ACVAE-VC : Non-Parallel Voice Conversion With Auxiliary Classifier Variational Autoencoder (2018)

–

Audio Signal Analysis course final project

Kamil Akesbi

Master MVA, France

15/03/2022

Table of content

Paper motivation

Paper Topic

Voice conversion : converting some aspects of a speech signal without changing its linguistic information.

Applications : speaker identity modification - speaking assistance - speech enhancement

Problem : Many VC methods requires parrallel source and target speechs which can be costly to collect and align.

$\Rightarrow$ **Goal :** develop a non-parallel VC method

Existing approaches and drawbacks

- Existing approach to Non-Parallel VC : **Conditional VAE** [Hsu et al. 2016]

$\Rightarrow$ Extended version of VAE where the encoder and decoder take an attribute class c as input.

Drawbacks :

- Don't capture time dependencies
- Over-smoothed outputs
- The encoder and decoder are free to ignore the attribute class label c

Solutions proposed in the paper :

- Fully convolutional architectures to capture time dependencies.
- Auxiliary classifier to ensure that the attribute class information is not lost.

Table of content

- **Goal** : $q_\phi(z|x) \iff p_\theta(z|x)$

- Conditional VAE modification :

$$q_\Phi(z|x,c)$$
$$p_\theta(x|z,c)$$

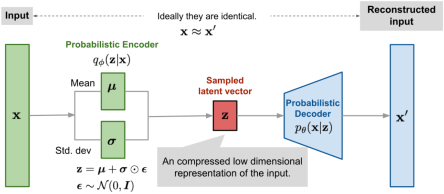

Figure 1: Variational Auto-Encoder

- **Lower bound** maximisation :

$$J(\phi,\theta) = \mathbb{E}_{(x,c)\sim p_d(x,c)}\left[\mathbb{E}_{z\sim q(z|x,c)}\left[\log p(x|z,c)\right]\right] - \mathrm{KL}\left[q(z|x,c)\ ||\ p(z)\right]$$

Table of content

- VAE's Inputs and Outputs :

$\Rightarrow$ Sequence of 36 **Mel-Cepstral coefficients** [Tokuda et al. 1992]

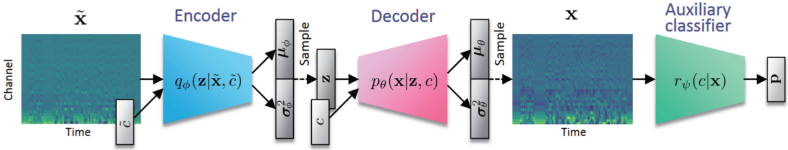

Figure 2: Structure of the proposed ACVAE-VC

About the auxiliary classifier

- The auxiliary classifier helps to maximize the **mutual information** between the output $x \sim p_\theta(x|z,c)$ and class attribute $c|z$.

- It can be seen as a **regularizer**.

- The final training criterion is given by :

$$\mathcal{J}(\phi, \theta) + \lambda_{\mathcal{Q}} \mathcal{Q}(\phi, \theta, \psi) + \lambda_{\mathcal{R}} \mathcal{R}(\psi)$$

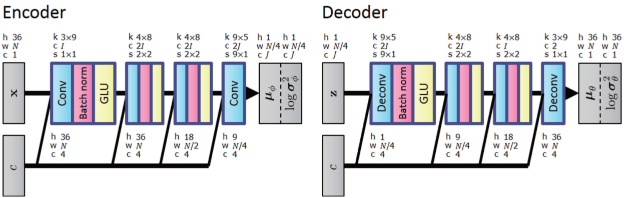

Figure 3: Encoder and Decoder architectures

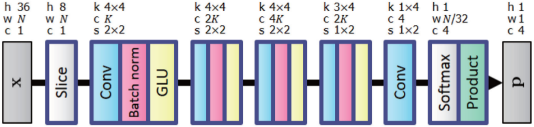

Figure 4: Auxiliary classifier architecture

Table of content

First : Which output consider at test time ?

- Different possibilities :

1. $\hat{x}_{mean} = \mu_\theta\big(\mu_\phi(x, c), \hat{c}\big)$
2. $\hat{x}_{diff} = x - \bar{x}_{mean} + \hat{x}_{mean}$
   where $\bar{x}_{mean} = \mu_\theta\big(\mu_\phi(x, c), c\big)$
3. $\hat{x}_{samp} \sim p_\theta\big(x|\hat{z}, \hat{c}\big)$

Second : Reconstructing the speech signal using the **WORLD** vocoder [Morise et al. 2016]

