Conditional Vocal Timbral Technique Conversion via Embedding-Guided Dual Attribute Modulation

Ting-Chao Hsu R12942155@NTU.EE.TW and **Yi-Hsuan Yang** YHYANGTW@NTU.EE.TW

Editors: D. Herremans, K. Bhandari, A. Roy, S. Colton, M. Barthet

Abstract

Vocal timbral techniques—such as whisper, falsetto, and vocal fry scream—uniquely shape the spectral properties of the human voice, presenting a complex challenge for converting between them while preserving the original speaker's identity. Traditional voice conversion methods, while effective at altering speaker identity or broad timbral qualities, often struggle to transform specialized timbral techniques without compromising speaker-specific traits. Similarly, existing style-transfer models, which are designed to capture broad categories like emotional expressiveness or singing styles, lack the necessary granularity to handle technique-specific variations. To address this, we propose FABYOL, a novel framework for timbral technique conversion built upon FACodec. FABYOL leverages supervised contrastive learning to generate embeddings that encode specific timbral techniques. These embeddings are then used to modulate timbre and prosody, enabling authentic technique conversion while preserving speaker identity. Experimental evaluation, using both tailored objective metrics and a user study, demonstrates that FABYOL achieves promising performance and offers significant improvements in fidelity and flexibility compared to state-of-the-art models. To support this task, we also introduce the EMO dataset, a high-quality, paired corpus developed with a specific focus on vocal fry scream. Audio samples, source code, pre-trained checkpoints, and the EMO dataset are available at https://alberthsu0509.github.io/FABYOL/.

Keywords: Vocal Timbral Technique, Voice Conversion

1. Introduction

Vocal timbre is fundamentally a combination of a speaker's identity and their applied timbral technique. These techniques are pervasive across diverse audio domains, including screaming in heavy metal music, whisper in cinema, and falsetto in voice acting. It is important to distinguish these techniques, which primarily alter the texture of the voice, from pitch-related techniques such as vibrato or trills. Defined by distinct vocal fold vibration patterns and spectral characteristics, they enhance expressiveness and convey specific artistic intent. Developing models that can controllably convert these techniques would unlock significant creative applications in the audio and music industry. For example, creators could instantly re-style a performance into a whisper to enhance narrative intimacy, while artists could generate extreme vocalizations—such as vocal fry screaming—without the need for extensive specialized training.

As this is a nascent research direction, we pioneer the task by strategically focusing on speech. This is because current speech corpora offer more distinct and high-contrast timbral technique variations than available singing datasets. For example, prominent singing datasets like VocalSet (Chou et al., 2018) or GTsinger (Hsu et al., 2024) are often limited to more subtle distinctions, such as "breathy" or "mixed voice". Furthermore, dedicated datasets for melodic screaming are non-existent. This scarcity of appropriate singing data motivates our decision to pioneer this task within the speech domain.

However, this task remains a significant challenge. While specialized methods exist for whisper or Lombard speech (Cotescu et al., 2020; Hu et al., 2021), they are typically treated as distinct tasks. Meanwhile, existing voice conversion models reveal clear limitations for this purpose. Existing voice conversion models reveal clear limitations for this task. Models like CosyVoice (Du et al., 2024) and FreeVC (Li et al., 2023), which are built for cross-speaker shifts, process timbre information to replace speaker identity, thereby discarding the original and neglecting technique control. Similarly, FACodec (Ju et al., 2024), while capable of broader timbre adjustments, lacks precision for techniques like scream or whisper and fails to retain speaker traits due to its generalized timbre handling. Meanwhile, style-transfer methods (Du et al., 2021; Zhou et al., 2021) are designed to disentangle style or emotion and overlook timbral techniques as distinct style elements. These gaps highlight the need for an approach tailored to self-retained timbral technique conversion that preserves speaker identity.

To address this, we propose FABYOL, a novel framework for timbral technique conversion built upon a pre-trained, frozen FACodec (Ju et al., 2024) architecture. FABYOL employs a BYOL-TT encoder, which uses targeted augmentation and contrastive learning to derive robust technique-specific embeddings. These embeddings guide a dual attribute modulation, implemented via adaptive layer normalization (AdaLN) (Peebles and Xie, 2023). This design stems from our key finding: while modulating timbre is an expected component of conversion, we found that prosody modulation is equally critical for achieving authentic results. This lightweight approach allows for targeted technique conversion while maintaining the speaker's core identity.

A further challenge is the evaluation of this task, given the lack of established metrics to quantify technique similarity. Our objective evaluation therefore employs a combination of proxy metrics from vocal analysis to assess technique-specific features and a specific protocol to measure speaker preservation. This is supplemented by a user study measuring perceptual authenticity across technique similarity, speaker similarity, and naturalness. To address the scarcity of paired screaming data and facilitate future research, we also introduce the EMO dataset, which is a one-hour paired dataset from a single speaker containing modal voice and vocal fry scream. The contributions of this paper are as follows:

- We demonstrate that prosody modulation is also essential for effective timbral technique conversion.
- We present FABYOL, the first model to our knowledge that performs timbral technique conversion while preserving the speaker identity.
- We introduce the EMO dataset and the evaluation framework that specifically address the unique challenges of assessing timbral technique conversion quality.

We encourage readers to visit our demo website for audio samples that demonstrate our model's performance.

2. Related Work

With the growing interest in vocal-related research, academic studies on vocal timbral techniques have gained popularity. Previous works have focused on vocal technique detection (Yamamoto et al., 2022; Kalbag and Lerch, 2022), style-controlled voice conversion (Dai et al., 2025; Du et al., 2021; Zhou et al., 2021), and representation learning for various vocal-related attributes (Ju et al., 2024; Elbanna et al., 2022; Yakura et al., 2022). These studies have laid the foundation for our work.

Vocal Technique Detection. Detecting vocal techniques is foundational for vocal manipulation. A previous study (Yamamoto et al., 2022) focused on detecting vocal techniques of J-POP solo singers, demonstrating the ability to distinguish between pitch and timbral techniques in J-POP singing. Moreover, the detection of harsh vocal effects such as screaming has been studied in (Kalbag and Lerch, 2022), which used spectral features to classify extreme timbres in heavy metal music.

Style-Controlled Voice Conversion. Singing voice conversion has progressed toward singing technique control (e.g., falsetto, vibrato) using diffusion models (Dai et al., 2025), yet rarely focuses on timbral techniques like whisper or scream. Speech style models (Du et al., 2021; Zhou et al., 2021) concentrate on emotional style transfer in spoken voice. While adept at disentangling emotional attributes, these approaches overlook timbral techniques as style components.

Representation Learning. Several general-purpose audio representation learning approaches (**?**Saeed et al., 2021; Niizumi et al., 2022) have shown success in various applications. Recently, specific representations for each audio source demonstrate superiority in target domains, including speech and singing voice representation learning (Elbanna et al., 2022; Yakura et al., 2022). Due to increasing requirements for further control of vocals, FACodec (Ju et al., 2024) focuses on attribute disentanglement, including timbre, prosody, and content. Though FACodec (Ju et al., 2024) demonstrates the ability for timbre disentanglement, it falls short on isolating timbral techniques from identity or content, limiting the further detailed control of the timbre attribute.

While prior work has made significant progress, timbral techniques remain largely unexplored across these domains. Our work addresses this gap by integrating *technique embeddings* into FACodec (Ju et al., 2024), enabling speaker-consistent conversion.

3. Timbral Technique Extractor

In our proposed method, FABYOL, we aim to develop a conditional generator, denoted as $\mathcal{G}(\mathbf{x}, \mathbf{h}_{\text{tech}})$, that transforms an input audio signal $\mathbf{x}$ into an output $\mathbf{y}$, where $\mathbf{x}$ and $\mathbf{y}$ are temporally aligned and share the same speaker identity and linguistic content, but exhibit different timbral techniques. The technique embedding $\mathbf{h}_{\text{tech}}$, a learnable representation of timbral techniques, is derived by the embedding extractor E from a reference audio signal $\mathbf{x}_{\text{ref}}$, such that $\mathbf{h}_{\text{tech}} = E(\mathbf{x}_{\text{ref}})$.

To build an effective timbral technique conversion system, we first need a representation of vocal techniques that generalizes across speakers and linguistic content.

3.1. Contrastive Objective of Timbral Technique

BYOL framework. To derive robust timbral technique representations, we adopt Bootstrap Your Own Latent (BYOL) (Grill et al., 2020), a self-supervised learning method that learns from positive pairs without needing negative samples, unlike traditional contrastive approaches such as SimCLR (Chen et al., 2020). BYOL uses two networks: the online network f_θ processes input $\mathbf{x}$, while the target network f_ξ handles an augmented version $\mathbf{u}$. A predictor q_θ aligns the online output to the target, with f_ξ's parameters updated as an exponential moving average of f_θ's parameters, controlled by a decay rate $\tau \in [0, 1]$. The loss minimizes the mean squared error:

$$\mathcal{L}_{\text{BYOL}} = \|q_\theta(f_\theta(\mathbf{x})) - f_\xi(\mathbf{u})\|_2^2.$$

Disentanglement objective. Building on this foundation, we introduce BYOL-TT (BYOL for Timbral Techniques), which leverages BYOL-A's audio encoder (Niizumi et al., 2022) to disentangle vocal signals into three distinct components: timbral technique, speaker identity, and linguistic content. This disentanglement is critical for enabling accurate timbral technique conversion while preserving speaker identity. Techniques like whisper or scream drastically alter vocal spectra and are often entangled with speaker characteristics, making it difficult for models to modify the technique without inadvertently changing who the speaker sounds like.

Therefore, the effectiveness of BYOL-TT relies on constructing meaningful positive pairs—samples that vary in speaker identity and linguistic content but share the same timbral technique, guiding the model to treat the technique as the invariant factor and learn robust, disentangled representations. To achieve this, we apply targeted augmentation (Yakura et al., 2022), crafting transformations that isolate timbral technique while varying other attributes. We propose two pair-generation methods: **DSP Augmentation** & **Real-world Data Selection**.

