ROBUST DISENTANGLED VARIATIONAL SPEECH REPRESENTATION LEARNING FOR ZERO-SHOT VOICE CONVERSION

Jiachen Lian^, Chunlei Zhang*, Dong Yu*

^UC Berkeley, BAIR, CA *Tencent AI Lab, Bellevue, WA

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

BAIR

Tencent AI Lab

ICASSP 2022
Session: SPE-19:Voice Conversion-Representation
May 7 - 13, 2022, Singapore

Outline

- Introduction
  - Limitations of current VC systems
  - Solutions achieving robust zero-shot VC
- Disentangled Sequential VAE for Zero-shot Voice Conversion
  - Overall framework
  - Disentanglement-aware probabilistic graphical models
  - Training Objectives
  - Noise-invariant VC
- Experimental results
- Conclusion

Voice Conversion and Limitations

Tencent AI Lab

VC Objective: swapping the speaker while keeping the content unchanged

Limitations of current VC systems:

1. Parallel training
2. Non-parallel training: speakers are pre-known
3. Zero-shot VC:
   (1) Speaker labels are used in data loader
   (2) Speaker embedding is pre-trained with labels

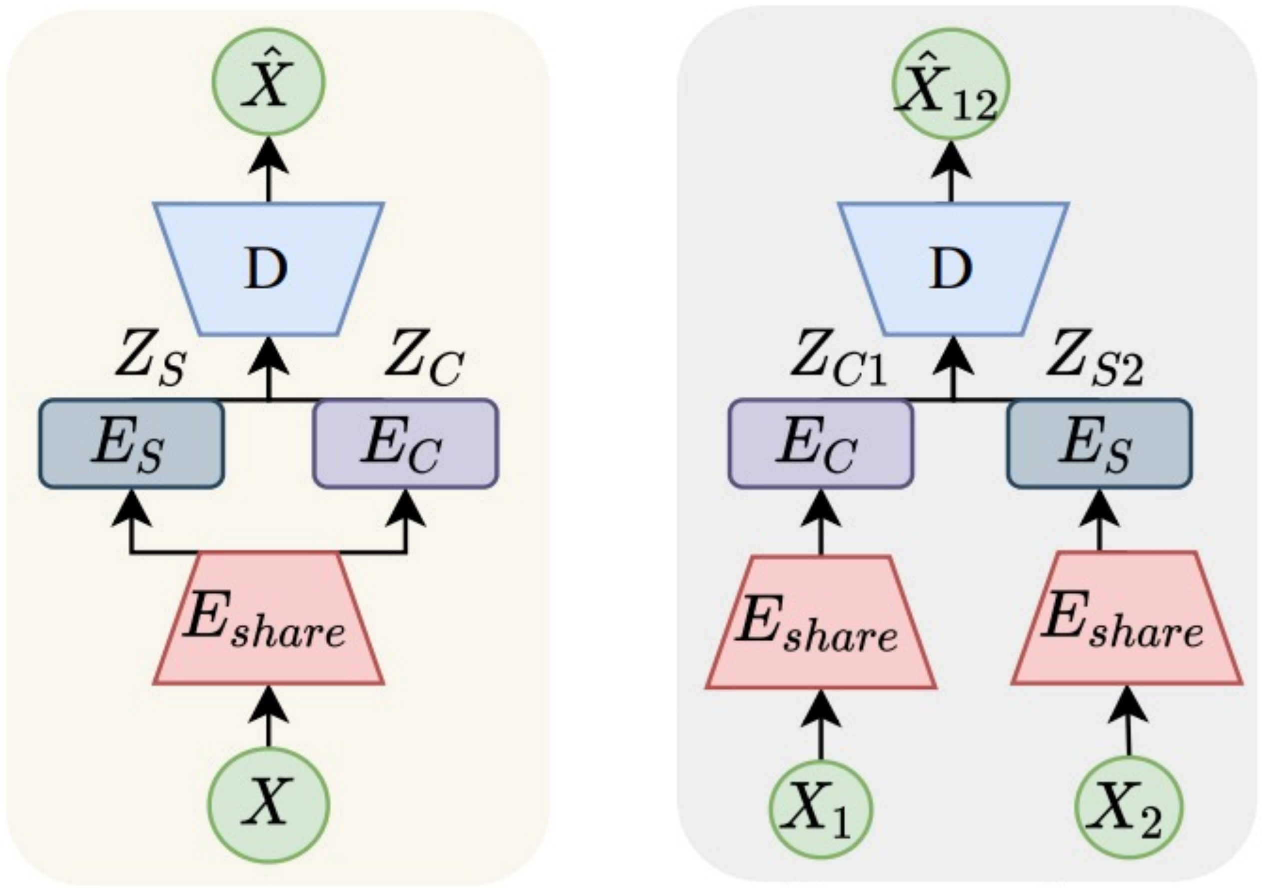
Ours:

1. Non-parallel training
2. Zero-shot
3. No speaker labels. No pre-trained speaker embeddings.
4. Noise-invariant

Disentangled Sequential Variational Auto-Encoder Framework

Tencent AI Lab

DSVAE-VC Diagram



a) training   b) conversion

X: *Melspec*

E_{share}: *Shared Encoder*

E_S: *Speaker Encoder*

E_C: *Content Encoder*

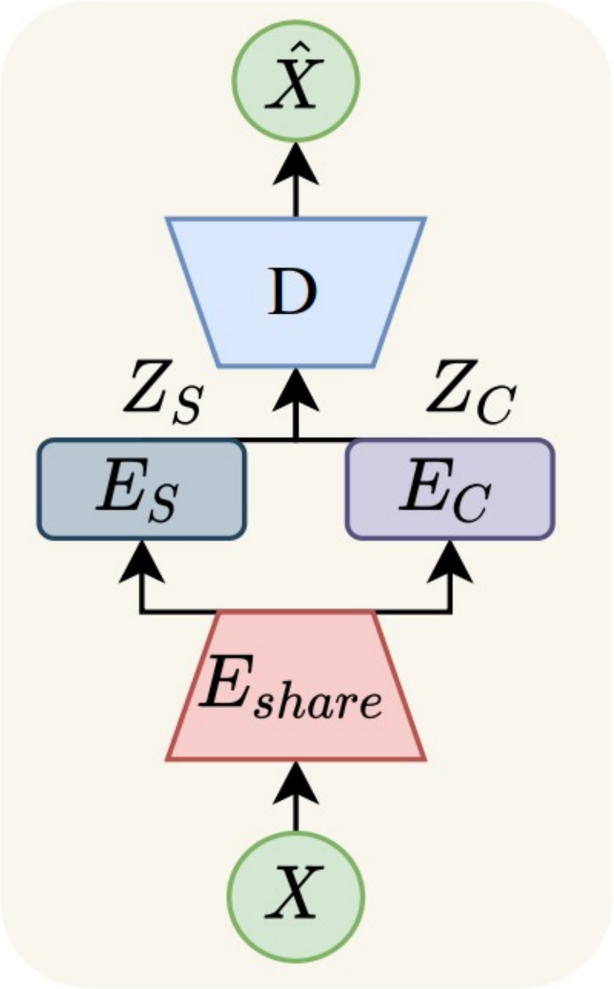
D: *Decoder*

Z_S: *Speaker Embedding*

Z_C: *Content Embedding*

Vocoder: *Wavenet*

Disentanglement-Aware Probabilistic Graphical Models



Independence (Disentangled) Factorization

Prior:

$$p_\theta(Z) = p(Z_S)p_\theta(Z_C) = p(Z_S)\prod_{t=1}^{T} p_\theta(z_{Ct}|z_{<t})$$

