DDDM-VC: Decoupled Denoising Diffusion Models with Disentangled Representation and Prior Mixup for Verified Robust Voice Conversion

KOREA UNIVERSITY

Ha-Yeong Choi * Sang-Hoon Lee * Seong-Whan Lee

AAAI Association for the Advancement of Artificial Intelligence

Scan and Enjoy our demo

Project page

Audio sample

Introduction

Objective

Proposing a decoupled denoising diffusion models (DDDMs) with disentangled representations, which can enable effective style transfers for each attribute in generative models

Voice Conversion

Converting the voice of a source speaker into the voice of a specific target speaker while preserving the source speaker's linguistic information

Motivations

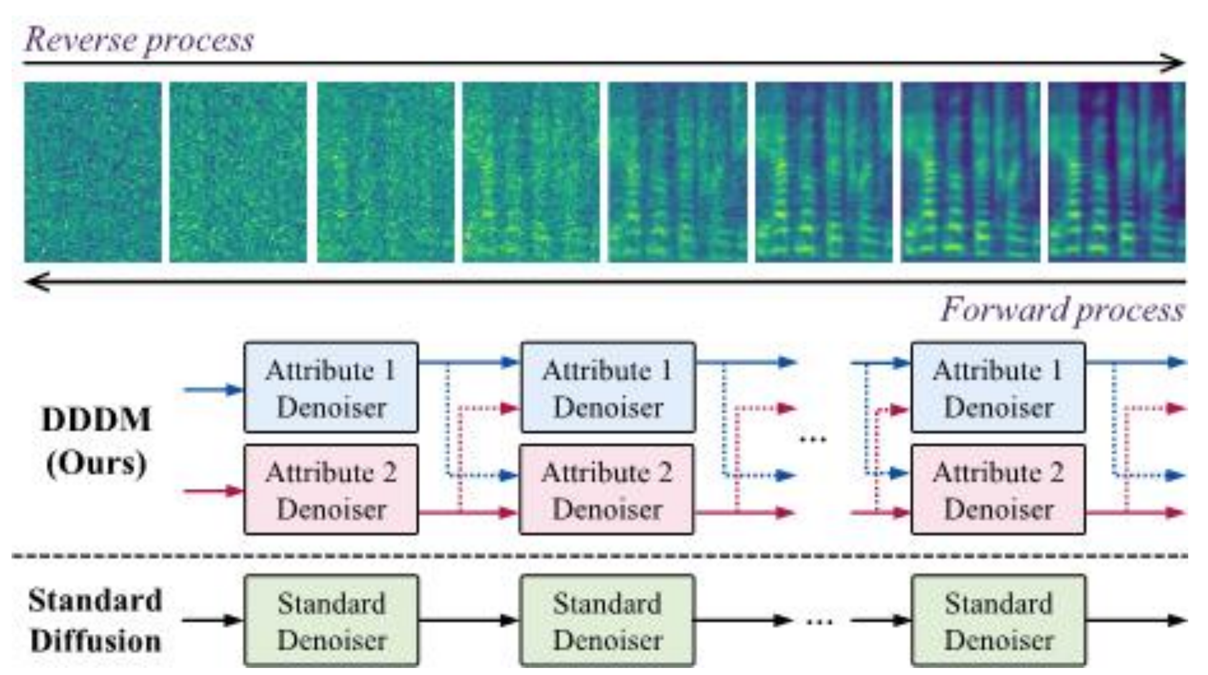

Fig 1. Speech synthesis in DDDM and standard diffusion model. Although a single denoiser with the same parameter is used for all denoising steps in standard diffusion models, we subdivide the denoiser into multiple denoisers for each attribute. For each intermediate time step, each denoiser focuses on removing the single noise from its attribute

* Controlling each attribute in speech
* Disentangling the speech components
* Improving the speaker similarity and intelligibility
* High-quality waveform reconstruction

Contribution

* Controlling the style for each attribute in generative models by decoupling attributes and adopting the disentangled denoisers
* Presenting DDDM-VC, which can disentangle and resynthesize speech for each attribute with self-supervised speech representation
* Proposing a prior mixup to improve VC performance
* Achieving superior performance in both many-to-many and zero-shot voice style transfer

Method

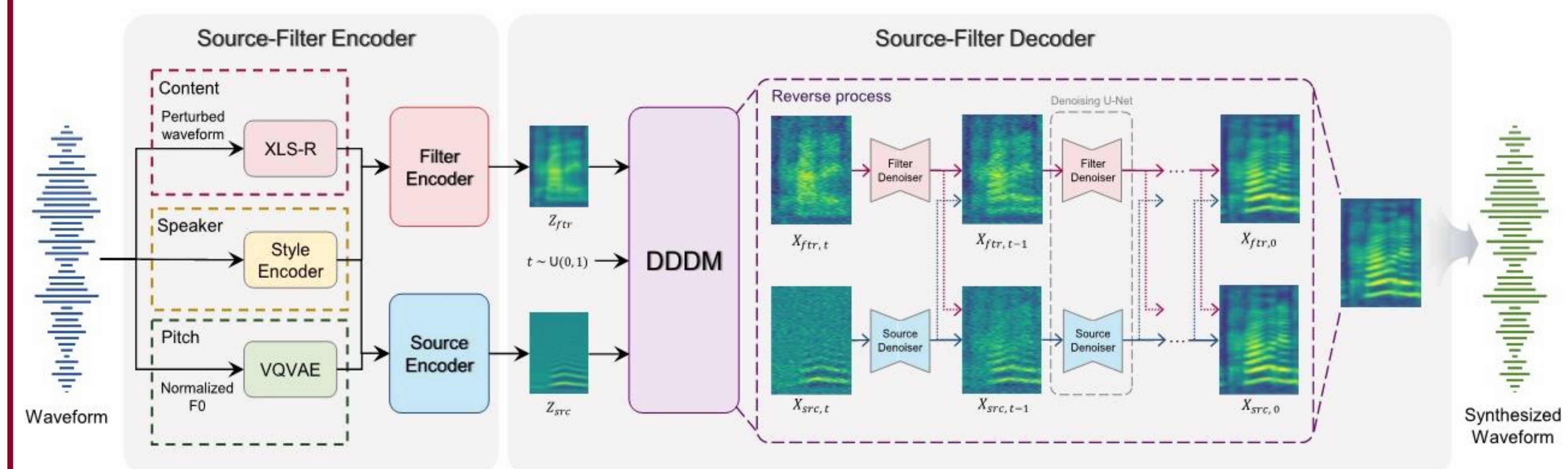

Fig2. proposed DDDM-VC

Speech Disentanglement

* Linguistic representation
  - Using a self-supervised speech representation from the 12th layer representation of XLS-R
  - Adopting speech perturbation to the 16 kHz audio used as the input for the XLS-R to eliminate content-irrelevant information (Using the formant shifting, pitch randomization, random frequency shaping)
* Pitch representation
  - Extracting the F0 using YAAPT to encode the intonation
  - Normalizing and vector quantizing the F0 for speaker-independent pitch representation
* Speaker representation
  - Utilizing a style encoder to extract the speech style from the Mel-spectrogram

Decoupled Denoising Diffusion Models

* Forward process

$$dX_{n,t} = \frac{1}{2}\beta_t(Z_n - X_{n,t})dt + \sqrt{\beta_t}d\overrightarrow{W_t}$$

* Reverse process of each disentangled denoiser

$$d\hat{X}_{n,t} = (\frac{1}{2}(Z_n - \hat{X}_{n,t}) - \sum_{n=1}^{N} s_{\theta_n}(\hat{X}_{n,t}, Z_n, t))\beta_t dt + \sqrt{\beta_t}d\overleftarrow{W_t}$$