3.2. DSP Augmentation

Our first augmentation strategy, DSP augmentation, generates synthetic positive pairs from a single audio sample by applying signal processing techniques that selectively modify different attributes of the audio while keeping the timbral technique unchanged.

We apply **Sequence Perturbation (SP)** (Deng et al., 2024) to change the linguistic content of the audio. This is done by splitting the audio into several segments and shuffling their order, which alters the content without affecting the speaker's voice or the technique being used. We denote the resulting audio as:

$$\mathbf{x}_{\text{SP}} = \text{SP}(\mathbf{x}),$$

In parallel, we apply **Vocal Tract Length Perturbation (VTLP)** (Jaitly and Hinton, 2013) to simulate a different speaker identity by warping the spectral characteristics of the voice. This changes how the speaker sounds, while preserving both the original content and the vocal technique. The perturbed version is denoted as:

$$\mathbf{x}_{\text{VTLP}} = \text{VTLP}(\mathbf{x}, \alpha),$$

By pairing $\mathbf{x}_{\mathrm{SP}}$ and $\mathbf{x}_{\mathrm{VTLP}}$, we construct a positive pair where the only shared attribute is the timbral technique. This encourages the model to learn representations that are invariant to speaker and content, and sensitive only to vocal technique. We apply both pre-norm and post-norm to the embeddings and compute the BYOL-style contrastive loss:

$$\mathcal{L}_{\mathrm{DSP}} = \left\| q_\theta(f_\theta(\mathbf{x}_{\mathrm{SP}})) - f_\xi(\mathbf{x}_{\mathrm{VTLP}}) \right\|_2^2 ,$$

3.3. Real-world Data Selection

While DSP Augmentation offers a controllable approach, it relies on synthetic transformations that might not fully capture natural variations. This led us to explore a complementary strategy: Real-world Data Selection.

Real-world data selection leverages our labeled dataset to create positive pairs from distinct audio clips $\mathbf{x}_1$ and $\mathbf{x}_2$ that share the same timbral technique but have different speaker identities and contain different linguistic content. By carefully selecting samples that meet these criteria, we ensure that the model focuses on technique-specific features. The loss for Selection is formulated as:

$$\mathcal{L}_{\mathrm{Sel}} = \| q_\theta(f_\theta(\mathbf{x}_1)) - f_\xi(\mathbf{x}_2) \|_2^2 .$$

This approach surpasses DSP Augmentation by enhancing diversity through varied samples as tested in our experiments, exposing the model to broader speaker and content ranges, potentially boosting the generalization of the technique embedding $\mathbf{h}_{\mathrm{tech}}$. It leverages the dataset's natural variability instead of synthetic changes, possibly yielding truer representations.

In our setup, the trained BYOL-TT encoder $E = f_\theta$ generates $\mathbf{h}_{\mathrm{tech}} = E(\mathbf{x}_{\mathrm{ref}})$ from a reference signal $\mathbf{x}_{\mathrm{ref}}$, conditioning the conversion module $\mathcal{G}(\mathbf{x}, \mathbf{h}_{\mathrm{tech}})$ in FABYOL.

4. FABYOL: Timbral Technique Conversion Framework

We selected FACodec (Ju et al., 2024) as the foundation for FABYOL due to its robust attribute disentanglement and high-fidelity audio reconstruction, effectively separating speaker identity, content, prosody, and acoustic details. FACodec employs a neural codec with factorized vector quantization (FVQ) to decompose speech into distinct subspaces, enabling precise manipulation of vocal attributes. However, FACodec was originally designed and trained for speech, not singing. In our preliminary tests, we observed that the pre-trained model does not generalize well to singing voice. For this reason, we limit the scope of this paper to speech-based techniques (e.g., modal speech, whisper, and non-melodic scream), leaving the extension to singing as a direction for future work

The FACodec process is briefly outlined as follows: An input waveform $\mathbf{x}$ is encoded into a latent representation $\mathbf{h}$, which is then factorized into content embeddings $\mathbf{z}_{\mathrm{c}}$, acoustic details embeddings $\mathbf{z}_{\mathrm{d}}$ using their respective quantizers, and timbre embeddings $\mathbf{h}_{\mathrm{tim}}$ extracted by a transformer encoder. Prosody embeddings $\mathbf{z}_{\mathrm{p}}$ is derived from frame-wise acoustic features using a separate transformer and quantizer. These components are recombined, conditioned by timbre via adaptive layer normalization, and decoded into the output waveform $\mathbf{y}$. This architecture disentangles speaker-specific traits from content and prosody, leveraging vector quantization to preserve nuanced vocal features, making it a strong base for our conversion system.