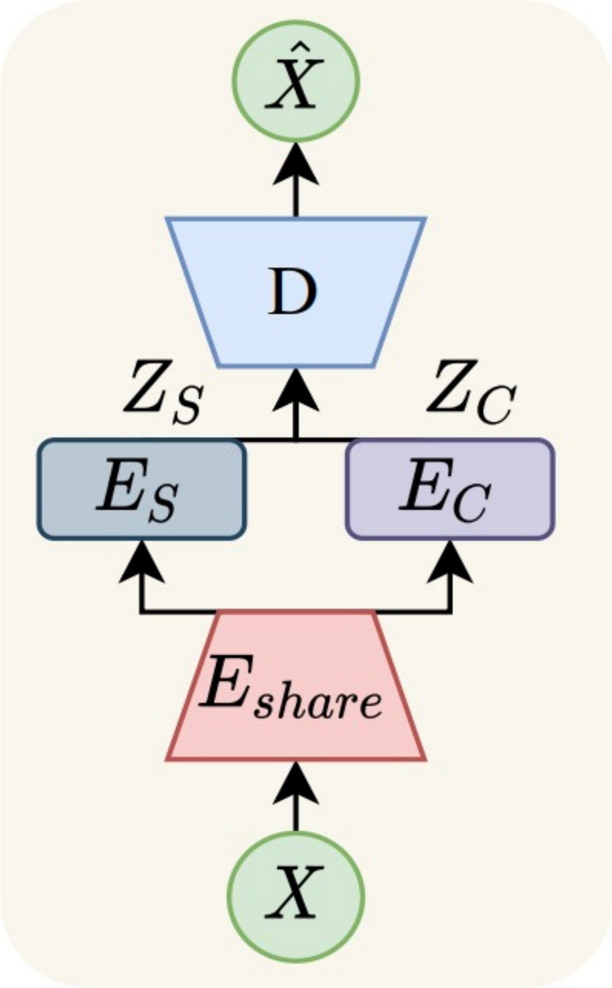
Standard Gaussian

Autoregressive LSTM

Posterior:

$$q_\theta(Z|X) = q_\theta(Z_S, Z_C|X) = q_\theta(Z_S|X)q_\theta(Z_C|X)$$

Training Objectives



Prior:

$$p_\theta(Z) = p(Z_S)p_\theta(Z_C) = p(Z_S)\prod_{t=1}^{T} p_\theta(z_{Ct}|z_{<t})$$

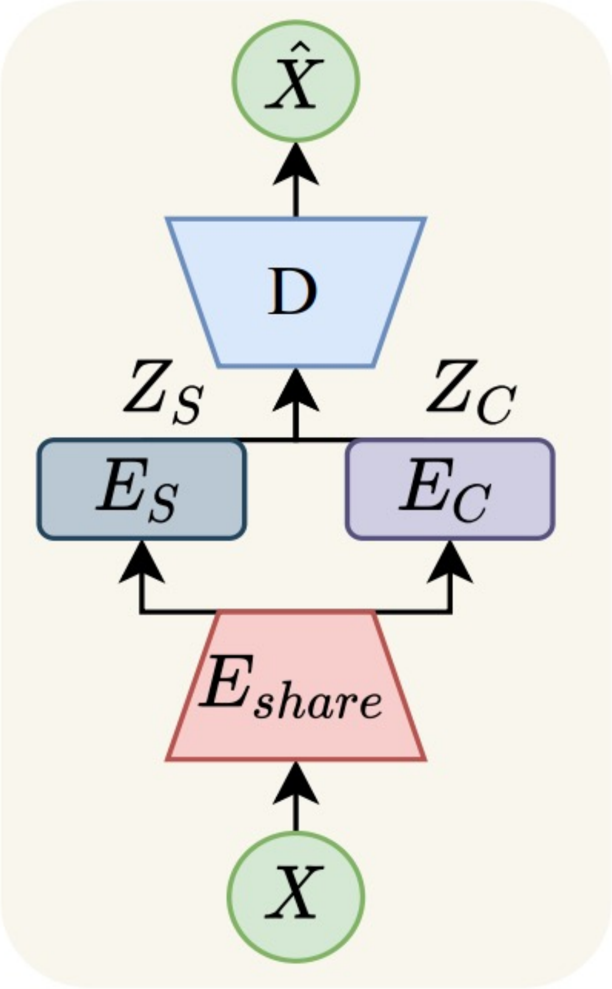
Posterior:

$$q_\theta(Z|X) = q_\theta(Z_S, Z_C|X) = q_\theta(Z_S|X)q_\theta(Z_C|X)$$

Loss Objectives:

$$\mathcal{L} = \mathbb{E}_{p(X)}\mathbb{E}_{q_\theta(X|Z_S,Z_C)}[-log(p_\theta(X|Z_S, Z_C))]+$$
$$\mathbb{E}_{p(X)}[\alpha kl(p(Z_S)||q_\theta(Z_S|X)) + \beta kl(p_\theta(Z_C)||q_\theta(Z_C|X))]$$

Training Objectives - Disentanglement



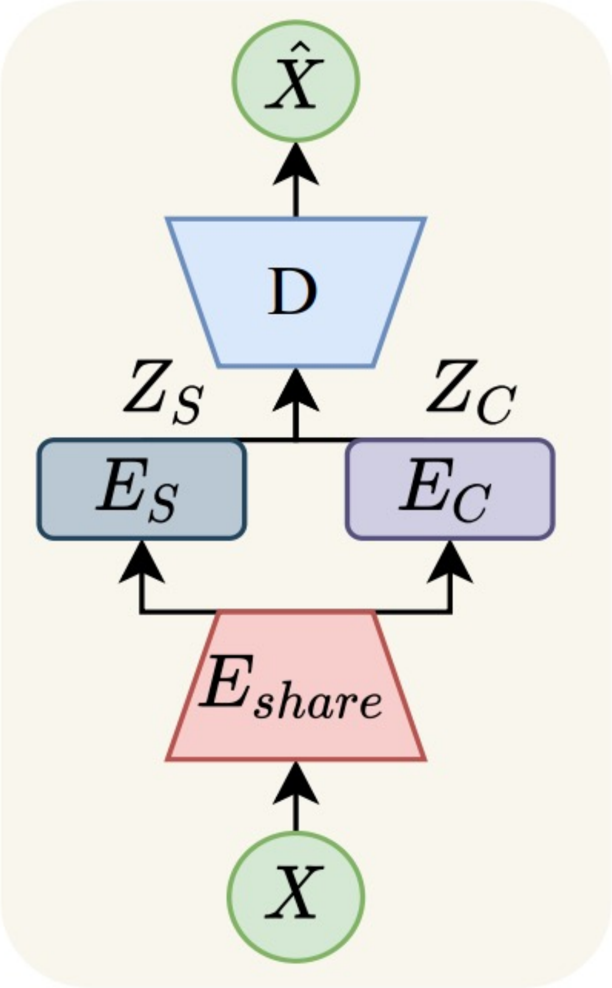
Loss Objectives:

$$\mathcal{L} = \mathbb{E}_{p(X)}\mathbb{E}_{q_\theta(X|Z_S,Z_C)}[-log(p_\theta(X|Z_S,Z_C))]+$$
$$\mathbb{E}_{p(X)}[\alpha kl(p(Z_S)||q_\theta(Z_S|X)) + \beta kl(p_\theta(Z_C)||q_\theta(Z_C|X))]$$

How is disentanglement achieved?

1. Balancing factors and KL vanishing
2. Time-domain Normalization

Noise-invariant Voice Conversion



Denoising Auto-Encoder!

clean utterance is augmented by MUSAN dataset with a balanced "noise", "music" and "babble" distribution

Dataset

1. TIMIT Dataset

   (1) Train/Test split
   The official training set/test set with 462 speakers/24 speakers.
   Following [1], all 18336 trials in test set are used for speaker verification.
   (2) Acoustic Features
   200 dimensional STFT features with 25ms/10ms framing configuration.
   During training, segment length is fixed to 20 frames.