* Optimizing Objective

$$\theta_n^* = \arg\min_{\theta_n} \int_0^1 \lambda_t \mathbb{E}_{X_0, X_{n,t}} \left\| \sum_{n=1}^{N} s_{\theta_n}(X_{n,t}, Z_n, s, t) - \nabla \log p_{t|0}(X_{n,t}|X_0) \right\|_2^2 dt$$

$t \in [0,1]$, $\overrightarrow{W_t}$, $\overleftarrow{W_t}$: two independent Wiener process in $\mathbb{R}^n$
β_t: non-negative function referred to as noise schedule
$X_{n,t}$: noisy sample and n: each prior attribute
s_θ: score function with parameter θ
$P_{t|0}(X_t|X_0)$: PDF of the conditional distribution
$\theta = [\theta_1, \dots, \theta_N]$ and $\lambda_t = 1 - e^{-\int_0^t \beta_s ds}$

Prior Mixup

* To address the training-inference mismatch, we use the randomly converted representation instead of the reconstructed representation as a prior
* Random speaker style selection

Table 1. Verify that the training-inference mismatch can be resolved in the decoder with prior mixup, we compare the VC results using the reconstructed (not converted) Mel-spectrogram

Method	Encoder Output (Prior)	EER (↓)	SECS (↑)
w/o prior mixup	Recon. Mel	48.34	0.677
w/ prior mixup	Recon. Mel	7.10	0.852