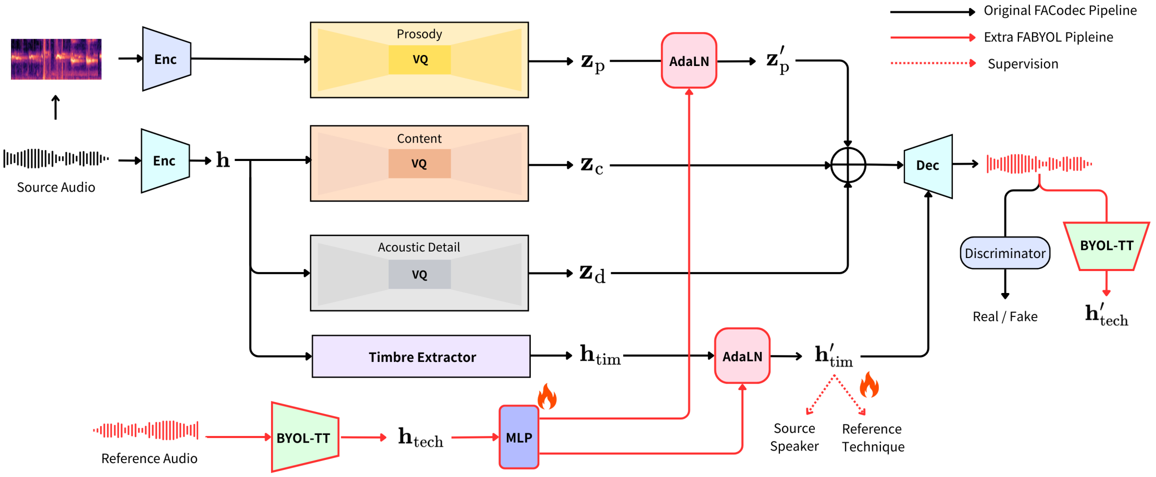

Figure 1: The FABYOL framework for timbral technique conversion. It integrates a BYOL-TT technique extractor and employs AdaLN to modulate both prosody and timbre. Black arrows depict the original FACodec pipeline, red arrows highlight the additional FABYOL pipeline, and the red dashed arrow indicates supervision. The fire icons mark the only components trained during the process while all other parts of the framework remain frozen.

4.1. Dual Attribute Modulation

Our analysis of FACodec revealed that its timbre subspace—originally intended to encode speaker identity—also entangles timbral technique. This entanglement complicates the task of converting specific vocal techniques. Moreover, our experiments show that realistic technique conversion cannot rely on timbre alone; Techniques like whisper and vocal fry exhibit distinct prosodic behaviors—whisper often features reduced pitch variation and energy, while vocal fry is characterized by irregular, low-frequency modulations. These observations suggest that technique modeling requires joint modulation of timbre and prosody.

To achieve that, we evaluated several conditioning strategies, including concatenation (Qian et al., 2019), cross-attention (Li et al., 2024), AdaIN (Chou et al., 2019), and FiLM (Perez et al., 2018), and ultimately chose AdaLN (Peebles and Xie, 2023), a frame-level AdaIN variant. AdaLN effectively removes source-specific global information before injecting target traits (Chou et al., 2019), which is beneficial for our setting where source techniques vary across training. Its frame-level modulation also better captures the temporal nuances of vocal techniques, which are not strictly global attributes like speaker identity.

FABYOL retains the original encoders, attribute vector quantizers, and decoder structure from FACodec, all kept frozen throughout the process. Central to our method, the technique embedding $\mathbf{h}_\mathrm{tech} \in \mathbb{R}^{C_\mathrm{tech}}$ is processed via a multilayer perceptron to produce scale and shift parameters:

$$[\boldsymbol{\gamma}_p(\mathbf{h}_\mathrm{tech}), \boldsymbol{\beta}_p(\mathbf{h}_\mathrm{tech}), \boldsymbol{\gamma}_t(\mathbf{h}_\mathrm{tech}), \boldsymbol{\beta}_t(\mathbf{h}_\mathrm{tech})] = \mathrm{MLP}(\mathbf{h}_\mathrm{tech}), \tag{1}$$

where $\boldsymbol{\gamma}_p(\mathbf{h}_{\text{tech}}), \boldsymbol{\beta}_p(\mathbf{h}_{\text{tech}}) \in \mathbb{R}^C$ are parameters for conditioning the prosody subspace, and $\boldsymbol{\gamma}_t(\mathbf{h}_{\text{tech}}), \boldsymbol{\beta}_t(\mathbf{h}_{\text{tech}}) \in \mathbb{R}^C$ condition the timbre subspace. The prosody parameters are later extended across the time dimension to match the length T'.

We apply AdaLN to modulate both prosody and timbre components, as depicted by the red arrows in Figure 1. For prosody, we normalize each time frame to zero-mean and unit-variance across the channel dimension. We compute the modulated prosody embedding frame as

$$\begin{aligned} \mathbf{z'}_{\text{p},t} &= \text{AdaLN}(\mathbf{z}_{\text{p},t}, \mathbf{h}_{\text{tech}}) \\ &= \boldsymbol{\gamma}_p(\mathbf{h}_{\text{tech}}) \cdot \frac{\mathbf{z}_{\text{p},t} - \mu(\mathbf{z}_{\text{p},t})}{\sigma(\mathbf{z}_{\text{p},t})} + \boldsymbol{\beta}_p(\mathbf{h}_{\text{tech}}), \end{aligned} \tag{2}$$

where $\mu(\mathbf{z}_{\text{p},t})$ and $\sigma(\mathbf{z}_{\text{p},t})$ represent the mean and standard deviation of $\mathbf{z}_{\text{p},t}$ over its channels.

