{c}).$$

Since $\hat{\mathbf{s}} \in T(\mathbb{S})$ and $\mathbf{c}$ were arbitrary, the equality holds on the entire learned latent support for every admissible condition. □

F Proof of Corollary 1

Proof. By applying Proposition 1: $d(e(\mathbf{x}), \mathbf{c}') = d(\hat{\mathbf{s}}, \mathbf{c}') = f(T^{-1}(\hat{\mathbf{s}}), \mathbf{c}') = f(\mathbf{s}, \mathbf{c}') = \mathbf{x}'$ □

G Proof of Proposition 2

Proof. Fix any two condition values $\mathbf{c}, \mathbf{c}' \in \mathrm{supp}(\mathbf{C})$. Recall that

$$T_{\mathbf{c}}(\mathbf{s}) := e(f(\mathbf{s}, \mathbf{c})), \qquad Q_{\mathbf{c}} := (T_{\mathbf{c}})_{\#} P_{\mathbf{S}} = P_{\hat{\mathbf{S}}|\mathbf{C}=\mathbf{c}}.$$

We prove the result in four steps.

Step 1: imperfect reconstruction preserves injectivity. We first show that $T_{\mathbf{c}}$ is injective for every $\mathbf{c} \in \mathrm{supp}(\mathbf{C})$. Suppose, for contradiction, that there exist two distinct latent states $\mathbf{s}_i \neq \mathbf{s}_j$ such that

$$T_{\mathbf{c}}(\mathbf{s}_i) = T_{\mathbf{c}}(\mathbf{s}_j) = \hat{\mathbf{s}}.$$

By definition of $T_{\mathbf{c}}$,

$$e(f(\mathbf{s}_i, \mathbf{c})) = e(f(\mathbf{s}_j, \mathbf{c})) = \hat{\mathbf{s}}.$$

Assumption 4 therefore gives $\|d(\hat{\mathbf{s}}, \mathbf{c}) - f(\mathbf{s}_i, \mathbf{c})\|_2 \leq \sqrt{\epsilon}$ and $\|d(\hat{\mathbf{s}}, \mathbf{c}) - f(\mathbf{s}_j, \mathbf{c})\|_2 \leq \sqrt{\epsilon}$. Using the triangle inequality,

$$\begin{aligned} \|f(\mathbf{s}_i, \mathbf{c}) - f(\mathbf{s}_j, \mathbf{c})\|_2 &\leq \|f(\mathbf{s}_i, \mathbf{c}) - d(\hat{\mathbf{s}}, \mathbf{c})\|_2 + \|d(\hat{\mathbf{s}}, \mathbf{c}) - f(\mathbf{s}_j, \mathbf{c})\|_2 \\ &\leq 2\sqrt{\epsilon}. \end{aligned}$$

However, by the definition of γ_X,

$$\|f(\mathbf{s}_i, \mathbf{c}) - f(\mathbf{s}_j, \mathbf{c})\|_2 \geq \gamma_X.$$

Assumption 6 states that $\epsilon < \gamma_X^2/4$, and hence

$$2\sqrt{\epsilon} < \gamma_X,$$

which gives a contradiction. Therefore, $T_{\mathbf{c}}$ is injective for every admissible condition value $\mathbf{c}$.

Step 2: total variation controls condition-dependent latent disagreement. Write

$$p_i := P_{\mathbf{S}}(\mathbf{s}_i), \qquad i = 1, \ldots, K,$$

and recall from Assumption 3 that

$$\Delta_p := \min_{i \neq j} |p_i - p_j| > 0, \qquad p_{\max} := \max_i p_i.$$

For the fixed pair $\mathbf{c}, \mathbf{c}'$, define

$$\hat{\mathbf{s}}_i := T_{\mathbf{c}}(\mathbf{s}_i), \qquad \hat{\mathbf{s}}'_i := T_{\mathbf{c}'}(\mathbf{s}_i),$$

and let

$$M_{\mathbf{c},\mathbf{c}'} := \{i : T_{\mathbf{c}}(\mathbf{s}_i) \neq T_{\mathbf{c}'}(\mathbf{s}_i)\}$$

be the set of latent states whose learned representation changes between the two condition values.

