BLOW: A SINGLE-SCALE HYPER-CONDITIONED FLOW FOR NON-PARALLEL RAW-AUDIO VOICE CONVERSION

Joan Serrà, Santiago Pascual, & Carlos Segura

INTRODUCTION

- Generating **raw audio** is hard; it requires of specific treatment.
- Parallel data does not scale; single-task models neither, specially for voice conversion. Need to move towards more unsupervised approaches, **many-to-many** and **non-parallel**.
- Normalizing flows are cool, and we show that they not only can handle these challenging situations, but also that they can outperform the state-of-the-art.

METHODS

Inheriting from Glow [1], but introducing **crucial improvements**:

1. **Single-scale structure** – Removing multi-scale structure yields better likelihoods and improves conversion. Gradient continues to flow thanks to factoring out layer-wise log-determinants.
2. **Many blocks** – We use 8 blocks of 12 flows each. This yields a large receptive field, necessary for raw audio.
3. **Forward-backward conversion** – Forward/backward passes remove/imprint the speaker identity. No operation in latent space (condition-free).
4. **Hyperconditioning** – We condition the coupling layers with a hypernetwork [2]. Traditional conditioning underperformed.
5. **Structure-wise shared embeddings** – Identity is expressed as a learnable embedding vector, shared across all structure and adapted for each hypernetwork.
6. **Data augmentation** – Performed in time domain: (a) temporal jitter, (b) random pre-/de-emphasis, (c) random amplitude scaling, and (d) magnitude flip.

RESULTS

Ablation study – Every introduced improvement is key.

Table 1: Objective scores and their relative difference for possible Blow alternatives (5 min per speaker, 100 epochs).

Configuration	L [nat/dim]	Spoofing [%]
Blow	**4.30**	**66.2**
1: with 3×32 structure	4.01 (− 6.7%)	17.2 (−74.0%)
2: with 3×32 structure (squeeze of 8)	4.21 (− 2.1%)	65.7 (− 0.8%)
3: with multi-scale structure	3.64 (−15.3%)	3.5 (−94.7%)
4: with multi-scale structure (5×19, squeeze of 4)	3.99 (− 7.2%)	16.6 (−74.9%)
5: with additive conditioning (coupling network)	4.28 (− 0.5%)	39.5 (−40.3%)
6: with additive conditioning (before ActNorm)	4.28 (− 0.5%)	22.5 (−66.0%)
7: without data augmentation	4.15 (− 3.5%)	28.3 (−57.2%)

Comparison – Similar or significantly better than the state-of-the-art.

Table 2: Objective and subjective voice conversion scores. For all measures, higher is better. The first two reference rows correspond to using original recordings from source or target speakers as target.

Approach	Objective		Subjective	
	L [nat/dim]	Spoofing [%]	Naturalness [1–5]	Similarity [%]
Source as target	n/a	1.1	4.83	10.6
Target as target	n/a	99.3	4.83	98.5
Glow	4.11	1.2	n/a	n/a
Glow-WaveNet	4.18	3.1	n/a	n/a
StarGAN	n/a	44.4	**2.87**	61.8
VQ-VAE	n/a	65.0	2.42	69.7
Blow	**4.45**	**89.3**	2.83	**77.6**

We improve **normalizing flows** to tackle the task of **voice conversion**.

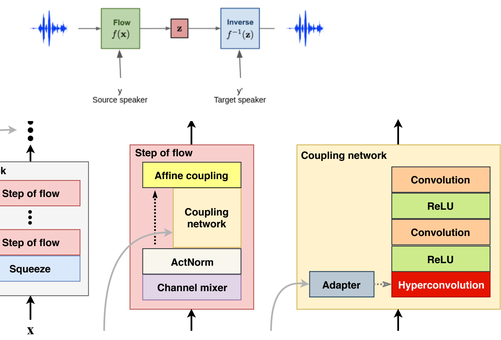
End-to-end, using non-parallel data, many-to-many speakers, and raw audio.

Audio examples, paper, code, slides, this poster, …

FURTHER DETAIL & RESULTS

Loss: likelihood $\log(p(\mathbf{x})) = \log(p(\mathbf{z})) + \sum_{i=1}^{k} \log\left|\det\left(\frac{\partial f_i(\mathbf{h}_{i-1})}{\partial \mathbf{h}_{i-1}}\right)\right|$

Blocks



Evaluation

- Approaches: Glow, Glow-WaveNet, StarGAN, and VQ-VAE.
- Data: VCTK (36 hours train, 108 speakers, avg. 20 min/speaker).
- Measures: objective (likelihood + speaker spoofing) and subjective (as in the voice conversion challenge [3]).

Further results

- Condition-free latent space. Spoofing results: Audio-based = 99.3% (MFCC+Linear) / $\mathbf{z}$-based = 1.8% (RF), 1.4% (MLP) / Chance = 1.1%

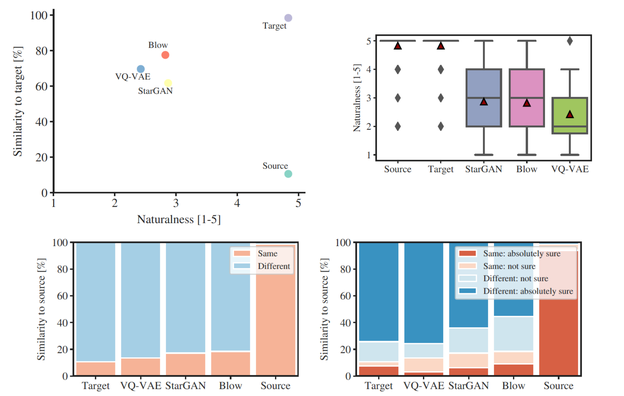


Figure 6: Similarity to the source ratings disregarding confidence (left) and including confidence (right) assessments.

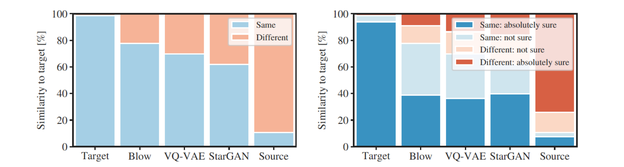


Figure 5: Similarity to the target ratings disregarding confidence (left) and including confidence assessment (right).

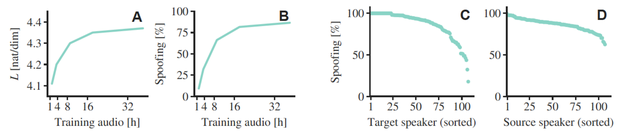


Figure 2: Objective scores with respect to amount of training (A–B) and target/source speaker (C–D).

REFERENCES

[1] D.P. Kingma & P. Dhariwal. *Glow: generative flow with invertible 1x1 convolutions*. In Advances in Neural Information Processing Systems (NeurIPS), volume 31, pages 10215–10224. Curran Associates, Inc., 2018.
[2] D. Ha, A. Dai, & Q.V. Le. *HyperNetworks*. In Proc. of the Int. Conf. on Learning Representations (ICLR), 2017.
[3] M. Wester, Z. Wu, & J. Yamagishi. *Analysis of the voice conversion challenge 2016 evaluation results*. In Proc. of the Int. Speech Communication Association Conf. (INTERSPEECH), pages 1637–1641, 2016.

Telefónica
UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH