ADAPTING A DIFFUSION-BASED MUSIC SYNTHESIS MODEL TO HUMAN VOICE CONVERSION

Ben Maman[⋆], *Frank Zalkow*[†], *Hans-Ulrich Berendes*[⋆], *Paolo Sani*[†], *Christian Dittmar*[†], *Meinard Müller*[⋆]

[⋆] International Audio Laboratories Erlangen, Germany
[†] Fraunhofer IIS, Erlangen, Germany

ABSTRACT

Recent generative models have shown promising results in audio generation across various domains, including human speech, singing voice, and multi-instrument music synthesis. Such acoustic models are typically specialized, with separate systems for speech, singing, and instrumental music. However, real-world audio often comprises multiple domains—for instance, music recordings may combine a sung melody or spoken lyrics with instrumental accompaniment. This highlights the need for more general-purpose approaches to audio synthesis that can handle cross-domain integration. As an initial step towards universal synthesis, in this work we compare different acoustic models originating from distinct domains—instrumental music synthesis and speech synthesis—on the task of human voice conversion. Through an extensive evaluation across singing and speech, we demonstrate that a diffusion-based instrumental music synthesis model can be effectively adapted to human voice conversion, achieving performance comparable to or surpassing that of a dedicated speech synthesis model. We show that off-the-shelf feature extractors for phonetic content, pitch and acoustics provide effective conditioning signals for the synthesizer, enabling self-supervised training on large-scale datasets. Project page: `https://benadar293.github.io/voice-conversion`

Index Terms— Diffusion, Voice Conversion, Music Synthesis

1. INTRODUCTION

Recent generative models have achieved promising results in visual content generation and are now widely applied to audio tasks such as speech [1], singing [2, 3, 4], and instrumental music synthesis [5, 6, 7]. Such models typically generate spectral representations conditioned on text, phonetic content, pitch, or musical score, which are later converted to waveforms. While each domain—speech, singing, and instrumental music—has seen significant progress with specialized models, real-world audio often combines multiple content types, including vocals, musical instruments, and hybrid styles like rap or Sprechgesang. Thus, specialized models do not account for the diverse nature of real-world audio.

Recent work in diffusion-based multi-instrument music synthesis [6, 7] demonstrates the ability to generate complex musical signals with a wide range of instruments, timbres and acoustics using a single diffusion model, which is entirely attention-based. Although limited to instrumental music, its broad generative capability suggests that integration of other audio types including speech and singing may be possible using similar conditioning techniques.

As an initial step towards integrated and controlled music synthesis combining singing with instrumental music, in this work we adapt the latter instrumental music synthesis model to *human voice conversion*, considering speech and singing in a single acoustic model. We demonstrate that principles and techniques from instrumental music synthesis readily transfer to human voice conversion.

As our main contribution, through an extensive evaluation covering both naturalness and performer similarity, we show that the purely attention-based diffusion model used for instrumental music synthesis performs comparably or even better than a recent convolution-based flow-matching voice conversion model, in both speech and singing voice conversion.

As a second contribution, we demonstrate how off-the-shelf automatic feature extractors can provide effective conditioning signals, rather than using manually annotated datasets. We show this approach to be effective even under domain shift, such as in the case of a phonetic feature extractor trained on speech data, applied to singing data. Most importantly, it enables training a synthesizer on diverse large-scale datasets in an entirely self-supervised manner.

2. RELATED WORK

Voice Conversion (VC) converts a source speech audio sample to match the speaker identity of a given target speaker [8, 9]. High-quality conversion is achieved using various generative models, such as generative adversarial networks (GANs) [8], diffusion models [1], or more recently flow-matching models [10, 11]. In order to capture time-varying content from the source recording, previous work applies existing feature extractors for phonetic content and pitch, which serve as conditioning signals [12, 13].

Singing Voice Conversion (SVC) adapts a source singing sample to match a target singer's identity. The analogy between VC and SVC allows shared feature representations such as time-aligned lyrics or transcripts for linguistic content, and pitch contours for melody or prosody. Building upon this, [4] proposes a unified speech and singing VC model trained jointly on both domains, enabling singing synthesis from a speech reference. To address data scarcity, prior work often resorts to source-separated vocals, followed by lyrics–audio alignment [2, 3].