Table of content

Experimental settings

- **Dataset** : Subset of the Voice Conversion Challenge 2018 dataset

  $\Rightarrow$ Two female speakers, two male speakers

- **Performance measure** : average mel-cepstral distortion (MCDs) along the DTW path :

$$MCD[dB] = \frac{10}{ln10}\sqrt{2\sum_{d=2}^{D}\left(x_d - y_d\right)^2}$$

Speakers		Layer type	
source	target	frame-independent	fully convolutional
SF1	SM1	9.07 ± 0.197	$\mathbf{6.79 \pm 0.088}$
	SF2	8.73 ± 0.121	$\mathbf{6.51 \pm 0.096}$
	SM2	9.25 ± 0.189	$\mathbf{7.05 \pm 0.081}$
SM1	SF1	8.94 ± 0.166	$\mathbf{7.03 \pm 0.090}$
	SF2	8.33 ± 0.214	$\mathbf{6.29 \pm 0.095}$
	SM2	8.68 ± 0.165	$\mathbf{6.67 \pm 0.071}$
SF2	SF1	8.78 ± 0.211	$\mathbf{6.95 \pm 0.104}$
	SM1	8.54 ± 0.198	$\mathbf{6.45 \pm 0.103}$
	SM2	8.75 ± 0.183	$\mathbf{6.87 \pm 0.102}$
SM2	SF1	9.03 ± 0.202	$\mathbf{7.17 \pm 0.098}$
	SM1	8.65 ± 0.182	$\mathbf{6.63 \pm 0.083}$
	SF2	8.43 ± 0.213	$\mathbf{6.58 \pm 0.088}$

Figure 5: Results obtained with or without taking into account time dependency

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

Figure 6: Results obtained with or without the auxiliary classifier

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	$\mathbf{6.70 \pm 0.129}$	6.91 ± 0.119
	SF2	7.53 ± 0.118	7.43 ± 0.124	7.54 ± 0.146	$\mathbf{6.57 \pm 0.134}$	6.70 ± 0.125
	SM2	8.06 ± 0.143	8.04 ± 0.145	8.11 ± 0.123	$\mathbf{6.97 \pm 0.124}$	7.06 ± 0.118
SM1	SF1	8.25 ± 0.104	8.20 ± 0.128	8.27 ± 0.119	$\mathbf{7.01 \pm 0.108}$	$\mathbf{7.01 \pm 0.114}$
	SF2	7.43 ± 0.111	7.23 ± 0.117	7.27 ± 0.134	$\mathbf{6.29 \pm 0.133}$	6.30 ± 0.108
	SM2	7.92 ± 0.106	7.82 ± 0.103	7.56 ± 0.106	6.64 ± 0.111	$\mathbf{6.58 \pm 0.099}$
SF2	SF1	7.97 ± 0.127	7.83 ± 0.121	7.99 ± 0.144	$\mathbf{6.94 \pm 0.115}$	7.21 ± 0.111
	SM1	7.38 ± 0.108	7.37 ± 0.097	7.28 ± 0.112	$\mathbf{6.36 \pm 0.108}$	6.77 ± 0.108
	SM2	7.92 ± 0.122	7.78 ± 0.109	7.75 ± 0.124	$\mathbf{6.85 \pm 0.137}$	$\mathbf{6.85 \pm 0.115}$
SM2	SF1	8.33 ± 0.148	8.20 ± 0.158	8.30 ± 0.189	$\mathbf{7.24 \pm 0.151}$	7.31 ± 0.116
	SM1	7.73 ± 0.138	7.66 ± 0.142	7.44 ± 0.122	$\mathbf{6.66 \pm 0.133}$	6.88 ± 0.114
	SF2	7.74 ± 0.135	7.65 ± 0.137	7.53 ± 0.154	$\mathbf{6.64 \pm 0.152}$	6.78 ± 0.146

Figure 7: Comparison to the conventional non-parallel and parallel methods

- Code adapted from the public repository : `https://github.com/ariacat3366/ACVAE-VC`

- **Main modification** : save data preprocessing step to speed up training :

  $\Rightarrow$ More then 20 hours training to 30 mins

- Some hyper-parameters (sampling frequency, λ_Q, λ_R) + Reconstruction loss computation **differs from the paper**...

  $\longrightarrow$ Still qualitatively gives good results.

- Code and Results on my github : `https://github.com/kamilakesbi/MVA-Voice-Conversion-with-ACVAE`

Table of content

Main contributions of the paper :

This paper proposed a non-parallel VC method using a VAE variant called auxiliary classifier VAE (ACVAE)

- Key ideas :

  $\Rightarrow$ capturing time dependencies.

  $\Rightarrow$ Using information-theoretic regularization with the auxiliary classifier.

Results :

- Better perfomances then conventional methods

Future work

- Incorporating a neural vocoder instead of the WORLD vocoder.
- Testing Transformer approachs rather than the convolutional ones

Hsu, Chin-Cheng et al. (2016). *Voice Conversion from Non-parallel Corpora Using Variational Auto-encoder*. URL: https://arxiv.org/abs/1610.04019.

Morise, M. et al. (2016). *WORLD : A vocoder-based highquality speech synthesis system for real-time applications*. URL: https://www.jstage.jst.go.jp/article/transinf/E99.D/7/E99.D_2015EDP7457/_article.

Tokuda, K. et al. (1992). *An adaptive algorithm for mel-cepstral analysis of speech*. URL: https://www.researchgate.net/publication/3532054_An_adaptive_algorithm_for_mel-cepstral_analysis_of_speech.

Thank you!