ACVAE-VC: Non-Parallel Voice Conversion With Auxiliary Classifier Variational Autoencoder by Kameoka et al.

Master MVA - Audio Signal Analysis, Indexing, and Transformations

Dorian Desblancs

March 15, 2021

Outline

1. Motivation

2. Methodology

3. Results

4. Wrap-up

Motivation

Voice Conversion

- Change one or more aspects of a speech signal while preserving linguistic information
- Applications: speaker-identity modification, speaking assistance, speech enhancement...
- Two types:
    - Parallel voice conversion
    - **Non-Parallel voice conversion**

Variational Auto-Encoder

-

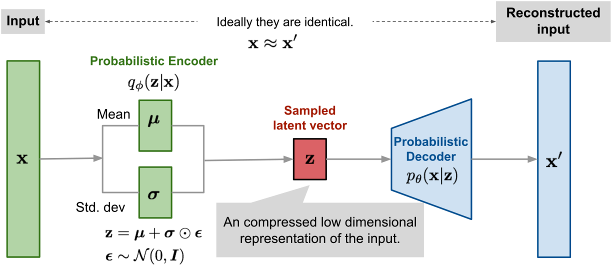

Input
Ideally they are identical.
$\mathbf{x} \approx \mathbf{x}'$
Reconstructed input
Probabilistic Encoder
$q_\phi(\mathbf{z}|\mathbf{x})$
$\mathbf{x}$
Mean
$\boldsymbol{\mu}$
$\boldsymbol{\sigma}$
Std. dev
$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$
$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$
Sampled latent vector
$\mathbf{z}$
An compressed low dimensional representation of the input.
Probabilistic Decoder
$p_\theta(\mathbf{x}|\mathbf{z})$
$\mathbf{x}'$

Illustration of the variational autoencoder model with the multivariate Gaussian assumption (2) (3)

Methodology

Auxiliary Classifier Variational Auto-Encoder

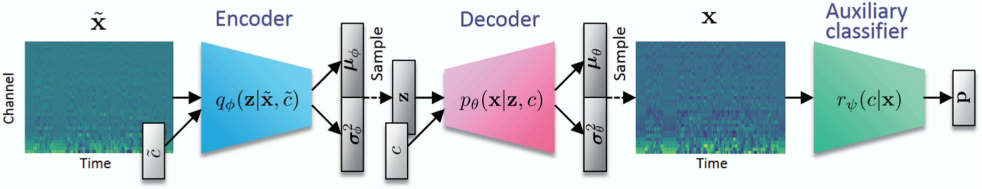
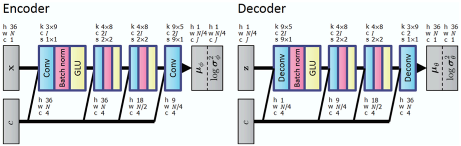
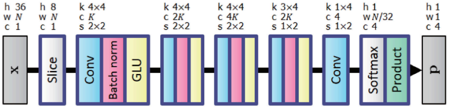


Illustration of the auxiliary class variational autoencoder model (1)

Network Architecture



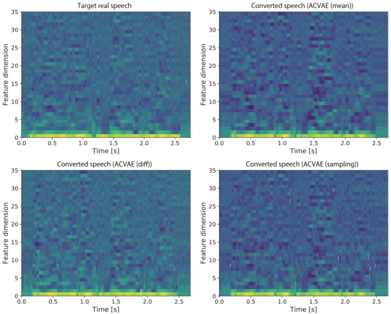
VAE Architecture (1)



Auxiliary Model Architecture (1)

VAE Inputs and Outputs

- Mel-Cepstral Coefficients



Sample Inputs and Outputs (1)

Experimental Design

- Loss function:

$$L(\phi, \theta, \psi) = VAE(\phi, \theta) + \lambda_Q Q(\phi, \theta, \psi) + \lambda_R R(\psi) \tag{1}$$

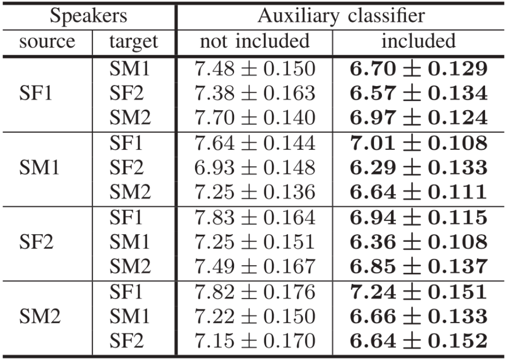
- Performance Measure: mel-cepstral distortion (MCD) along DTW Path