Since both $T_{\mathbf{c}}$ and $T_{\mathbf{c}'}$ are injective, each point $\hat{\mathbf{s}}_i = T_{\mathbf{c}}(\mathbf{s}_i)$ has probability

$$Q_{\mathbf{c}}(\hat{\mathbf{s}}_i) = p_i.$$

Under $Q_{\mathbf{c}'}$, there are two possibilities.

First, suppose that $\hat{\mathbf{s}}_i$ does not belong to the image of $T_{\mathbf{c}'}$. Then $Q_{\mathbf{c}'}(\hat{\mathbf{s}}_i) = 0$, and therefore

$$|Q_{\mathbf{c}}(\hat{\mathbf{s}}_i) - Q_{\mathbf{c}'}(\hat{\mathbf{s}}_i)| = p_i.$$

Second, suppose that $\hat{\mathbf{s}}_i$ also belongs to the image of $T_{\mathbf{c}'}$. Since $T_{\mathbf{c}'}$ is injective, there exists a unique index j such that $T_{\mathbf{c}'}(\mathbf{s}_j) = \hat{\mathbf{s}}_i$. Thus, $Q_{\mathbf{c}'}(\hat{\mathbf{s}}_i) = p_j$. If $i \in M_{\mathbf{c},\mathbf{c}'}$, then $j \neq i$, because $j = i$ would imply $T_{\mathbf{c}}(\mathbf{s}_i) = T_{\mathbf{c}'}(\mathbf{s}_i)$, contrary to the definition of $M_{\mathbf{c},\mathbf{c}'}$. By Assumption 3, the prior probabilities of any two distinct latent states are separated by at least

$$\Delta_p = \min_{k \neq \ell} |p_k - p_\ell|.$$

Since $j \neq i$, it follows that

$$|p_i - p_j| \geq \Delta_p.$$

Finally, using

$$Q_{\mathbf{c}}(\hat{\mathbf{s}}_i) = p_i, \qquad Q_{\mathbf{c}'}(\hat{\mathbf{s}}_i) = p_j,$$

we obtain

$$|Q_{\mathbf{c}}(\hat{\mathbf{s}}_i) - Q_{\mathbf{c}'}(\hat{\mathbf{s}}_i)| = |p_i - p_j| \geq \Delta_p.$$

Since $|Q_{\mathbf{c}}(\hat{\mathbf{s}}_i) - Q_{\mathbf{c}'}(\hat{\mathbf{s}}_i)| \geq \Delta_p$, we have $\frac{|Q_{\mathbf{c}}(\hat{\mathbf{s}}_i) - Q_{\mathbf{c}'}(\hat{\mathbf{s}}_i)|}{\Delta_p} \geq 1$. Multiplying by $p_{\max}$ gives

$$p_{\max} \leq \frac{p_{\max}}{\Delta_p} |Q_{\mathbf{c}}(\hat{\mathbf{s}}_i) - Q_{\mathbf{c}'}(\hat{\mathbf{s}}_i)|.$$

Finally, since $p_i \leq p_{\max}$, for every $i \in M_{\mathbf{c},\mathbf{c}'}$,

$$p_i \leq \frac{p_{\max}}{\Delta_p} |Q_{\mathbf{c}}(\hat{\mathbf{s}}_i) - Q_{\mathbf{c}'}(\hat{\mathbf{s}}_i)|.$$

Because $T_{\mathbf{c}}$ is injective, the points $\{\hat{\mathbf{s}}_i\}_{i=1}^K$ are distinct. Summing over $i \in M_{\mathbf{c},\mathbf{c}'}$ therefore gives

$$\begin{aligned}
P_{\mathbf{S}}\left[T_{\mathbf{c}}(\mathbf{S}) \neq T_{\mathbf{c}'}(\mathbf{S})\right] &= \sum_{i \in M_{\mathbf{c},\mathbf{c}'}} p_i \\
&\leq \frac{p_{\max}}{\Delta_p} \sum_{i \in M_{\mathbf{c},\mathbf{c}'}} |Q_{\mathbf{c}}(\hat{\mathbf{s}}_i) - Q_{\mathbf{c}'}(\hat{\mathbf{s}}_i)| \\
&\leq \frac{p_{\max}}{\Delta_p} \sum_{\hat{\mathbf{s}}} |Q_{\mathbf{c}}(\hat{\mathbf{s}}) - Q_{\mathbf{c}'}(\hat{\mathbf{s}})| \\
&= \frac{2p_{\max}}{\Delta_p} \operatorname{TV}(Q_{\mathbf{c}}, Q_{\mathbf{c}'}),
\end{aligned}$$