The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Grant No. 500643750 (MU 2686/15-1). The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) under the NHR project b265dc10. NHR funding is provided by federal and Bavarian state authorities. NHR@FAU hardware is partially funded by the German Research Foundation (DFG)—440719683

© 2025 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

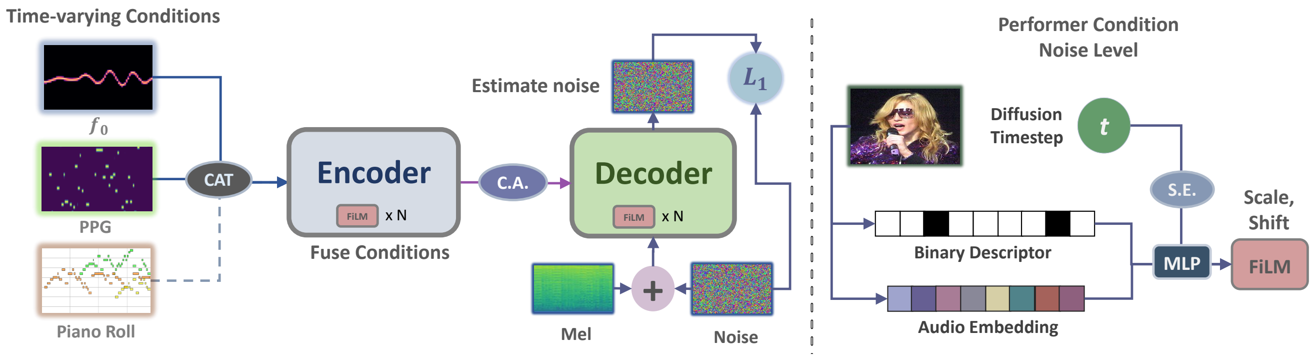

Fig. 1. Overview of our model. "CAT": Concatenation along the channel axis, "C.A.": Cross-attention, "S.E.": Sinusoidal embedding.

Instrumental Music Synthesis generates musical performances from musical scores. While earlier work focused on single instruments or monophonic music, more recent work enables polyphonic, multi-instrument synthesis [5], conditioning on the version to capture acoustic and stylistic aspects [6, 7]. We adapt similar architectures to support speech and singing voice conversion, incorporating conditioning on phonetic content, pitch, speaker and singer identity.

Audio Foundation Models [14, 15] can generate rich and complex audio including singing with instrumental accompaniment. Such models are typically prompted by a descriptive text input or similar meta-level information, providing a form of weak conditioning. However, they often lack explicit and fine-grained control over the different time-varying aspects of the generated audio, such as melody, phonetic content, and musical score.

3. METHOD

Following previous work in speech, singing, and instrumental music synthesis, our acoustic model is based on mel spectrogram diffusion. To convert the generated mel spectrogram into a waveform, we use an off-the-shelf BigVGAN vocoder [16].[1] We choose a general purpose vocoder rather than a vocal-only one —although this may come at the expense of quality—in order to assess potential for generating vocal–instrumental mixtures. Figure 1 presents an overview of our acoustic model. Generation is factorized into three components:

(i) Spectrogram generation is done using a diffusion-based spectral decoder trained to estimate noise from noisy mel spectrograms, serving as the generative backbone.

(ii) Time-varying conditioning is done via an encoder that learns a fused representation of the given time-varying conditions, which is provided to the spectral decoder as auxiliary input to guide generation. We replace the piano roll representation typically used to represent musical score in instrumental music synthesis [5, 7] with a phonetic posteriorgram (PPG) representing phonetic content, and an f_0 contour representing vocal pitch.