2. VCTK Dataset

   (1) Train/Test split
   90% of the speakers are randomly selected for training and the remaining 10% as for testing.
   Randomly generate 36900 trials from test set for speaker verification.
   (2) Acoustic Features
   80 dimensional melspectrogram as features with 64ms/16ms framing configuration.
   During training, segment length is fixed to 100 frames.

[1] Yingzhen, Li, and Stephan Mandt. "Disentangled sequential autoencoder." *International Conference on Machine Learning*. PMLR, 2018.

Evaluation Metrics

Tencent AI Lab

1. Equal Error Rate

Speaker Embedding ⬇

Content Embedding ⬆

2. MOS

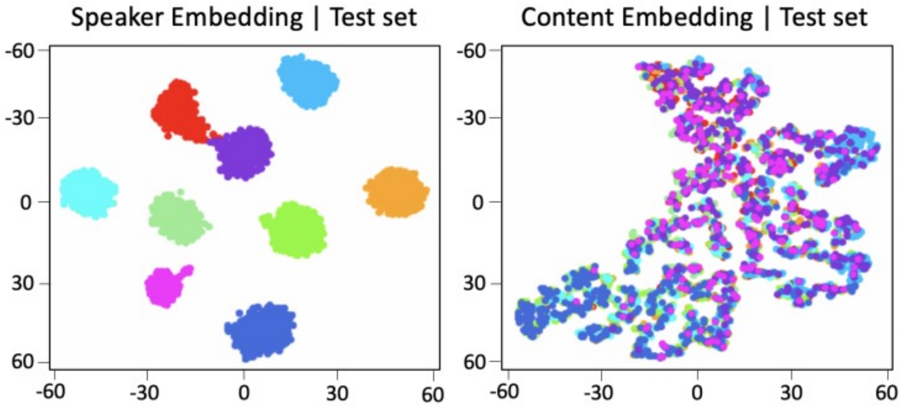
1 = Bad; 2 = Poor; 3 = Fair; 4 = Good; 5 = Excellent

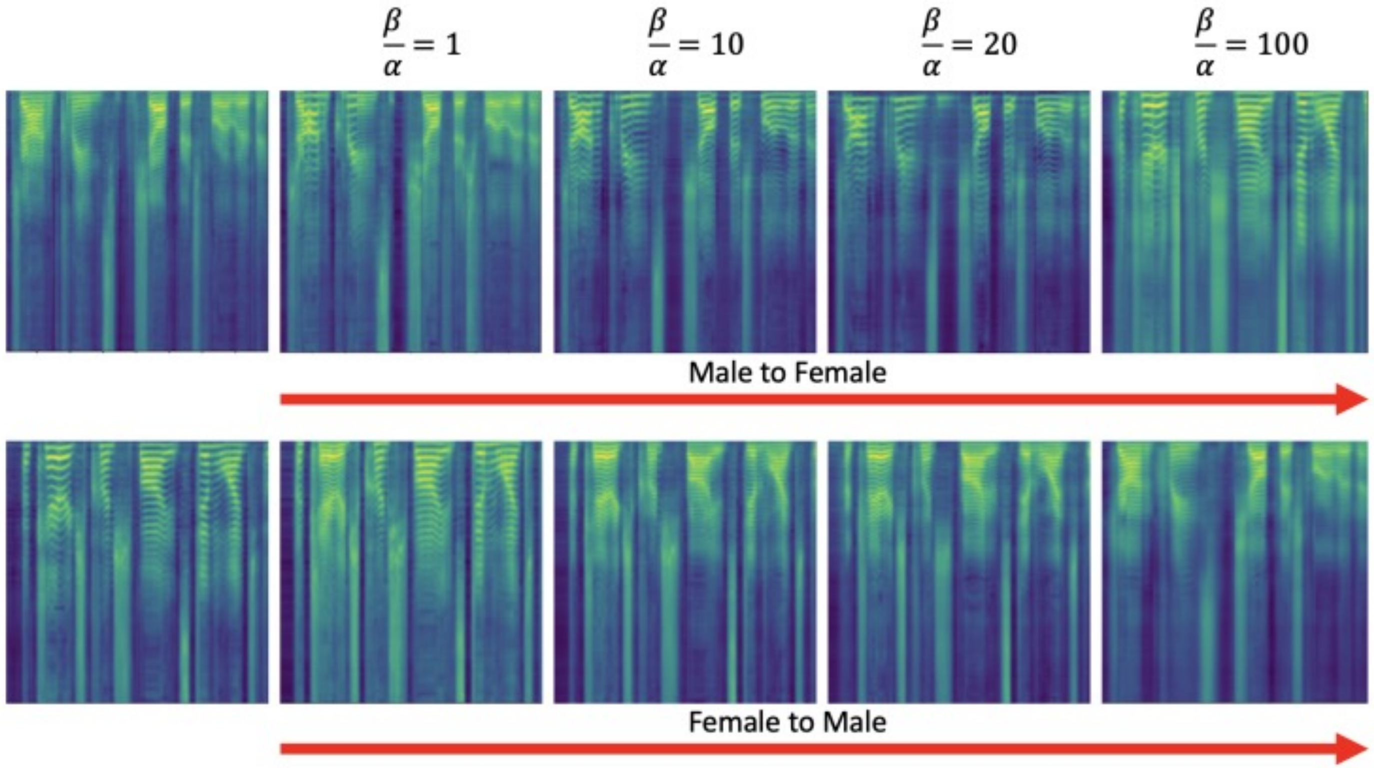
VCTK only: 6 speakers (3 females and 3 males), one utterance per speaker.

