Real Non-Volume Preserving Voice Conversion

Santiago Pascual[1], Joan Serrà[2], Antonio Bonafonte[1]

[1]Universitat Politècnica de Catalunya, Barcelona, Spain
[2]Telefónica Research, Barcelona, Spain

Abstract

Voice conversion binds a transformation between two speakers such that the contents uttered by a source speaker are transferred to a target speaking style and identity. Voice conversion is challenging if we have unaligned data to train a mapping model, where source samples do not have labeled conversions to targets. In this work, we present a voice conversion model that deals with unaligned data by leveraging the use of normalizing flows. We show the potential effectiveness and simplicity of this technique, with which we build a non-autoregressive generative model. Once we train it, we can manipulate the latent characteristics of a source speaker to become more like the target one. We thus show preliminary effective conversion results between male and female speakers, using a simple technique to manipulate embeddings in the normalizing flow manifold. This opens new possibilities for speech generative models under the hood of normalizing flows.

Introduction

Voice conversion (VC) systems transform a source speaker recording into another speaker's voice by preserving the information contents (what is said). A typical VC approach is to learn a conversion function bewteen acoustic frames $y_t = f(x_t)$ with aligned linguistic contents in x_t and y_t. Hence, these systems are trained with acoustic pairs $[(x_1, y_1), (x_2, y_2), ..., (x_T, y_T)]$. Previous works used this approach to transform acoustic parameters like mel frequency cepstral coefficients (MFCCs) and/or excitation parameters with linear and non-linear models. A recent trend is to use neural networks of different kinds [1, 12, 14, 11, 9], owing to their success in many supervised tasks where a highly non-linear mapping is required. The problem becomes even more challenging when we face unaligned corpora, where even language between two speakers may differ. There is active research on this problem to go beyond dataset restrictions, either with frequency warping techniques [3] or non-linear neural network systems, such as variational autoencoders [5, 15] or generative adversarial networks [7, 6].

A recent advance in the field of generative models comes with the so called normalizing flows [13]. These allow us to build a composition of bijective funcitons $f : k$, to transform a density distribution x to another one, z, and vice-versa: $x \xleftrightarrow{f_1} h_1 \xleftrightarrow{f_2} h_2 \cdots \xleftrightarrow{f_K} z$. New powerful generative models have been built upon this mechanism in the image generation domain, like the real non-volume preserving (RNVP) system [2] or its follow up work Glow [8]. In this work, we build a voice conversion system following the RNVP structure, thus making, as far as we know, the first step towards using normalizing flows in speech-related tasks and, more concretely, voice conversion. This system works unsupervisedly and without alignment between speaker contents, where x will be our acoustic parameter vectors, and z their latent factors that can enable our manipulation of abstract concepts of the source x like the identity.

Proposed RNVP Voice Conversion

Our initial system works with acoustic features for the sake of simplicity. Hence we make use of the Ahocoder [4] to extract acoustic contents out of waveforms, thus obtaining a set of

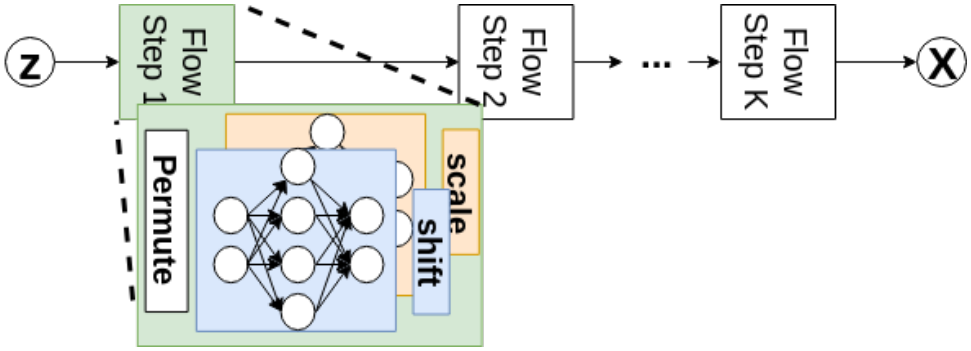

Figure 1: Real non-volume preserving voice conversion system, composed of K normalizing flow steps, each containing a permutation module, a scale network and a shift network.

frames per waveform $x_n^S \in \mathbb{R}^{43}$ (40 MFCCs, 1 voiced-frequency, 1 log-F0, and 1 voiced/unvoiced flag). These source speaker frames x_n^S are transformed into latent space features $z_n^S \in \mathbb{R}^{43}$ with the reverse flow that we will call f, such that $z_n^S = f(x_n^S)$. We will then be able to operate a transformation in the latent space that converts z_n^S into z_n^T, the latent space features of the target speaker, and convert the result back to the original acoustic space with a function $y_n^T = g(z_n^T) = f^{-1}(z_n^T)$, i.e. the forward flow. The proposed generative model is shown in figure , where we can see a composition of many flow steps. Each step contains the permutation block as in [2] to ensure the transformation of all dimensions after all steps. Also, we have a pair of networks in each step that predict the scale and shift, respectively, having half of their inputs masked out. This pair of networks with their masking is called the affine coupling layer. The process to build the flow and be able to make conversions follows the steps:

1. Train the RNVP-VC system with gradient descent. This is achieved projecting x frames from any speaker to z with the reverse flow f, and computing the likelihood loss between our z samples and a prior isotropic unit norm Gaussian distribution. We minimize the negative log-likelihood $\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} -\log p_\theta(x_i)$, being θ the parameters of our model and N the size of the training batches. This is an unsupervised learning task where we learn abstractions about the acoustic data for all the speakers in the dataset, making an independence assumption between acoustic frames.

2. Infer the z values of all training frames from our speakers and compute each speaker's z_{mean} over his/her frames. We assume the resulting z_{mean} contains a dominant activation pattern of certain dimensions that expose identity traits, because content-related features should not be dominant across all frames of each speaker. This suggests that after averaging all frames of a certain speaker, identity should prevail.

3. To convert from speaker S to speaker T, we get the source utterance, infer the latent vector z_n^S of each frame x_n^S, apply the transformation $z_n^T = z_n^S + \alpha(z_{\text{mean}}^T - z_{\text{mean}}^S)$, and forward the resulting representations z^T back to the acoustic space to obtain the converted frames x_T. The α parameter controls how far we go from source towards destination,. These are then transformed into the waveform via vocoder decoding.

Initial Results

We train the system with two speakers from CMU Arctic datset [10]: awb (male) and slt (female). We post some initial conversion results between these speakers online [1], using a model with $K = 6$ steps of flow and neural networks of three layers with sizes $h_1 = 256$, $h_2 = 256$ and $h_2 = 43$ with LeakyReLU activations.

[1] http://veu.talp.cat/rnvpvc

References

[1] Srinivas Desai, E Veera Raghavendra, B Yegnanarayana, Alan W Black, and Kishore Prahallad, *Voice conversion using artificial neural networks*, Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on, IEEE, 2009, pp. 3893–3896.

[2] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio, *Density estimation using Real NVP*, International Conference on Learning Representations (2017).

[3] Daniel Erro, Asunción Moreno, and Antonio Bonafonte, *Inca algorithm for training voice conversion systems from nonparallel corpora*, IEEE Transactions on Audio, Speech, and Language Processing **18** (2010), no. 5, 944–953.

[4] Daniel Erro, Iñaki Sainz, Eva Navas, and Inma Hernáez, *Improved hnm-based vocoder for statistical synthesizers*, Twelfth Annual Conference of the International Speech Communication Association, 2011.

[5] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang, *Voice conversion from non-parallel corpora using variational auto-encoder*, Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific, IEEE, 2016, pp. 1–6.

[6] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo, *Stargan-vc: Non-parallel many-to-many voice conversion with star generative adversarial networks*, arXiv preprint arXiv:1806.02169 (2018).

[7] Takuhiro Kaneko and Hirokazu Kameoka, *Parallel-data-free voice conversion using cycle-consistent adversarial networks*, arXiv preprint arXiv:1711.11293 (2017).

[8] Diederik P Kingma and Prafulla Dhariwal, *Glow: Generative flow with invertible 1x1 convolutions*, arXiv preprint arXiv:1807.03039 (2018).

[9] Kazuhiro Kobayashi, Tomoki Hayashi, Akira Tamamori, and Tomoki Toda, *Statistical voice conversion with wavenet-based waveform generation*, Proc. Interspeech, vol. 2017, 2017, pp. 1138–1142.

[10] John Kominek and Alan W Black, *The cmu arctic speech databases*, Fifth ISCA workshop on speech synthesis, 2004.

[11] Runnan Li, Zhiyong Wu, Helen Meng, and Lianhong Cai, *DBLSTM-based multi-task learning for pitch transformation in voice conversion*, Chinese Spoken Language Processing (ISCSLP), 2016 10th International Symposium on, IEEE, 2016, pp. 1–5.

[12] Seyed Hamidreza Mohammadi and Alexander Kain, *Voice conversion using deep neural networks with speaker-independent pre-training*, Spoken Language Technology Workshop (SLT), 2014 IEEE, IEEE, 2014, pp. 19–23.

[13] Danilo Jimenez Rezende and Shakir Mohamed, *Variational inference with normalizing flows*, Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37, 2015, pp. 1530–1538.

[14] Lifa Sun, Shiyin Kang, Kun Li, and Helen Meng, *Voice conversion using deep bidirectional long short-term memory based recurrent neural networks*, Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, IEEE, 2015, pp. 4869–4873.

[15] Aaron van den Oord, Oriol Vinyals, et al., *Neural discrete representation learning*, Advances in Neural Information Processing Systems, 2017, pp. 6306–6315.