where $\operatorname{TV}(\cdot,\cdot)$ denotes the total variation distance between two probability distributions, defined for discrete distributions P and Q as

$$\operatorname{TV}(P,Q) := \frac{1}{2} \sum_z |P(z) - Q(z)|.$$

Hence,

$$P_{\mathbf{S}}\left[T_{\mathbf{c}}(\mathbf{S}) \neq T_{\mathbf{c}'}(\mathbf{S})\right] \leq \frac{2p_{\max}}{\Delta_p} \operatorname{TV}(Q_{\mathbf{c}}, Q_{\mathbf{c}'}). \tag{5}$$

By the definition of $D_{\hat{S}}$,

$$\|T_{\mathbf{c}}(\mathbf{S}) - T_{\mathbf{c}'}(\mathbf{S})\|_2^2 \leq D_{\hat{S}}^2 \mathbf{1}\left\{T_{\mathbf{c}}(\mathbf{S}) \neq T_{\mathbf{c}'}(\mathbf{S})\right\}.$$

Taking expectation over $\mathbf{S}$ and applying equation 5 yields

$$\begin{aligned}
\mathbb{E}_{\mathbf{S}}\left[\|T_{\mathbf{c}}(\mathbf{S}) - T_{\mathbf{c}'}(\mathbf{S})\|_2^2\right] &\leq D_{\hat{S}}^2 P_{\mathbf{S}}\left[T_{\mathbf{c}}(\mathbf{S}) \neq T_{\mathbf{c}'}(\mathbf{S})\right] \\
&\leq \frac{2D_{\hat{S}}^2 p_{\max}}{\Delta_p} \operatorname{TV}(Q_{\mathbf{c}}, Q_{\mathbf{c}'}).
\end{aligned} \tag{6}$$

Step 3: latent disagreement propagates to conversion error. For a realization $\mathbf{s}$ of $\mathbf{S}$, define the source and corresponding target samples

$$\mathbf{x}_0 = f(\mathbf{s}, \mathbf{c}), \qquad \mathbf{x}' = f(\mathbf{s}, \mathbf{c}'),$$

and the converted output

$$\hat{\mathbf{x}}' := d(e(\mathbf{x}_0), \mathbf{c}').$$

The conversion error is therefore

$$\epsilon_{\text{conv}}(\mathbf{s}, \mathbf{c}, \mathbf{c}') = \|\mathbf{x}' - d(e(\mathbf{x}_0), \mathbf{c}')\|_2^2.$$

By Assumption 4, applied to the target sample $\mathbf{x}' = f(\mathbf{s}, \mathbf{c}')$,

$$\|d(e(\mathbf{x}'), \mathbf{c}') - \mathbf{x}'\|_2^2 \le \epsilon.$$

Define the reconstruction residual

$$\mathbf{r} := \mathbf{x}' - d(e(\mathbf{x}'), \mathbf{c}'), \qquad \|\mathbf{r}\|_2^2 \le \epsilon.$$

Then

$$\mathbf{x}' = d(e(\mathbf{x}'), \mathbf{c}') + \mathbf{r},$$

and consequently

$$\begin{aligned}\epsilon_{\text{conv}}(\mathbf{s}, \mathbf{c}, \mathbf{c}') &= \|d(e(\mathbf{x}'), \mathbf{c}') - d(e(\mathbf{x}_0), \mathbf{c}') + \mathbf{r}\|_2^2 \\ &\le 2\left(\|d(e(\mathbf{x}'), \mathbf{c}') - d(e(\mathbf{x}_0), \mathbf{c}')\|_2^2 + \|\mathbf{r}\|_2^2\right) \\ &\le 2\left(\|d(e(\mathbf{x}'), \mathbf{c}') - d(e(\mathbf{x}_0), \mathbf{c}')\|_2^2 + \epsilon\right).\end{aligned}$$