(iii) Performer conditioning is implemented via feature-wise linear modulations (FiLM) [17] applied to hidden features at each block (scale, shift). These are conditioned on an audio embedding and on an explicit binary descriptor, both representing the vocal performer, i.e., the *speaker or singer*. While this mechanism was used in instrumental synthesis to condition on timbre and acoustics [6, 7], we show it is similarly effective for conditioning on the performer.

3.1. Feature Extraction

We estimate vocal f_0 using CREPE [18], and PPG using a variant of wav2vec 2.0 [19, 20]. Although the PPG extractor was trained on speech data, we demonstrate it can provide meaningful conditioning for singing as well. Lastly, we obtain audio embeddings from TRILL [21], which was trained with a triplet loss such that snippets closer in time are closer in the embedding space.

3.2. Architecture

We experiment with the following models which are based on publicly available implementations: A T5-based diffusion model used in multi-instrument music synthesis [5, 6, 7] and the state-of-the-art FlowMAC [11], which is based on MatchaTTS [10], with conditioning done using the voice conversion system PAD-VC [12] based on Forward Tacotron, as done in [22]. We choose the music synthesis model for its ability to generate complex, multi-instrument music signals under varied acoustic conditions, suggesting potential for singing with accompaniment. We choose the speech model as a baseline for comparison and assessment of the music model's capability to handle vocal and phonetic content. While we use off-the-shelf architectures, many similar architectural alternatives exist—such as 1D convolution–attention hybrids previously used in singing and music synthesis [3, 4, 6, 7]. Task-specific architecture optimization for speech and singing remains important future work.

In the **T5 Transformer**-based diffusion model [7], both the decoder and the encoder are stacks of self-attention layers. The decoder receives the encoded conditioning information via interleaved cross-attention layers. Performer conditioning is applied to both the encoder and decoder through FiLM layers [17]. Conditioning on the noise level is done only in the decoder.

The **PAD-VC** model consists of 1D convolution blocks and long short-term memory (LSTM) layers. It is based on the ForwardTacotron decoder without the text alignment encoder which is not necessary in our setting, since our conditioning PPGs and f_0 contours are time-aligned. It is trained with a spectral L1 reconstruction loss. The resulting coarse spectrogram estimate serves as a conditioning signal for the subsequent flow-matching model.

The **FlowMAC** [11] model combines 1D convolutional residual blocks with attention layers trained as a flow matching model conditioned on the output of PAD-VC. It is based on the decoder part of MatchaTTS, without the text encoder, since the conditioning PPG and f_0 contour are time-aligned.

All models were trained as described in the original publications.

[1] `https://huggingface.co/nvidia/bigvgan_v2_22khz_80band_256x`

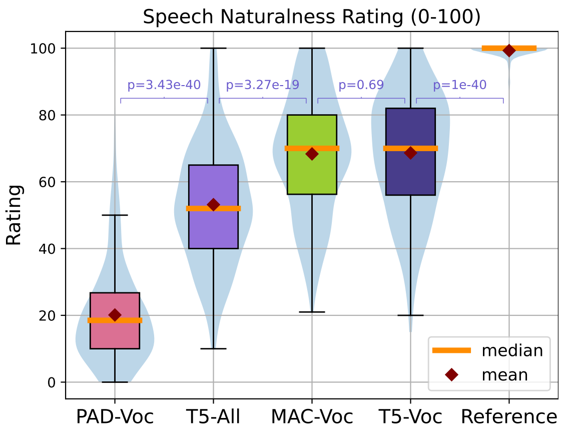

Fig. 2. Speech naturalness listening test results.

4. EXPERIMENTS

4.1. Datasets

We use the following datasets (only audio without annotation):

(i) **Speech ($\mathcal{D}_{\text{speech}}$):** A ∼33-hour compound dataset comprising proprietary and public sources, featuring recordings in the English language from two male and three female speakers. As a held-out set, we randomly sample 50 full-utterance excerpts from each of the five speakers, each excerpt 2–12 seconds long, totaling ∼19m, yielding a split $\mathcal{D}_{\text{speech}} = \mathcal{D}^{\text{train}}_{\text{speech}} \cup \mathcal{D}^{\text{test}}_{\text{speech}}$.