Similarly, for the timbre embedding, we use

$$\begin{aligned} \mathbf{h}'_{\text{tim}} &= \text{AdaLN}(\mathbf{h}_{\text{tim}} \mid \mathbf{h}_{\text{tech}}) \\ &= \boldsymbol{\gamma}_t(\mathbf{h}_{\text{tech}}) \cdot \frac{\mathbf{h}_{\text{tim}} - \mu(\mathbf{h}_{\text{tim}})}{\sigma(\mathbf{h}_{\text{tim}})} + \boldsymbol{\beta}_t(\mathbf{h}_{\text{tech}}), \end{aligned} \tag{3}$$

where $\mu(\mathbf{h}_{\text{tim}})$ and $\sigma(\mathbf{h}_{\text{tim}})$ are computed across the timbre vector's dimension. This dual application of AdaLN ensures both prosody and timbre are reconfigured to reflect the target technique's characteristics.

In the final stage, the modulated prosody embeddings $\mathbf{z}'_{\text{p}}$, along with $\mathbf{z}_{\text{c}}$ and $\mathbf{z}_{\text{d}}$, is summed to form $\mathbf{z'}_{\text{sum}} \in \mathbb{R}^{C \times T'}$. This is then conditioned by conditional layer normalization using the modulated timbre embeddings, $\mathbf{z}_{\text{cond}} = \text{AdaLN}(\mathbf{z'}_{\text{sum}}, \mathbf{h}'_{\text{tim}})$, and passed through the frozen decoder to synthesize the output waveform $\hat{\mathbf{x}}$. Inspired by adaptive normalization in style transfer (Chou et al., 2019) and diffusion transformers (Peebles and Xie, 2023), our AdaLN-based $\mathcal{G}(\mathbf{x}, \mathbf{h}_{\text{tech}})$ efficiently transfers target technique traits with minimal modification to the existing architecture.

4.2. Cross-Speaker Unpaired Reference

In our design, we tackled the real-world challenge of users providing unrelated reference samples by adopting an unpaired reference method during training, similar to (Chen et al., 2024). We transform a source utterance $\mathbf{x}$ using a randomly chosen reference $\mathbf{x}_{\text{ref}}$ with the target technique t_{ref}, selected from different speakers and linguistic content than $\mathbf{x}$. This mirrors practical use cases, unlike traditional supervised training that relies on ground truth audio (Liu, 2024). By decoupling technique-specific learning from speaker identity and content, this approach improves disentanglement and generalization. Our training integrates reconstruction and bidirectional paired data conversion across all transformations, boosting the model's adaptability to diverse, unseen speaker scenarios.

4.3. Training Objective

Our loss-driven optimization strategy builds upon FACodec with a total loss:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{mel}}\mathcal{L}_{\text{mel}} + \lambda_{\text{aux}}\big(\mathcal{L}_{\text{p}} + \mathcal{L}_{\text{tim}} + \mathcal{L}_{\text{tech}} + \mathcal{L}_{\text{cls\_spkr}} + \mathcal{L}_{\text{cls\_tech}} + \mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{feat}}\big).$$

We retain the original FACodec losses $\mathcal{L}_{\text{mel}}$, $\mathcal{L}_{\text{adv}}$, and $\mathcal{L}_{\text{feat}}$. Additionally, we minimize L_1 distances for prosody $\mathcal{L}_{\text{p}} = \|\mathbf{Z}'_{\text{p}} - \mathbf{Z}^{\text{GT}}_{\text{p}}\|_1$, timbre $\mathcal{L}_{\text{tim}} = \|\mathbf{h}'_{\text{tim}} - \mathbf{h}^{\text{GT}}_{\text{tim}}\|_1$, and technique $\mathcal{L}_{\text{tech}} = \|\mathbf{h}'_{\text{tech}} - \mathbf{h}^{\text{GT}}_{\text{tech}}\|_1$. Supervision is applied via cross-entropy losses $\mathcal{L}_{\text{cls\_spkr}}$ and $\mathcal{L}_{\text{cls\_tech}}$ to enforce speaker identity and technique accuracy respectively. Empirically, we set $\lambda_{\text{mel}} = 10$ and $\lambda_{\text{aux}} = 5$.

5. Experimental Setup

5.1. Dataset

Our study utilized audio data from a total of 130 unique speakers from several datasets. After manually quality-filtering, the JVS dataset (Takamichi et al., 2019) provided 1,500 parallel clips (500 modal voice, 500 whisper, 500 falsetto) from 100 speakers. The EMVD dataset (Tailleur et al., 2024) provided 270 clips (135 modal voice and 135 scream) from 27 vocalists. Our self-recorded data included the EMO dataset—consisting of 706 clips (353 modal voice and 353 scream) recorded by a single professional metal singer in both Chinese and English—and 6 clips (3 modal voice and 3 scream) from two other speakers. Finally, we used one scream sample from the `Genera Studios Metal Screams`[1] sample pack.