By Assumption 5,

$$\|d(e(\mathbf{x}'), \mathbf{c}') - d(e(\mathbf{x}_0), \mathbf{c}')\|_2 \le L\|e(\mathbf{x}') - e(\mathbf{x}_0)\|_2.$$

Using

$$e(\mathbf{x}') = T_{\mathbf{c}'}(\mathbf{s}), \qquad e(\mathbf{x}_0) = T_{\mathbf{c}}(\mathbf{s}),$$

we obtain

$$\epsilon_{\text{conv}}(\mathbf{s}, \mathbf{c}, \mathbf{c}') \le 2\left(L^2\|T_{\mathbf{c}'}(\mathbf{s}) - T_{\mathbf{c}}(\mathbf{s})\|_2^2 + \epsilon\right).$$

Taking expectation over $\mathbf{S}$ gives

$$\mathbb{E}_{\mathbf{S}}\left[\epsilon_{\text{conv}}(\mathbf{S}, \mathbf{c}, \mathbf{c}')\right] \le 2\epsilon + 2L^2\mathbb{E}_{\mathbf{S}}\left[\|T_{\mathbf{c}'}(\mathbf{S}) - T_{\mathbf{c}}(\mathbf{S})\|_2^2\right].$$

Applying Eq. equation 6, we obtain the intermediate total-variation bound

$$\mathbb{E}_{\mathbf{S}}\left[\epsilon_{\text{conv}}(\mathbf{S}, \mathbf{c}, \mathbf{c}')\right] \le 2\epsilon + \frac{4L^2 D_{\hat{S}}^2 p_{\max}}{\Delta_p}\,\text{TV}(Q_{\mathbf{c}}, Q_{\mathbf{c}'}). \tag{7}$$

Step 4: approximate independence yields the uniform bound. Let

$$Q := P_{\hat{\mathbf{S}}}.$$

By the triangle inequality for total variation,

$$\text{TV}(Q_{\mathbf{c}}, Q_{\mathbf{c}'}) \le \text{TV}(Q_{\mathbf{c}}, Q) + \text{TV}(Q, Q_{\mathbf{c}'}). \tag{8}$$

Assumption 6 states that

$$\sup_{\tilde{\mathbf{c}} \in \text{supp}(\mathbf{C})} \text{TV}(Q_{\tilde{\mathbf{c}}}, Q) \le \delta.$$

Therefore, for the arbitrary values $\mathbf{c}, \mathbf{c}'$ fixed above,

$$\text{TV}(Q_{\mathbf{c}}, Q) \le \delta, \qquad \text{TV}(Q_{\mathbf{c}'}, Q) \le \delta,$$

and Eq. equation 8 gives

$$\text{TV}(Q_{\mathbf{c}}, Q_{\mathbf{c}'}) \le 2\delta.$$

Substituting this result into Eq. equation 7 yields

$$\mathbb{E}_{\mathbf{S}}\left[\epsilon_{\text{conv}}(\mathbf{S}, \mathbf{c}, \mathbf{c}')\right] \le 2\epsilon + \frac{8L^2 D_{\hat{S}}^2 p_{\max}}{\Delta_p}\delta.$$

Since $\mathbf{c}$ and $\mathbf{c}'$ were arbitrary elements of $\text{supp}(\mathbf{C})$, the bound holds for every admissible conversion pair, completing the proof. □

Table 6: Comparison of the continuous and discrete IVC variants on objective evaluation metrics. The reference speaker is given by a single **3-second** utterance.

Method	WER (%) ↓	CER (%) ↓	EER (%) ↑
Target	3.39	0.84	47.83
IVC (continuous $\hat{\mathbf{S}}$)	5.30	1.60	20.50
IVC (discrete $\hat{\mathbf{S}}$)	5.29	1.59	**23.00**