(ii) **Singing** ($\mathcal{D}_{\text{sing}}$)**:** A ∼31-hour compound dataset consisting of the following: The SingStyle111 dataset [23] containing ∼13 hours of solo singing from four male and four female singers, together with the source-separated vocals from the following $\mathcal{D}_{\text{mix}}$ which includes four female and 33 male singers, using only voice-active regions totaling ∼18h. Source separation was verified to be of high quality through informal listening tests, to ensure it does not affect evaluation. As a held-out set, we randomly sample three songs for each of the eight singers in SingStyle111, each song 1–7 minutes long, totaling ∼1.3h, yielding a split: $\mathcal{D}_{\text{sing}} = \mathcal{D}^{\text{train}}_{\text{sing}} \cup \mathcal{D}^{\text{test}}_{\text{sing}}$.

(iii) **Vocal–Instrumental Mix** ($\mathcal{D}_{\text{mix}}$)**:** ∼90h including the Schubert Winterreise dataset [24] (∼11h, nine male singers with piano accompaniment), popular music (∼34h, 24 male and four female singers) and instrumental Western classical music (∼47h).

4.2. Evaluation

We conduct qualitative listening tests focusing on two aspects: (i) quality and naturalness, and (ii) control over performer identity. We complement the listening tests by a quantitative evaluation based on the Fréchet Audio Distance [25]. We compare the following three models trained on the vocal data $\mathcal{D}_{\text{voc}}$: `T5-Voc`: A T5 diffusion-based model; `PAD-Voc`: The ForwardTacotron-based model trained with a spectral reconstruction loss; `MAC-Voc`: The FlowMAC model conditioned on the output of `PAD-Voc`.

Lastly, to assess the ability to handle mixed data, we include in our evaluation `T5-All`—the T5 diffusion model trained on $\mathcal{D}_{\text{all}}$, excluding the held-out sets. For this model, we condition on the instrumental musical score using a piano roll [7], which we simply concatenate to the vocal conditioning features (f_0 and PPG). We further condition it by concatenating a one-hot encoding of the data type (vocal only or with instrumentals) to the performer condition.

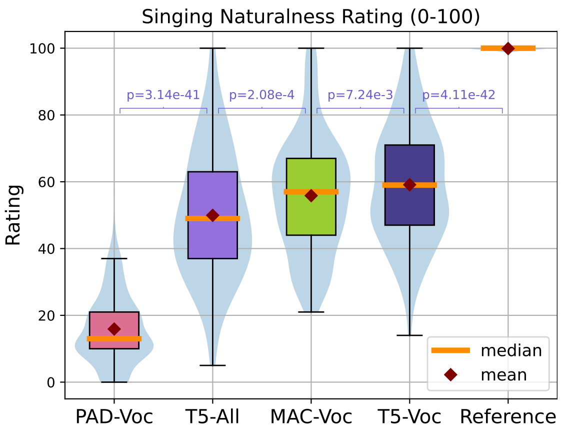

Fig. 3. Singing naturalness listening test results.

	Naturalness Rating↑	
	Speech	Singing
`T5-Voc`	**68.63±18.11**	**59.11±17.54**
`T5-All`	53.15±17.66	49.94±19.27
`PAD-Voc`	20.12±15.55	15.90±10.57
`MAC-Voc`	68.34±17.98	55.84±17.53
`Ref.`	99.28±4.49	99.89±1.64

Table 1. Speech and singing naturalness listening tests results.

4.2.1. Naturalness Listening Tests

To evaluate the quality of the synthesized audio in terms of naturalness and realism, we follow a listening test protocol with a hidden reference and a lower anchor similar to a standard Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) test [26], which usually produces statistically significant results with a relatively small number of participants. We perform two listening tests, one for speech and one for singing, comparing the four aforementioned models, in the task of reconstructing the original audio from the conditioning features, namely f_0, PPG, and performer identity. In this setting, the target identity is the same as in the source audio excerpt from which the f_0 and PPG were extracted.

As a reference we use the vocoded version of the original audio (i.e., the audio generated by the vocoder from the mel spectrogram of the original audio), which is the upper bound on the attainable quality. Listeners who rated the hidden reference lower than 95 more than once were discarded (post-screening). In the speech listening test, 26 listeners participated, one of which was post-screened, leaving 25 listeners. In the singing listening test, 29 listeners participated, four of which were post-screened, leaving 25 listeners.

Results for speech and singing appear in box plot and violin plot form in Figures 2 and 3, with pairwise p-values using a Wilcoxon signed-rank test. Mean and standard deviation values appear in Table 1. It can be seen that `T5-Voc` performs comparably or slightly better than `MAC-Voc`. `T5-Voc` was rated slightly higher than `MAC-Voc` in singing—with a mean rating of 59.11 compared to 55.84, and a p-value of $7.24e^{-3}$. In speech, the difference (68.63 compared to 68.34) is not significant ($p = 0.69$). We observe generally lower ratings for singing, which may indicate that expressive singing is harder to generate than speech.

A key takeaway from this experiment is that `T5-Voc` bor-

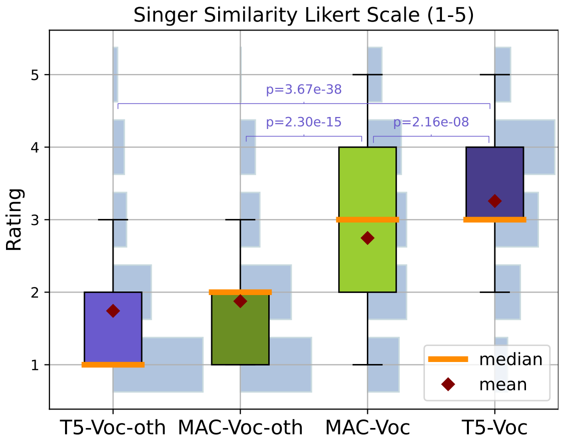

Fig. 4. Singer similarity listening test results.

rowed from instrumental synthesis can be applied to vocal synthesis, suggesting its potential for singing with instrumental accompaniment. However, adding instrumental data into training (T5-All) degrades perceived quality for both speech and singing by a considerable amount of 10–15 points. This can be expected— adding data from a distinct domain can increase complexity. and require higher capacity. Mitigating this effect remains important future work.

4.2.2. *Singer Similarity Listening Test*

In this test we investigate how closely the generated audio resembles recordings of a target singer, comparing the T5-Voc and MAC-Voc from the previous subsection in singing voice conversion.

For each question we randomly sample a reference singer ref, and another singer other of the same gender. A random source excerpt provides the f_0 and the PPG. Using each of the T5-Voc and MAC-Voc models, we sonify the source f_0 and PPG conditioned on ref and other, yielding four generated samples. Listeners are then presented with three random excerpts of ref to familiarize themselves with the target voice, and are asked to rate the similarity of each generated sample to ref, according to the following Likert scale: (1) "completely different person," (2) "probably different person," (3) "similar," (4) "probably the same person," and (5) "exactly the same person."

The test comprised ten randomly sampled questions, each with four generated and three reference samples. Excerpts were 3–12 seconds long, of full utterances drawn from a test set balanced between male and female singers. In total, twenty listeners participated.

Results appear in Figure 4, including p-values, and rating histograms for each Likert value and model (light-colored bars). For both T5-Voc and MAC-Voc, conditioning on the target singer substantially increases perceived similarity. For instance, T5-Voc ratings rise on average from 1.76 ± 1.06 when conditioned on another singer (T5-Voc-oth) to 3.25 ± 1.09 when conditioned on the reference singer (T5-Voc), with the most frequent rating being (4) "probably the same person".

We observe that T5-Voc achieves higher similarity ratings than MAC-Voc with a mean rating of 3.25 versus 2.75 and a p-value below 10^{-7}, indicating stronger conditioning. We hypothesize this stems from the conditioning in T5-Voc being applied both in the encoder and decoder, whereas conditioning of MAC-Voc is done only in its ForwardTacotron-based encoder. Conditioning its decoder could potentially improve performer similarity.