All audio was resampled to 16 kHz, and trimmed via voice activity detection (VAD) (Silero Team, 2024). Paired data alignment varied by source: JVS and EMVD clips were time-stretched with `pyrubberband`[2] using a DTW time map (Müller, 2007), while the EMO clips were manually aligned in a Digital Audio Workstation. For normalization, audio was handled differently by technique: modal, falsetto, and scream were set to a modal RMS reference, while whisper was normalized to its own RMS reference. The data was then partitioned; JVS and EMVD were split into speaker-disjoint training and test sets at approximately a 9:1 ratio, and the EMO dataset was randomly split at the same 9:1 ratio. A reference set was constructed by selecting two clips for each technique from the test set speakers. Our evaluation protocol involved two distinct tasks: 1) Reconstruction: each test file was processed using itself as the reference; 2) Conversion: each test file was converted using all files in the reference set to generate the final outputs.

5.2. Implementation Details

We trained the BYOL-TT encoder on a NVIDIA RTX 3090 GPU. It transforms 1-second Mel-spectrograms (16 kHz, 1024-point FFT, 1024-sample window, 160-sample hop) into 1024-dimensional technique embeddings $\mathbf{h}_{\text{tech}}$. Training took one day with a batch size of 256 and a learning rate of 10^{-4}. The encoder remains frozen for FABYOL after training. FABYOL is trained on a NVIDIA RTX 3090 GPU. It processes 1-second Mel-spectrograms with the same parameters, using a 10-layer MLP with SiLU activation to project $\mathbf{h}_{\text{tech}}$ into 256-dimensional parameters conditioning prosody and timbre subspaces. Training lasted approximately two days with a batch size of 16 and a learning rate of 2×10^{-4}.

[1] https://generastudios.com/products/metal-screams

[2] https://github.com/bmcfee/pyrubberband

5.3. Evaluation Metrics

Objective Metrics. We first employ Mel Cepstral Distortion (MCD) (Kubichek, 1993) to assess overall audio quality between converted and ground truth audio samples. To evaluate whisper conversion, we use Harmonic-to-Noise Ratio (HNR) (Fernandes et al., 2018), a well-established metric in traditional vocal analysis research, as whispering typically results in reduced harmonic content—values closer to GT indicate better conversion. For falsetto conversion, we employ the average fundamental frequency (AF0) as a proxy. While falsetto is a complex acoustic phenomenon, F0 is its most obvious characteristic, with falsetto voice typically exhibiting a significantly higher F0 than modal voice (Keating, 2014). Therefore, results nearer to the ground truth AF0 reflect higher fidelity in the conversion. To assess speaker identity preservation, we evaluate existing speaker verification (SV) models across various timbral techniques. We find that SV models often assign lower similarity scores when comparing a speaker's ground truth modal voice to their own utterances in other techniques than when comparing modal utterances from different speakers. This suggests a bias toward modal conditions and a limited ability to capture speaker identity across techniques. To address this, we propose a more robust, technique-agnostic evaluation: cross-gender modal-to-modal conversion. We compute speaker embedding cosine similarity (**SEC**) by Resemblyzer[3] between source and converted samples.
Subjective Metrics. We employ three subjective metrics in our user study: TSMOS to evaluate the timbral technique similarity, NMOS to assess the perceptual quality and naturalness of the audio, and SSMOS to measure the speaker similarity. Twenty listeners evaluated 8 sets of samples, comparing baseline and proposed models with a reference audio.

5.4. Baseline Models

We compare FABYOL against three state-of-the-art baseline models representing distinct paradigms in voice conversion and timbral control: 1) FreeVC (Li et al., 2023): a text-free, one-shot voice conversion system with VITS framework; 2) CosyVoice (Du et al., 2024): a scalable, zero-shot TTS system based on supervised semantic tokens; 3) FACodec (Ju et al., 2024): a neural codec factorizes speech into multiple attributes; 4) FABYOL: The timbral technique conversion model proposed in this paper.

6. Experimental Result

6.1. Reconstruction

Table 1 presents the MRSTFT loss (Yamamoto et al., 2020) across various vocal techniques, illuminating the reconstruction strengths of FACodec and FABYOL. FACodec excels at reconstructing diverse timbral techniques—like whisper, falsetto, and scream—with quality matching modal voice, despite their absence from its training data. This highlights its robust disentanglement of content, prosody, timbre, and acoustic details, showcasing its adaptability to new vocal styles and solidifying its role as a key foundation for our research.

However, FACodec's success depends on how well its timbre extractor works. If it only captures speaker identity and misses the unique sound characteristics of these techniques,

[3] https://github.com/resemble-ai/Resemblyzer

Table 1: Comparison of reconstruction MRSTFT loss (Yamamoto et al., 2020) across vocal techniques: M = Modal, F = Falsetto, W = Whisper, S = Scream.

Model	M	F	W	S	Overall
FACodec (Ju et al., 2024)	0.86	0.91	0.86	1.14	0.948
FABYOL	1.26	1.24	1.01	1.56	1.281

which are mixed with speaker identity, the quality of reconstruction could drop. The fact that it handles such variety well suggests its timbre representation is flexible and has good potential for our work. FABYOL builds on this by adding technique embeddings as conditioning inputs, maintaining strong reconstruction quality.

6.2. Efficacy of Timbral Technique Conversion

Objective Performance. As shown in Table 2, FABYOL achieves the lowest MCD, delivering superior spectral fidelity across techniques like whisper, falsetto, and scream. It outperforms FreeVC and CosyVoice, which favor broad timbral shifts over technique-specific details, and FACodec, which struggles to fully convert timbral techniques due to its lack of prosody modulation. In whisper conversion, FABYOL's HNR closely matches ground truth, capturing the weak tonality of whispers. Baselines, lacking prosody modulation, miss this nuance and produce overly harmonic outputs. Similarly, FABYOL's AF0 aligns near-perfectly with ground truth, thanks to prosody embedding modulation enabling precise spectral control—an area where other baseline models lag behind. FABYOL also excels in preserving speaker identity, as demonstrated by its strong performance in SEC. Its success is driven by targeted augmentations, unpaired reference training, and classifier-guidance, which together effectively separate technique from identity.

Subjective Performance. Subjective results (Table 2) reinforce FABYOL's core strengths. Its technique similarity (TSMOS) was rated significantly higher than all baselines, approaching ground truth and confirming the successful direction of the conversion. Furthermore, FABYOL achieved the highest speaker preservation (SSMOS) scores, validating its ability to robustly preserve speaker identity. However, it is important to note the lower naturalness (NMOS), as listeners perceived synthesis artifacts. We hypothesize this is a limitation of our lightweight conditioning approach. While our AdaLN successfully guides the type of technique, its affine transformation may lack the capacity to fully condition the backbone's internal features, a step required for artifact-free synthesis, especially for extreme techniques. Training a generative model from scratch would be ideal but is currently unfeasible given the scarcity of large-scale, diverse timbral datasets. Our approach therefore prioritized data-efficiency and establishing foundational feasibility. While the lower NMOS is a clear limitation, the high TSMOS and SSMOS scores confirm we met our primary objective for this fundamental study: establishing that controlled, speaker-preserving technique conversion is possible. Improving synthesis quality by exploring more powerful conditioning methods or by collecting the diverse data needed to train a dedicated model, is the clear next step for future work.

Table 2: Performance comparison of timbral technique conversion across models.

Models	MCD ↓	HNR	AF0	SEC ↑	TSMOS ↑	NMOS ↑	SSMOS ↑
GT	—	1.32	381	—	3.93 ± 0.86	4.08 ± 0.69	—
FreeVC (Li et al., 2023)	8.93	17.01	248	0.58	—	—	—
CosyVoice (Du et al., 2024)	8.28	11.97	328	0.68	1.49 ± 0.25	**4.50 ± 0.20**	2.32 ± 1.67
FACodec (Ju et al., 2024)	8.55	10.55	298	0.74	2.18 ± 0.78	3.23 ± 0.47	2.45 ± 0.35
FABYOL (ours)	**7.59**	**-1.27**	**379**	**0.79**	**3.64 ± 0.43**	2.86 ± 0.49	**4.66 ± 0.25**

Table 3: Ablation study results of dual modulation.

Modulations	MCD ↓	HNR	AF0	SEC ↑
w/o prosody	9.27	**–0.94**	198	0.78
FABYOL	**7.59**	–1.27	**379**	**0.79**

Table 4: Ablation study results of different augmentations.

Augmentations	MCD ↓	WHNR	AF0	SEC ↑
BYOL-TT-DSP	7.90	**–1.04**	392	0.71
BYOL-TT-SEL	**7.59**	–1.27	**379**	**0.79**

6.3. Ablation Study

Dual Attribute Modulation. As detailed in Table 3, removing prosody modulation noticeably degrades spectral fidelity, especially for techniques like falsetto. Speaker identity remains stable with or without prosody modulation, indicating that prosody primarily enhances technique accuracy rather than identity consistency. These results underscore the importance of prosody-aware modulation for faithfully transferring fine-grained vocal timbral techniques.

Augmentation Strategies. Results in Table 4 indicate that DSP-based augmentations improve technique representation through targeted transformations, but methods like VTLP often compromise speaker consistency due to weaker disentanglement. In contrast, real-data selection strikes the best balance—capturing subtle vocal techniques, preserving spectral detail, and maintaining speaker identity—making it the most effective strategy for timbral technique transfer.

7. Conclusion

We present FABYOL, the first model for vocal timbral technique conversion while preserving speaker identity. Our approach surpasses prior work in spectral fidelity, technique similarity, and identity consistency by modulating both timbre and prosody via embedding-guided AdaLN. We also introduce the EMO dataset to provide a high-quality, paired corpus for this task, with a specific focus on vocal fry scream. Future work should focus on improving output naturalness and developing scream-specific metrics. A key priority is to extend the model to handle melodic content, enabling the application of timbral techniques to singing. We also plan to expand datasets to include more timbral techniques and languages to enable real-time applications.

References

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020.

Yu-Hua Chen, Yen-Tung Yeh, Yuan-Chiao Cheng, Jui-Te Wu, Yu-Hsiang Ho, Jyh-Shing Roger Jang, and Yi-Hsuan Yang. Towards zero-shot amplifier modeling: One-to-many amplifier modeling via tone embedding control. In *Proc. Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, 2024.