H Proof of Corollary 2

Proof. Let $\hat{\mathbf{s}}, \hat{\mathbf{s}}' \in \mathbb{R}^K$ be any two categorical probability vectors. Since their entries are nonnegative and sum to one,

$$\|\hat{\mathbf{s}}\|_2^2 = \sum_{k=1}^{K} \hat{s}_k^2 \leq \left(\sum_{k=1}^{K} \hat{s}_k \right)^2 = 1,$$

and similarly,

$$\|\hat{\mathbf{s}}'\|_2^2 \leq 1.$$

Moreover, because both vectors are component-wise nonnegative,

$$\hat{\mathbf{s}}^\top \hat{\mathbf{s}}' \geq 0.$$

Therefore,

$$\begin{aligned} \|\hat{\mathbf{s}} - \hat{\mathbf{s}}'\|_2^2 &= \|\hat{\mathbf{s}}\|_2^2 + \|\hat{\mathbf{s}}'\|_2^2 - 2\hat{\mathbf{s}}^\top \hat{\mathbf{s}}' \\ &\leq 2. \end{aligned}$$

Hence, the latent diameter appearing in Proposition 2 satisfies

$$D_{\hat{S}}^2 \leq 2.$$

Substituting this bound into

$$\mathbb{E}_{\mathbf{S}} \left[\epsilon_{\text{conv}} \right] \leq 2\epsilon + \frac{8L^2 D_{\hat{S}}^2 p_{\max}}{\Delta_p} \delta$$

gives

$$\mathbb{E}_{\mathbf{S}} \left[\epsilon_{\text{conv}} \right] \leq 2\epsilon + \frac{16L^2 p_{\max}}{\Delta_p} \delta,$$

which proves the result. □

I Continuous Latent Variant

For completeness, we also evaluate a continuous variant of the latent representation. In this setting, the encoder outputs

$$e(x) = \hat{\mathbf{S}} \in \mathbb{R}^{T \times D},$$

using a continuous bottleneck at each time step of dimension $D = 1$. This variant follows the same reconstruction and independence training principle as the discrete model. We include it only as an empirical ablation.

Table 6 compares the continuous and discrete variants. The discrete representation achieves slightly better WER, CER, and target-speaker EER.

J Speech Units Analysis

To provide empirical evidence that Assumption 3 holds in real-world speech data, we analyze the variable $\tilde{\mathbf{S}}$ constructed from the K-means labels of samples $\{\mathbf{x}_i\}_{i=1}^N$. We estimate the prior probability

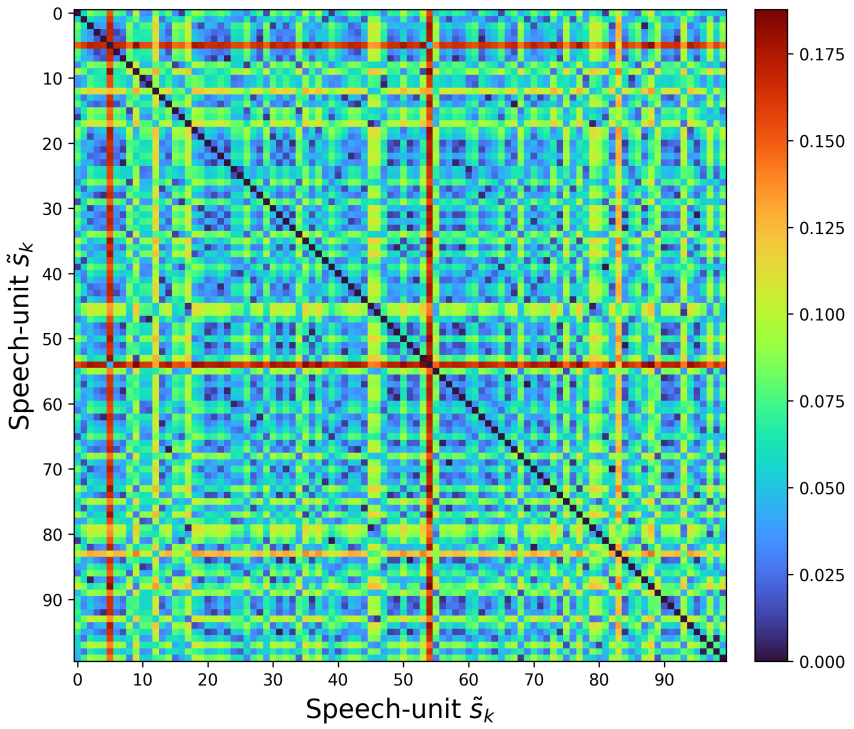


Figure 2: Empirical analysis of the asymmetric-prior condition in Assumption 3. We compute the pairwise distance matrix $\mathbf{D}$ over the empirical prior probabilities p_k of the speech unit categories $\tilde{s}_k$, for $k \in \{1, \ldots, K\}$. Since $\tilde{\mathbf{S}}$ serves as a proxy for the latent content variable $\mathbf{S}$ and $\mathbf{D}_{ii} = 0$ for all i while $\mathbf{D}_{ij} \neq 0$ for all $i \neq j$, the results indicate pairwise-distinct empirical category priors in real-world speech data.

of each category by measuring the empirical frequency of the corresponding label in the training dataset,

$$p_k \;=\; \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\{\tilde{s}_i = k\}, \qquad k \in \{1, \ldots, K\}.$$

From the resulting probability vector $\mathbf{p} = [p_1, p_2, \ldots, p_K]$, we compute the pairwise square-root l_1 distance matrix $\mathbf{D} \;=\; \left|\, \mathbf{p}\mathbf{1}^\top - \mathbf{1}\mathbf{p}^\top \,\right|^{\frac{1}{2}}$, where $\mathbf{1} \in \mathbb{R}^K$ denotes the all-ones column vector.

Figure 2 visualizes the matrix $\mathbf{D}$. All off-diagonal entries are non-zero and clearly separated from the diagonal, indicating that the empirical prior probabilities of the speech unit categories are distinct. Since the unobserved content variable $\mathbf{S}$ is approximated by the proxy $\tilde{\mathbf{S}}$, which captures phonetic structure, this analysis provides empirical support for a positive empirical identifiability margin Δ_p in the speech-unit proxy, while not establishing the assumption for the unobserved true content variable $\mathbf{S}$.

K Prior Work on Independence Objectives

Several prior works have explored different strategies to enforce the independence condition. We review them briefly and compare them to our approach.

Hilbert-Schmidt Independence Criterion Ma et al. (2020) proposed to use an empirical estimate of the Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., 2005) objective, which measures the statistical dependence between two random variables using kernel methods. Hilbert-Schmidt norm of the cross-covariance operator between the distributions in the Reproducing Kernel Hilbert Space (RKHS) defined by

$$(N-1)^{-2}\mathrm{tr}(\mathbf{K}_{\hat{\mathbf{S}}}\mathbf{H}\mathbf{K}_{\mathbf{C}}\mathbf{H}), \tag{9}$$

where $\mathbf{K}_{\hat{\mathbf{S}}}, \mathbf{K}_{\mathbf{C}} \in \mathbb{R}^{N \times N}$ are kernel matrices computed over the sets of samples $\hat{\mathbb{S}} = \{\hat{\mathbf{s}}_1, \ldots, \hat{\mathbf{s}}_N\}$ and $\mathbb{C} = \{\hat{\mathbf{c}}_1, \ldots, \hat{\mathbf{c}}_N\}$ using a positive-definite kernel function (e.g., Gaussian/RBF), $\mathbf{H} \in \mathbb{R}^{N \times N}$ is the centering matrix defined by $\mathbf{H} = \mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T$ to ensure that the kernels are computed on zero-mean features and N is the number of samples. Intuitively, HSIC measures the covariance between features in two RKHSs induced by the kernels on $\hat{\mathbf{S}}$ and $\mathbf{C}$. If the variables are statistically independent, the cross-covariance operator vanishes, and thus the objective in Eq. 9 approaches zero. While this biased HSIC estimator is often preferred for its lower variance, it corresponds to a V-statistic and exhibits a finite-sample bias due to the inclusion of self-similarity (diagonal) terms. In contrast, one might employ the unbiased U-statistic estimator, which removes these terms and provides a consistent estimate of statistical dependence (Gretton et al., 2007).

Contrastive Log-ratio Upper Bound Cheng et al. (2020) proposed a contrastive log-ratio upper bound (vCLUB) of mutual information for a variational autoencoder, which is trained with the following objective function:

$$\min_{\boldsymbol{\theta}} \left[\mathbb{E}_{\mathbb{P}(\hat{\mathbf{S}},\mathbf{C})}[\log q_{\boldsymbol{\theta}}(\mathbf{C}|\hat{\mathbf{S}})] - \mathbb{E}_{\mathbb{P}(\hat{\mathbf{S}})}\mathbb{E}_{\mathbb{P}(\mathbf{C})}[\log q_{\boldsymbol{\theta}}(\mathbf{C}|\hat{\mathbf{S}})] \right],$$

where $q_{\boldsymbol{\theta}}(\mathbf{C}|\hat{\mathbf{S}})$ is a variational distribution parametrized by $\boldsymbol{\theta}$ that approximates $\mathbb{P}(\mathbf{C}|\hat{\mathbf{S}})$. The intuition behind the vCLUB objective is to estimate the mutual information between two variables by contrasting how well a variational model can predict the true paired samples versus mismatched (independent) samples. The first term encourages a high likelihood of true pairs $(\mathbf{S}, \mathbf{C})$. In contrast, the second term penalizes a high likelihood of randomly paired $\mathbf{S}$ and $\mathbf{C}$, sampled independently from their marginals. The difference between these two expectations provides an upper bound on the mutual information, which can be minimized to encourage statistical independence between the variables.

Adversarial Independence Another approach was proposed by Shaham et al. (2022) where the model is trained in an adversarial way, and a discriminator $g(e(\mathbf{x}))$ is trained to predict a condition $\mathbf{c}$ from the latent $\hat{\mathbf{s}} = e(\mathbf{x})$ by maximizing the objective $-\mathcal{I}(g(\hat{\mathbf{s}}), \mathbf{c})$, e.g. minimizing cross entropy loss term $\sum_{i=1}^{K} c_i \log(g(\hat{\mathbf{s}}_\mathbf{i}))$ between condition $\mathbf{c}$ and discriminator prediction $g(\hat{\mathbf{s}})$. The autoencoder is trained to confuse the discriminator by minimizing $-\mathcal{I}(g(e(\mathbf{x})), \mathbf{c})$. Assuming a problem with a single conditional source $\mathbf{C}$, the training objective becomes:

$$\min_{e,d} \max_{g} [\mathcal{R}(\mathbf{x}, d(\hat{\mathbf{s}}, \mathbf{c})) - \lambda \mathcal{I}(g(\hat{\mathbf{s}}), \mathbf{c})].$$

While this approach has been proven to recover the target latent component up to an entropy-preserving transformation, it critically relies on the capacity and stability of the discriminator. In practice, weak or poorly trained discriminators may suffer from mode collapse, where the discriminator focuses only on a subset of easily distinguishable modes in the conditional variable and ignores others. As a result, the encoder may exploit this weakness by only obfuscating the modes to which the discriminator is sensitive, while still leaking conditional information through other dimensions. This undermines the goal of achieving valid conditional invariance and can lead to incomplete or biased disentanglement in the learned representation.

L LIMITATIONS

- **Model Architecture** The proposed method relies on the quality and expressiveness of the pretrained self-supervised learning (SSL) encoder and acoustic decoder. Since our autoencoder operates on the outputs of the SSL encoder and its reconstructions serve as inputs to the acoustic decoder, any limitations or biases in these components can affect the performance and fidelity of the conversion. However, this dependence also becomes a strength in low-resource scenarios: the SSL and acoustic models are foundational models trained on large-scale, diverse datasets, enabling strong generalization even when the trainable part of our method is relatively lightweight. As a result, our approach remains effective and data-efficient in domains with limited labeled or supervised data.
- **Evaluation and Baselines** This work prioritizes the theoretical analysis and convergence guarantees of our proposed framework over achieving state-of-the-art (SOTA) empirical performance. Due to limited research resources and computational constraints, we train and evaluate our method on publicly available datasets with reduced scale. For fair comparison,

we benchmark against baselines trained under the same conditions. While some recent models (e.g., NANSY, SelfVC) report strong empirical results, they were trained on large-scale or proprietary datasets and do not publicly release complete code or training details, making direct comparison infeasible. Our goal is to provide a principled and reproducible foundation that can support future extensions and scaling efforts.