	FAD↓					
	Speech			Singing		
	$\mathcal{D}^{\text{test}}_{\text{speech}}$	$\mathcal{D}^{\text{train}}_{\text{speech}}$	Perf.	$\mathcal{D}^{\text{test}}_{\text{sing}}$	$\mathcal{D}^{\text{train}}_{\text{sing}}$	Perf.
T5-Voc	**0.162**	0.118	**0.156**	**0.108**	0.093	**0.141**
T5-All	0.187	0.141	0.190	0.142	0.128	0.192
PAD-Voc	0.345	0.288	0.343	0.271	0.257	0.356
MAC-Voc	0.171	**0.113**	0.163	0.110	**0.092**	0.178
Vocoder	0.002	0.054	0.573	0.004	0.065	0.369

Table 2. FAD results for speech and singing.

A key takeaway from this experiment is that the same conditioning technique used for acoustics in instrumental music synthesis [7] is similarly effective for conditioning on the singer in human VC.

4.2.3. *Fréchet Audio Distance (FAD)*

We complement the listening tests with a quantitative evaluation using the Fréchet Audio Distance (FAD) [25] which measures a distributional distance between two sets. Following [5, 6, 7], we use the TRILL model [21] which produces five audio embeddings per second. We apply two variants: **(i) All-FAD** corresponds to overall quality, measuring the distance between all generated audio, and the entire reference set of real recordings; **(ii) Performer-FAD** corresponds to performer similarity, measuring the distance between audio generated with a specific performer condition, and the set of real recordings of the same performer. To evaluate conversion quality, we render the held-out sets with randomly sampled other target performers—speakers for speech and singers for singing.

Table 2 reports the results. For All-FAD, we compare generated audio to the train and test sets (cols. 2–3, 5–6). For Performer-FAD, we compare synthesized and real subsets corresponding to performers (cols. 4, 7). T5-Voc produces best overall scores, usually surpassing MAC-Voc. The difference is most prominent for Performer-FAD—for example, in singing, T5-Voc yields 0.141 compared to 0.178 of MAC-Voc, consistent with the similarity listening test.

We hypothesize the generally lower distances w.r.t. the train sets are due to the train set being larger, thus of smoother distribution.

While the vocoded source yields best All-FAD scores, it yields worst Performer-FAD scores, since the source and target performers differ, confirming that performer conditioning shifts the generated audio distribution towards the target performer.

It is also clear that adding instrumental data (T5-All) worsens scores, consistent with the listening tests (Section 4.2.1).

5. CONCLUSION

In this work we evaluated the performance of an attention-based diffusion model adapted from instrumental music synthesis to human voice conversion. Through an extensive evaluation across speech and singing we have shown it can match—or even surpass—a dedicated voice conversion model in terms of quality and performer similarity. While our results also indicate a decline in vocal synthesis quality when including training data with instrumental music, the model's demonstrated ability to handle both instrumental and vocal synthesis within a unified framework underscores its potential for generating complex signals, such as singing with instrumental accompaniment. Exploring and evaluating generation of such signals remains an important direction for future work.

6. REFERENCES

[1] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, Mikhail Sergeevich Kudinov, and Jiansheng Wei, “Diffusion-based voice conversion with fast maximum likelihood sampling scheme,” *in International Conference on Learning Representations (ICLR)*, 2022.

[2] Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao, “DiffSinger: Singing voice synthesis via shallow diffusion mechanism,” in *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

[3] Shuqi Dai, Ming-Yu Liu, Rafael Valle, and Siddharth Gururani, “Expressivesinger: Multilingual and multi-style score-based singing voice synthesis with expressive performance control,” in *Proc. ACM Multimedia*, pp. 3229-3238, 2024.

[4] Shuqi Dai, Yunyun Wang, Roger B. Dannenberg, and Zeyu Jin, “Everyone-can-sing: Zero-shot singing voice synthesis and conversion with speech reference,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[5] Curtis Hawthorne, Ian Simon, Adam Roberts, Neil Zeghidour, Josh Gardner, Ethan Manilow, and Jesse H. Engel, “Multi-instrument music synthesis with spectrogram diffusion,” in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2022.