Results - Disentanglement

Table 1. EER (%) for TIMIT test trials on varying $\frac{\beta}{\alpha}$.

$\frac{\beta}{\alpha}$	1	10	20	100	DSVAE [16]
μ_S	5.40	3.25	4.16	5.01	4.94
μ_C	31.09	38.83	37.16	38.79	17.49





1. Adjusting $\frac{\beta}{\alpha}$ could control disentanglement.
2. Higher EER on content and lower EER on speaker not necessarily conrrespond to better disentanglement.

Loss Objectives:

$$\mathcal{L} = \mathbb{E}_{p(X)}\mathbb{E}_{q_\theta(X|Z_S,Z_C)}[-log(p_\theta(X|Z_S,Z_C))]+$$
$$\mathbb{E}_{p(X)}[\alpha kl(p(Z_S)||q_\theta(Z_S|X)) + \beta kl(p_\theta(Z_C)||q_\theta(Z_C|X))]$$

Results – VC quality and demo

Table 2. The results of the MOS (95% CI) test on different models.

model	seen to seen		unseen to unseen	
	naturalness	similarity	naturalness	similarity
AUTOVC [13]	2.65±0.12	2.86±0.09	2.47±0.10	2.76±0.08
AdaIN-VC [14]	2.98±0.09	3.06±0.07	2.72±0.11	2.96±0.09
Ours	3.40±0.07	3.56±0.06	3.22±0.09	3.54±0.07
Ours(noisy)	3.23±0.09	3.43±0.07	3.12±0.08	3.47±0.08

Vocoder is WaveNet.
HiGi-GAN is also used in the updated demo.

Demo sample:

Source	Target	VC-wavenet	VC-HiFi-GAN

Unconditional Speech Generation!

Complete demo: https://jlian2.github.io/Robust-Voice-Style-Transfer/

Conclusions

Tencent AI Lab

1. DSVAE-VC: Non-parallel, zero-shot, no speaker labels, no pre-trained speaker embeddings

2. Disentanglement is adjustable and controllable

3. State-of-the-art performance on both SV and VC

4. A unified framework that can be beneficial to ASR, TTS, etc.

Tencent AI Lab

Thank You!