M Broader Impacts

This work presents a model for speech attributes conversion, e.g. voice or emotion, offering benefits such as improved accessibility, expressive speech synthesis, and enhanced human-computer interaction. However, it also poses risks, including potential misuse for impersonation, emotional manipulation, and audio deepfakes. These concerns are particularly relevant to disinformation and privacy violations. Our work is intended for controlled research use, and we emphasize the need for future safeguards, such as watermarking, detection tools, and responsible access policies, to mitigate misuse and uphold ethical standards.

N Additional Results

The cosine distance between the generated and real samples is very low, indicating that the generated samples closely match real speaker characteristics. We demonstrate this by plotting the distribution of cosine distances between real and generated samples in Figure 3.

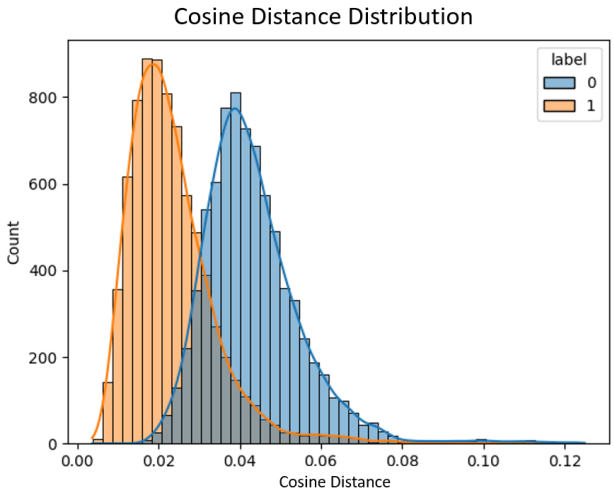


Figure 3: Distributions of cosine distances calculated on the pairs constructed from real samples (orange) and on the pairs constructed such that sample one comes from generated samples and sample two comes from a real sample (blue). The distance is defined by $1 - \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{||\mathbf{x}_i||_2 ||\mathbf{x}_j||_2}$ where $\mathbf{x}_i$, $\mathbf{x}_j$ are speaker embeddings. The results obtained from the model trained with the RedimNet speaker encoder. The plot shows that the generated samples are very close to the real samples and almost indistinguishable.

O Training Details

Table 7 summarizes the hyperparameters used for training. The model is trained on source audio segments of 3 seconds in length, ensuring that all features are extracted from short-duration waveforms. All models were trained using the Adam optimizer to reduce the memory demands of the moments.

Table 7: Training hyperparameters

Hyperparameter	Value
Number of parameters	21.4M
Epochs	100
Batch Size	256
LR (start)	1e-3
LR (end)	1e-5
Scheduler	cosine
Segment Length (sec)	3
Condition Encoder	RedimNet
Optimizer	Adam

Table 8: Objective evaluation of the mel-spectrogram autoencoder variant with Vocos vocoder.

Method	Params (M)	EER (%) ↑	WER (%) ↓	CER (%) ↓
IVC-mel	20.5	**11.05**	29.40	13.77
MeanVC	14.0	6.14	**29.26**	**13.25**

P MEL-SPECTROGRAM IVC VARIANT WITH VOCOS

We also evaluate an acoustic-domain variant of IVC that removes the pretrained WavLM feature extractor. Instead, the observed variable $\mathbf{X}$ is defined as the mel-spectrogram of the input waveform. The autoencoder encodes the mel-spectrogram into $\hat{\mathbf{S}} = e(\mathbf{X})$ and reconstructs it with the decoder conditioned on the target attribute $\mathbf{C}$, using the same conditioning setup as in the main experiments.

Waveform synthesis is performed with Vocos. Training is done in two stages. First, the mel-spectrogram autoencoder is trained with the reconstruction objective and independence penalty between $\hat{\mathbf{S}}$ and $\mathbf{C}$ as in Eq. 1. Second, the autoencoder is frozen and Vocos is fine-tuned on converted mel-spectrograms produced by the frozen model. This adapts the vocoder to the distribution of converted rather than natural mel-spectrograms. Notably, our model contains only 20.5M parameters, making it a compact alternative to SSL-based conversion systems.

This variant serves as an ablation of the representation choice: it tests whether IVC can operate directly on mel-spectrograms without SSL speech features. We compare against MeanVC, a recently proposed lightweight voice-conversion method in Table 8.