[6] Ben Maman, Johannes Zeitler, Meinard Müller, and Amit H. Bermano, “Performance conditioning for diffusion-based multi-instrument music synthesis,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2024.

[7] Ben Maman, Johannes Zeitler, Meinard Müller, and Amit H. Bermano, “Multi-aspect conditioning for diffusion-based music synthesis: Enhancing realism and acoustic control,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 33, pp. 1–14, 2024.

[8] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo, “Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks,” in *IEEE Spoken Language Technology Workshop (SLT)*, 2018.

[9] Berrak Sisman, Junichi Yamagishi, Simon King, and Haizhou Li, “An overview of voice conversion and its challenges: From statistical modeling to deep learning,” *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 132–157, 2021.

[10] Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, and Gustav Eje Henter, “Matcha-TTS: A fast TTS architecture with conditional flow matching,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[11] Nicola Pia, Martin Strauss, Markus Multrus, and Bernd Edler, “FlowMAC: Conditional flow matching for audio coding at low bit rates,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[12] Arunava Kr. Kalita, Christian Dittmar, Paolo Sani, Frank Zalkow, Emanuël A. P. Habets, and Rusha Patra, “PAD-VC: A prosody-aware decoder for any-to-few voice conversion,” in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2024.

[13] Cameron Churchwell, Max Morrison, and Bryan Pardo, “High-fidelity neural phonetic posteriorgrams,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW), 2024.*

[14] Kai Shen, Zeqian Ju, Xu Tan, Eric Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian, “NaturalSpeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers,” in *International Conference on Learning Representations (ICLR)*, 2024.

[15] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez, “Simple and controllable music generation,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[16] Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon, “BigVGAN: A universal neural vocoder with large-scale training,” in *International Conference on Learning Representations (ICLR)*, 2023.

[17] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville, “FiLM: Visual reasoning with a general conditioning layer,” in *Proc. AAAI Conference on Artificial Intelligence*, 2018.

[18] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello, “CREPE: A convolutional representation for pitch estimation,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[19] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *Advances in Neural Information Processing Systems (NeurIPS)* , 2020.

[20] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli, “XLS-R: Self-supervised cross-lingual speech representation learning at scale,” in *Proc. Interspeech*, pp. 2278–2282, 2022.

[21] Joel Shor, Aren Jansen, Ronnie Maor, Oran Lang, Omry Tuval, Félix de Chaumont Quitry, Marco Tagliasacchi, Ira Shavitt, Dotan Emanuel, and Yinnon Haviv, “Towards learning a universal non-semantic representation of speech,” in *Proc. Interspeech*, pp. 140–144, 2020.

[22] Frank Zalkow, Paolo Sani, Kishor Kayyar Lakshminarayana, Emanuël A.P̃. Habets, Nicola Pia, and Christian Dittmar, “Bridging the training–inference gap in TTS: Training strategies for robust generative postprocessing for low-resource speakers,” in *Proc. Interspeech*, pp. 2470-2474, 2025.

[23] Shuqi Dai, Yuxuan Wu, Siqi Chen, Roy Huang, and Roger B. Dannenberg, “Singstyle111: A multilingual singing dataset with style transfer,” in *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2023.

[24] Christof Weiß, Frank Zalkow, Vlora Arifi-Müller, Meinard Müller, Hendrik Vincent Koops, Anja Volk, and Harald Grohganz, “Schubert Winterreise dataset: A multimodal scenario for music analysis,” *ACM Journal on Computing and Cultural Heritage (JOCCH)*, vol. 14, no. 2, pp. 25:1–18, 2021.

[25] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi, “Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms,” in *Proc. Interspeech*, pp. 2350–2354, 2019.

[26] International Telecommunications Union, “ITU-R Rec. BS.1534-3: Method for the subjective assessment of intermediate quality levels of coding systems,” 2015.