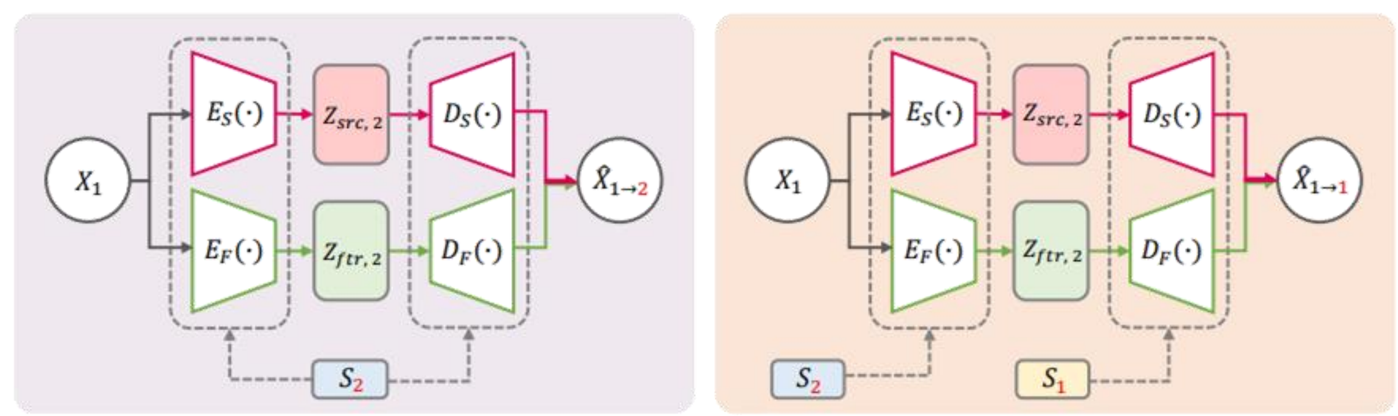

Fig 3. (left) Voice conversion (right) Prior mixup

Experiment and Result

Many-to-Many VC

Table 2. Many-to-many VC results on seen speakers from LibriTTS dataset

Method	iter.	nMOS (↑)	sMOS (↑)	CER (↓)	WER (↓)	EER (↓)	SECS (↑)	Params. (↓)	Real-time (↑)
GT	-	3.82±0.05	3.44±0.03	0.54	1.84	-	-	-	-
GT (Mel + Vocoder)	-	3.81±0.05	3.23±0.05	0.60	2.19	-	0.986	13M	-
AutoVC (Qian et al. 2019)	-	3.62±0.05	2.44±0.04	5.34	8.53	33.30	0.703	30M	×99.13
VoiceMixer (Lee et al. 2021a)	-	3.75±0.05	2.74±0.05	2.39	4.20	16.00	0.779	52M	×123.03
SR (Polyak et al. 2021)	-	3.62±0.05	2.55±0.04	6.63	11.72	33.30	0.693	15M	×177.22
DiffVC (Popov et al. 2022)	6	3.77±0.05	2.72±0.05	7.28	12.80	10.50	0.817	123M	× 20.06
DiffVC (Popov et al. 2022)	30	3.77±0.05	2.77±0.05	7.99	13.92	11.00	0.817	123M	×4.63
DDDM-VC-Small (Ours)	6	3.75±0.05	2.75±0.05	3.25	5.80	6.25	0.826	21M	×28.73
DDDM-VC-Small (Ours)	30	**3.79±0.05**	**2.81±0.05**	4.25	7.51	6.25	0.827	21M	×6.65
DDDM-VC-Base (Ours)	6	3.75±0.05	2.75±0.05	**1.75**	**4.09**	**4.00**	0.843	66M	×22.75
DDDM-VC-Base (Ours)	30	**3.79± 0.05**	2.80±0.05	2.60	5.32	4.24	**0.845**	66M	×5.09

Zero-shot VC

Table 3. Zero-shot VC results on unseen speakers from VCTK dataset.

Method	iter.	nMOS (↑)	sMOS (↑)	CER (↓)	WER (↓)	EER (↓)	SECS (↑)	MCD_{13} (↓)
GT	-	4.28±0.06	3.87±0.03	0.21	2.17	-	-	-
GT (Mel + Vocoder)	-	4.03±0.07	3.82±0.03	0.21	2.17	-	0.989	0.67
AutoVC (Qian et al. 2019)	-	2.49±0.09	1.88±0.08	5.14	10.55	37.32	0.715	5.01
VoiceMixer (Lee et al. 2021a)	-	3.43±0.08	2.63±0.08	1.08	3.31	20.75	0.797	4.49
SR (Polyak et al. 2021)	-	2.58±0.10	2.03±0.07	2.12	6.18	27.24	0.750	5.12
DiffVC (Popov et al. 2022)	6	3.48±0.07	2.62±0.08	5.82	11.76	25.30	0.786	4.82
DiffVC (Popov et al. 2022)	30	3.62±0.07	2.50±0.07	6.92	13.19	24.01	0.785	5.00
DDDM-VC-Small (Ours)	6	3.76±0.07	2.99±0.07	1.27	3.77	6.51	0.852	**4.39**
DDDM-VC-Small (Ours)	30	3.84±0.06	2.96±0.07	1.95	4.70	6.89	0.851	4.55
DDDM-VC-Base (Ours)	6	3.74±0.07	2.98±0.07	**1.00**	**3.49**	**6.25**	0.856	4.42
DDDM-VC-Base (Ours)	30	**3.88±0.06**	**3.05±0.07**	1.77	4.35	6.49	**0.858**	4.54
DDDM-VC-Fine-tuning (Ours)	6	3.74±0.07	**3.07±0.07**	1.26	3.80	**0.81**	**0.910**	**4.27**
DDDM-VC-Fine-tuning (Ours)	30	3.86±0.07	3.06±0.07	1.87	4.63	0.82	0.913	4.38

Ablation Study

Table 4. Results of ablation study on many-to-many VC tasks with seen speakers from LibriTTS

Method	iter.	nMOS (↑)	sMOS (↑)	CER (↓)	WER (↓)	EER (↓)	SECS (↑)	Params. (↓)
DDDM-VC-Small (Ours)	30	-	-	4.25	7.51	6.25	0.827	21M
DDDM-VC-Base (Ours)	30	3.76±0.05	3.08±0.05	2.60	5.32	4.24	0.845	66M
w/o Prior Mixup	30	3.79±0.05	3.03±0.05	3.28	5.66	7.99	0.821	66M
w/o Disentangled Denoiser	30	3.76±0.05	3.00±0.05	3.20	5.57	9.75	0.815	36M
w/o Normalized F0	30	3.78±0.05	3.00±0.05	3.27	5.88	10.25	0.811	33M
w/o Data-driven Prior	30	3.83±0.05	2.87±0.05	2.32	4.86	19.25	0.786	66M

Application: DDDM-TTS

Based on the DDDM-VC, we train the text-to-vec (TTV) model which can generate the self-supervised speech representation (the representation from the middle layer of XLS-R) from the text as a content representation.

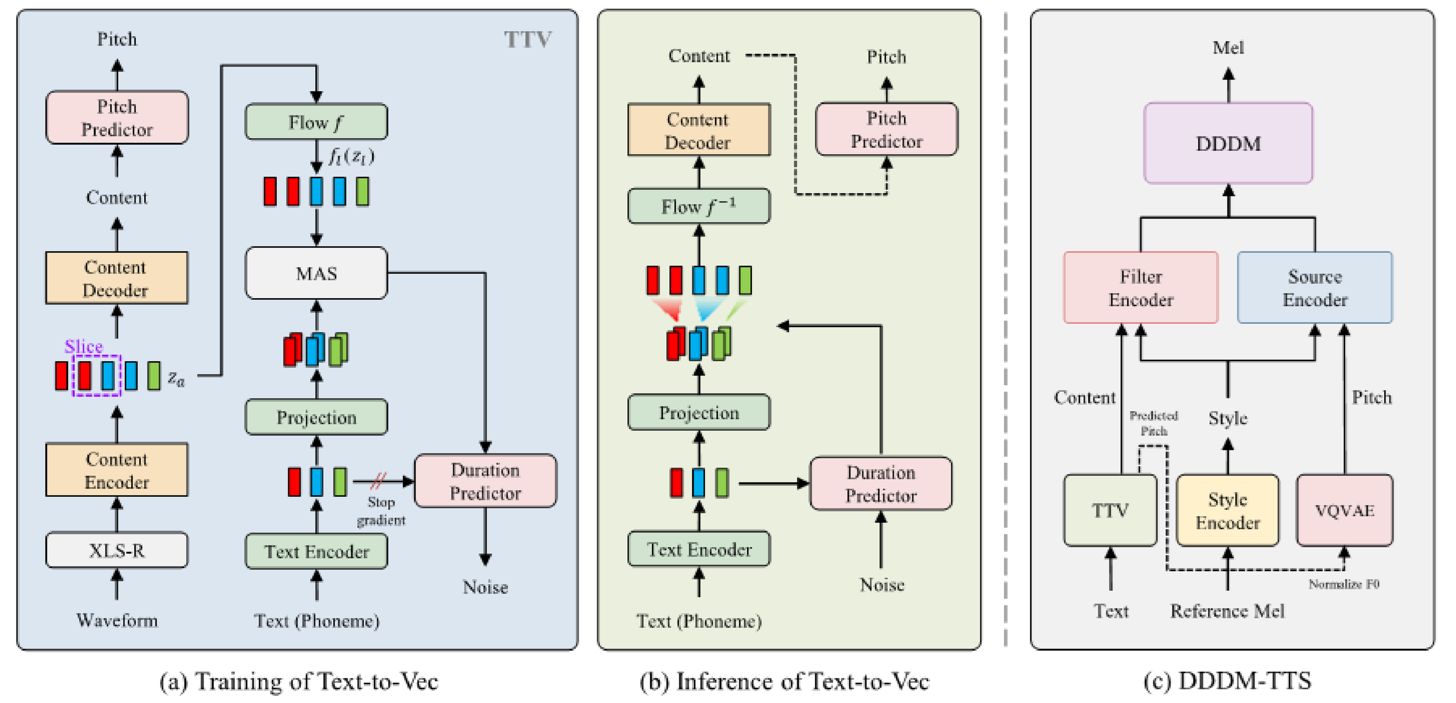

Fig 4. Overall framework of DDDM-TTS