Ju-Chieh Chou, Cheng-chieh Yeh, and Hung-yi Lee. VocalSet: A singing voice dataset. In *Proc. Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, pages 493–500, 2018.

Ju-chieh Chou, Cheng-chieh Yeh, and Hung-yi Lee. One-shot voice conversion by separating speaker and content representations with instance normalization. In *Proc. Interspeech*, 2019.

Marius Cotescu, Thomas Drugman, Goeric Huybrechts, Jaime Lorenzo-Trueba, and Alexis Moinet. Voice conversion for whispered speech synthesis. *IEEE Signal Processing Letters*, 27:186–190, 2020.

Shuqi Dai, Yunyun Wang, Roger B Dannenberg, and Zeyu Jin. Everyone-Can-Sing: Zero-shot singing voice synthesis and conversion with speech reference. *arXiv preprint arXiv:2501.13870*, 2025.

Yimin Deng, Huaizhen Tang, Xulong Zhang, Ning Cheng, Jing Xiao, and Jianzong Wang. Learning disentangled speech representations with contrastive learning and time-invariant retrieval. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2024.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. CosyVoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*, 2024.

Zongyang Du, Berrak Sisman, Kun Zhou, and Haizhou Li. Disentanglement of emotional style and speaker identity for expressive voice conversion. In *Proc. Interspeech*, 2021.

Gasser Elbanna, Neil Scheidwasser-Clow, Mikolaj Kegler, Pierre Beckmann, Karl El Hajal, and Milos Cernak. BYOL-S: Learning self-supervised speech representations by bootstrapping. In *HEAR: Holistic Evaluation of Audio Representations*, 2022.

Joana Fernandes, Felipe Teixeira, Vitor Guedes, Arnaldo Junior, and João Paulo Teixeira. Harmonic to noise ratio measurement-selection of window and length. *Procedia Comput. Sci.*, 138, 2018.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33, 2020.

Yu-Tung Hsu, Jun-You Wang, and Jyh-Shing Roger Jang. GTsinger: A global multi-technique singing corpus with realistic music scores for all singing tasks. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2024.

Qiong Hu, Tomi Bleisch, Petko Petkov, Tuomo Raitio, Erik Marchi, and V Lakshminarasimhan. Whispered and lombard neural speech synthesis. In *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2021.

Navdeep Jaitly and Geoffrey E Hinton. Vocal tract length perturbation (VTLP) improves speech recognition. In *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, 2013.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, et al. NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. In *Proc. Int. Conf. Mach. Learn. (ICML)*, 2024.

Vedant Kalbag and Alexander Lerch. Scream detection in heavy metal music. In *Proc. Sound Music Comput. (SMC)*, 2022.

Patricia A. Keating. Acoustic measures of falsetto voice. Presentation slides from the Annual Meeting of the Acoustical Society of America, May 2014. URL https://pdfs.semanticscholar.org/0e85/ae9390039cc8b1be2fe4cb32ad2486b865c4.pdf.

Robert Kubichek. Mel-cepstral distance measure for objective speech quality assessment. In *Proc. IEEE Pac. Rim Conf. Commun. Comput. Signal Process.*, 1993.

Jingyi Li, Weiping Tu, and Li Xiao. FreeVC: Towards high-quality text-free one-shot voice conversion. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2023.

Junjie Li, Yiwei Guo, Xie Chen, and Kai Yu. SEF-VC: Speaker embedding free zero-shot voice conversion with cross attention. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2024.

Songting Liu. Zero-shot voice conversion with diffusion transformers. *arXiv preprint arXiv:2411.09943*, 2024.

Meinard Müller. *Information Retrieval for Music and Motion.* Springer, 2007.

Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. BYOL for audio: Exploring pre-trained general-purpose audio representations. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 31, 2022.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proc. ICCV*, 2023.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2018.

Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. AutoVC: Zero-shot voice style transfer with only autoencoder loss. In *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019.

Aaqib Saeed, David Grangier, and Neil Zeghidour. Contrastive learning of general-purpose audio representations. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2021.

Silero Team. Silero VAD: pre-trained enterprise-grade voice activity detector (VAD), number detector and language classifier. `https://github.com/snakers4/silero-vad`, 2024.

Modan Tailleur, Julien Pinquier, Laurent Millot, Corsin Vogel, and Mathieu Lagrange. EMVD dataset: a dataset of extreme vocal distortion techniques used in heavy metal. In *Proc. Int. Conf. Content-Based Multimedia Indexing (CBMI)*, 2024.

Shinnosuke Takamichi, Kentaro Mitsui, Yuki Saito, Tomoki Koriyama, Naoko Tanji, and Hiroshi Saruwatari. JVS corpus: free Japanese multi-speaker voice corpus. *arXiv preprint arXiv:1908.06248*, 2019.

Hiromu Yakura, Kento Watanabe, and Masataka Goto. Self-supervised contrastive learning for singing voices. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 30, 2022.

Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pages 6199–6203, 2020.

Yuya Yamamoto, Juhan Nam, and Hiroko Terasawa. Analysis and detection of singing techniques in repertoires of J-POP solo singers. In *Proc. Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, 2022.

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 *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2021.