$$MCD[dB](x, y) = \frac{10}{ln10}\sqrt{2\sum_{d=2} D(x_d - y_d)^2} \tag{2}$$

- Data set: Voice Conversion Challenge (VCC) 2018

Results

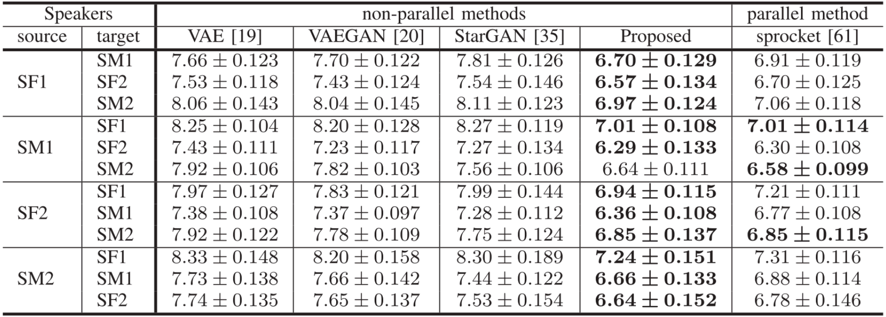
...From Paper (1/2)



Speakers		Auxiliary classifier	
source	target	not included	included
SF1	SM1	7.48 ± 0.150	**6.70 ± 0.129**
	SF2	7.38 ± 0.163	**6.57 ± 0.134**
	SM2	7.70 ± 0.140	**6.97 ± 0.124**
SM1	SF1	7.64 ± 0.144	**7.01 ± 0.108**
	SF2	6.93 ± 0.148	**6.29 ± 0.133**
	SM2	7.25 ± 0.136	**6.64 ± 0.111**
SF2	SF1	7.83 ± 0.164	**6.94 ± 0.115**
	SM1	7.25 ± 0.151	**6.36 ± 0.108**
	SM2	7.49 ± 0.167	**6.85 ± 0.137**
SM2	SF1	7.82 ± 0.176	**7.24 ± 0.151**
	SM1	7.22 ± 0.150	**6.66 ± 0.133**
	SF2	7.15 ± 0.170	**6.64 ± 0.152**

Results obtained with and without auxiliary classifier (1)

...From Paper (2/2)



Speakers		non-parallel methods				parallel method
source	target	VAE [19]	VAEGAN [20]	StarGAN [35]	Proposed	sprocket [61]
SF1	SM1	7.66 ± 0.123	7.70 ± 0.122	7.81 ± 0.126	**6.70 ± 0.129**	6.91 ± 0.119
	SF2	7.53 ± 0.118	7.43 ± 0.124	7.54 ± 0.146	**6.57 ± 0.134**	6.70 ± 0.125
	SM2	8.06 ± 0.143	8.04 ± 0.145	8.11 ± 0.123	**6.97 ± 0.124**	7.06 ± 0.118
SM1	SF1	8.25 ± 0.104	8.20 ± 0.128	8.27 ± 0.119	**7.01 ± 0.108**	**7.01 ± 0.114**
	SF2	7.43 ± 0.111	7.23 ± 0.117	7.27 ± 0.134	**6.29 ± 0.133**	6.30 ± 0.108
	SM2	7.92 ± 0.106	7.82 ± 0.103	7.56 ± 0.106	6.64 ± 0.111	**6.58 ± 0.099**
SF2	SF1	7.97 ± 0.127	7.83 ± 0.121	7.99 ± 0.144	**6.94 ± 0.115**	7.21 ± 0.111
	SM1	7.38 ± 0.108	7.37 ± 0.097	7.28 ± 0.112	**6.36 ± 0.108**	6.77 ± 0.108
	SM2	7.92 ± 0.122	7.78 ± 0.109	7.75 ± 0.124	**6.85 ± 0.137**	**6.85 ± 0.115**
SM2	SF1	8.33 ± 0.148	8.20 ± 0.158	8.30 ± 0.189	**7.24 ± 0.151**	7.31 ± 0.116
	SM1	7.73 ± 0.138	7.66 ± 0.142	7.44 ± 0.122	**6.66 ± 0.133**	6.88 ± 0.114
	SF2	7.74 ± 0.135	7.65 ± 0.137	7.53 ± 0.154	**6.64 ± 0.152**	6.78 ± 0.146

Results compared to other methods (1)

...Generated

All code and results on my GitHub!

Username: d-dawg78, repo: MVA_ASAIT

Wrap-up

Discussion

Advantages:

- Great results
- Clear distinction between outputs depending on class

Improvement:

- Improve data processing and loading for large-scale training feasibility

Future Work

- Auxiliary classifier added to a GAN ?
- Improve pre-processing
- Data augmentation to create more training samples out of small data sets

Thank You

Questions?

Bibliography

1. Kameoka, H., Kaneko, T., Tanaka, K., & Hojo, N. (2019). ACVAE-VC: Non-parallel voice conversion with auxiliary classifier variational autoencoder. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(9), 1432-1443.
2. Weng, L. (2018, August 12). From Autoencoder to Beta-VAE. GitHub. https://lilianweng.github.io/lil-log/2018/08/12/from-autoencoder-to-beta-vae.html#vae-variational-autoencoder
3. Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.