BLOW: SINGLE-SCALE HYPERCONDITIONED FLOW FOR NON-PARALLEL RAW-AUDIO VOICE CONVERSION

Joan Serrà[1], Santiago Pascual[2], & Carlos Segura[1]
@serrjoa @santty128 @cseguraperales

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

Preliminary presentation, September 2019

OUTLINE

- Introduction
  - Flow-based models
  - Glow
- Blow
- Experimental setup
- Results
- Conclusion

INTRODUCTION

INTRODUCTION

VOICE CONVERSION

- One to one models
- Parallel data
- Intermediate representation

INTRODUCTION

VOICE CONVERSION

- One to one models
  - Not exploiting other identities
- Parallel data
  - Non-scalable, problematic collection
- Intermediate representation
  - Raw audio is a challenge!

INTRODUCTION

VOICE CONVERSION

- One to one models
  - Not exploiting other identities
- Parallel data
  - Non-scalable, problematic collection
- Intermediate representation
  - Raw audio is a challenge!

Many-to-many + Non-parallel + Raw audio

INTRODUCTION

GENERATION OF RAW AUDIO

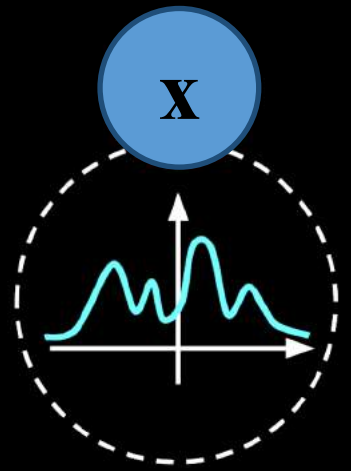
- Autoregressive models: WaveNet, SampleRNN, WaveRNN, …
- Generative adversarial networks: SEGAN, WaveGAN, …

INTRODUCTION

GENERATION OF RAW AUDIO

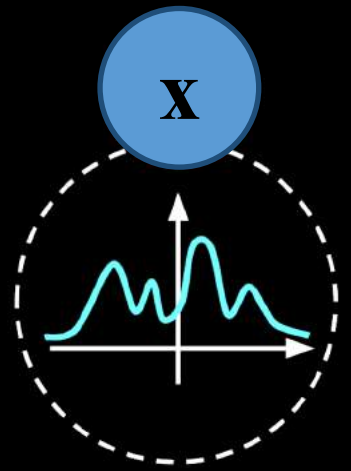
- Autoregressive models: WaveNet, SampleRNN, WaveRNN, …
- Generative adversarial networks: SEGAN, WaveGAN, …
- Flow-based models (normalizing flows): WaveGlow

INTRODUCTION
FLOW-BASED MODELS

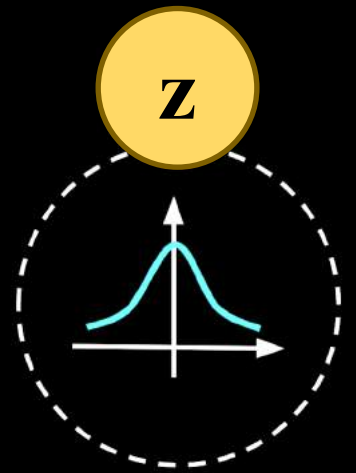


$$\mathbf{x} \sim p(\mathbf{x})$$

INTRODUCTION
FLOW-BASED MODELS

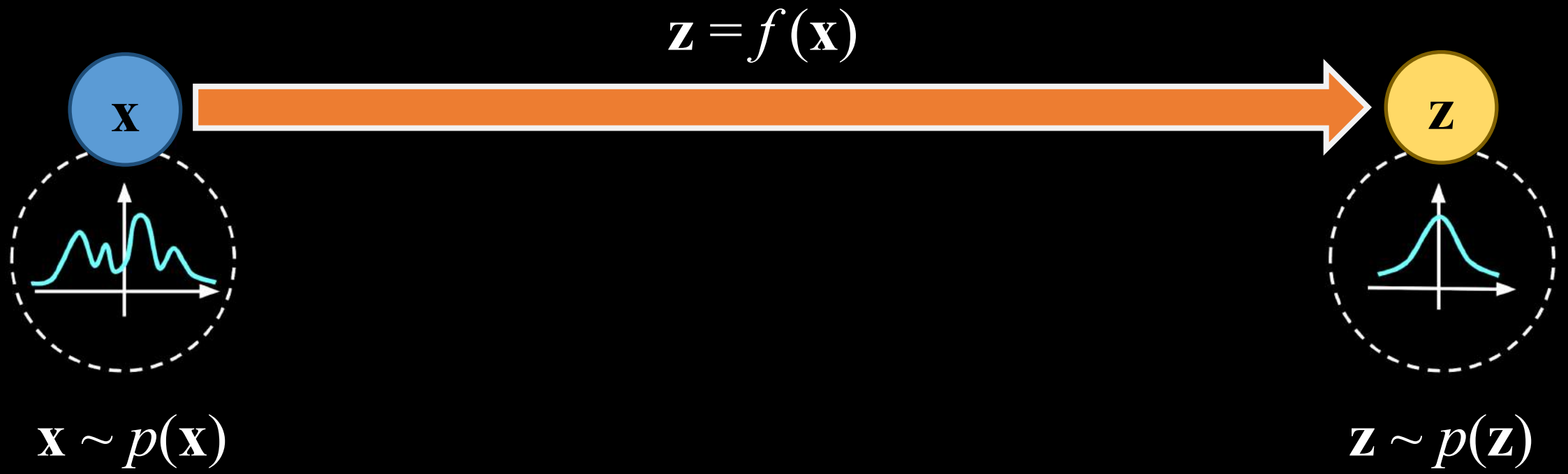


$\mathbf{x} \sim p(\mathbf{x})$



$\mathbf{z} \sim p(\mathbf{z})$

INTRODUCTION
FLOW-BASED MODELS



INTRODUCTION
FLOW-BASED MODELS

$$\mathbf{z} = f(\mathbf{x}), \quad \mathbf{x} = f^{-1}(\mathbf{z})$$

$\mathbf{x} \sim p(\mathbf{x})$

$\mathbf{z} \sim p(\mathbf{z})$

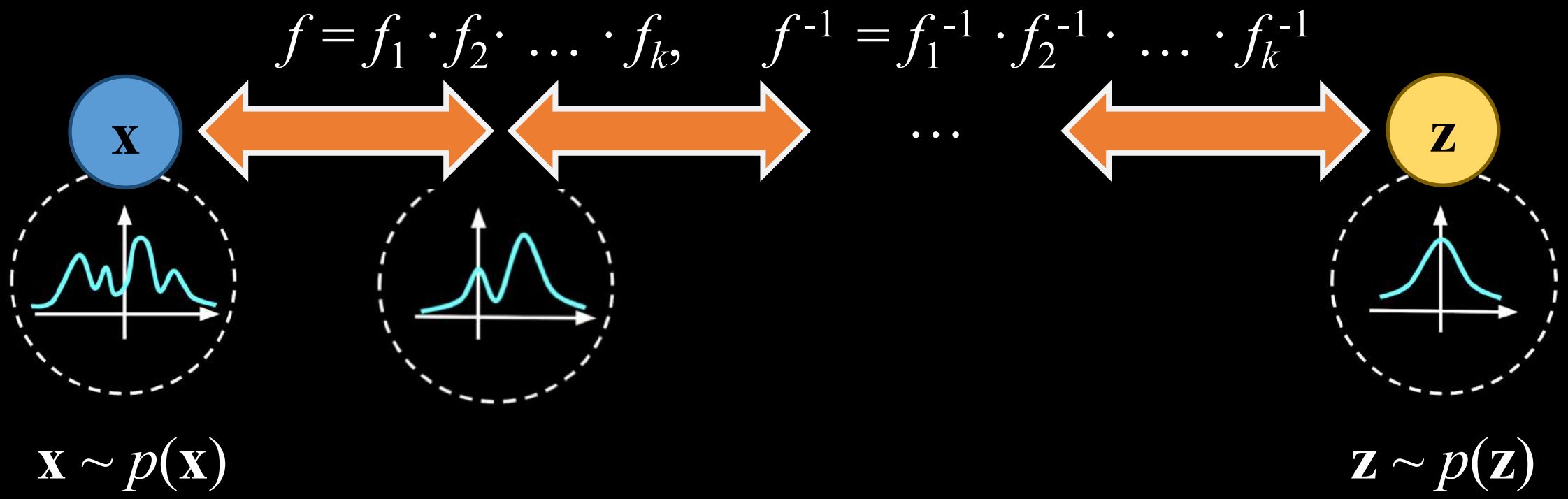
INTRODUCTION
FLOW-BASED MODELS

$$\mathbf{z} = f(\mathbf{x}), \quad \mathbf{x} = f^{-1}(\mathbf{z})$$

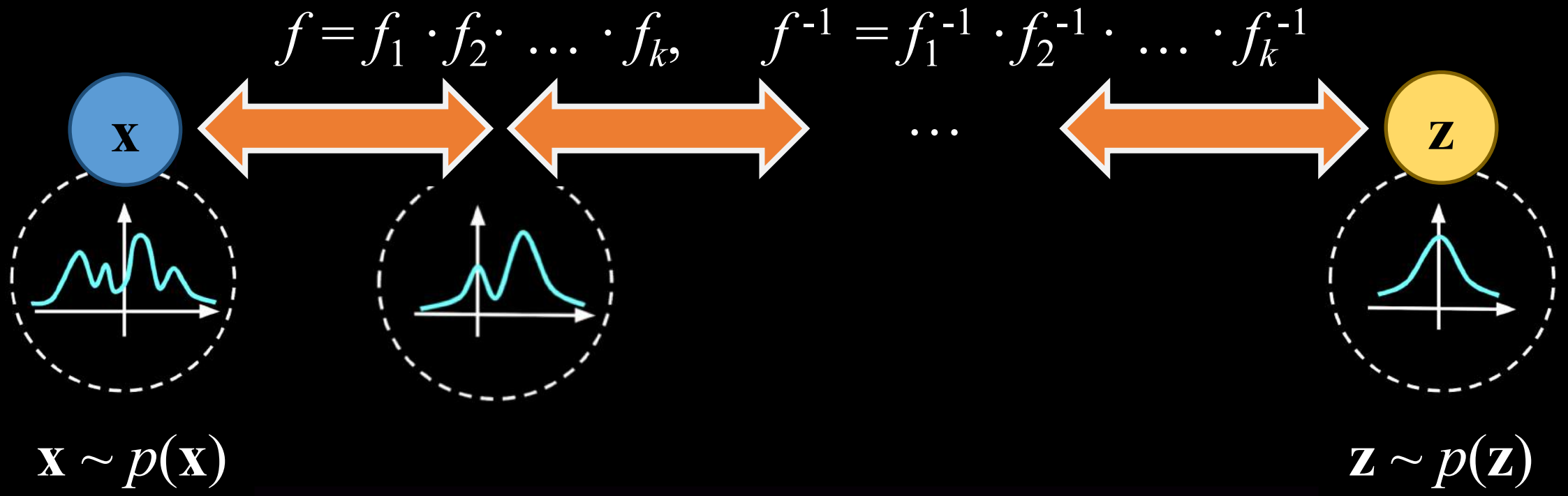
$\mathbf{x} \sim p(\mathbf{x})$ $\mathbf{z} \sim p(\mathbf{z})$

$$p(\mathbf{z}) = p(\mathbf{x}) \left| \det \left(\frac{\partial f(\mathbf{z})}{\partial \mathbf{z}} \right) \right|^{-1}$$

INTRODUCTION
FLOW-BASED MODELS



INTRODUCTION
FLOW-BASED MODELS



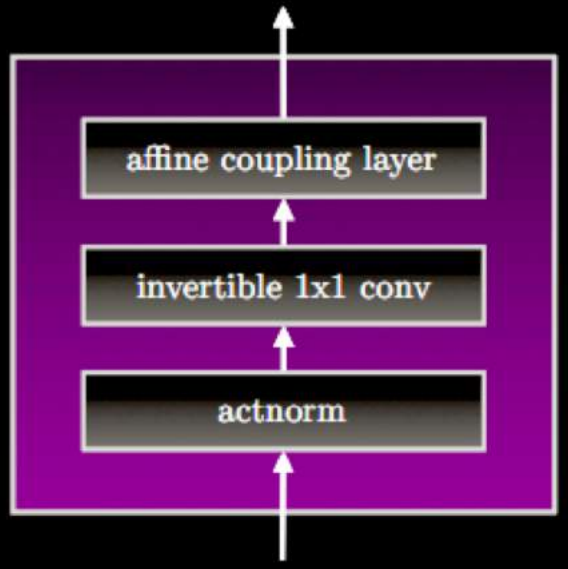
$$\log\left(p\left(\mathbf{x}\right)\right) = \log\left(p\left(\mathbf{z}\right)\right) + \sum_{i=1}^{k} \log \left| \det \left(\frac{\partial f_i(\mathbf{h}_{i-1})}{\partial \mathbf{h}_{i-1}} \right) \right|$$

PROPERTIES OF FLOW-BASED MODELS

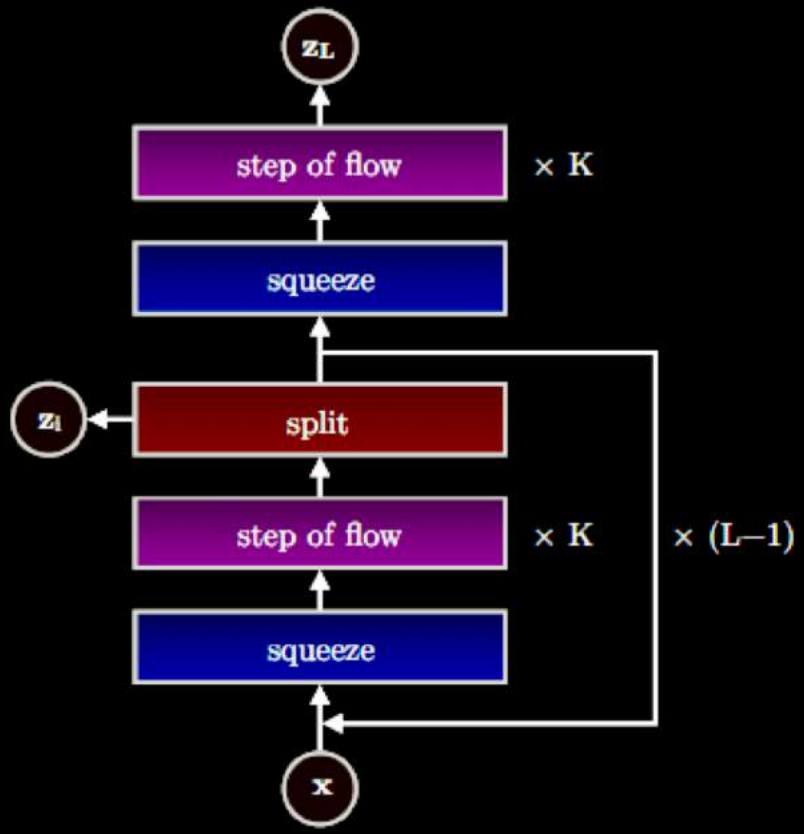
- Standard training (with validation data)
- Exact latent variable inference
- “Meaningful” loss (likelihood)
- Exact likelihood evaluation
- Efficient inference/synthesis
- Useful latent space
- …

INTRODUCTION

POPULAR FLOW-BASED MODEL: GLOW



(a) One step of our flow.



(b) Multi-scale architecture (Dinh et al., 2016).

INTRODUCTION

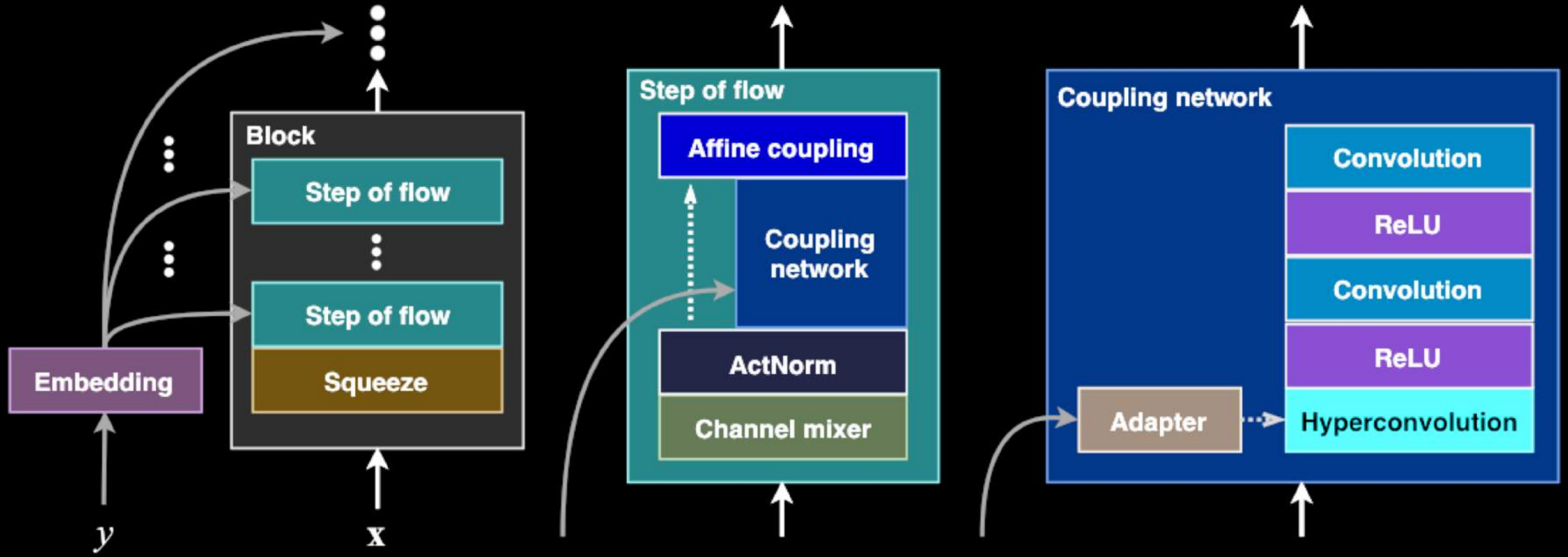
POPULAR FLOW-BASED MODEL: GLOW

Description	Function	Reverse Function	Log-determinant
Actnorm. See Section 3.1.	$\forall i,j : \mathbf{y}_{i,j} = \mathbf{s} \odot \mathbf{x}_{i,j} + \mathbf{b}$	$\forall i,j : \mathbf{x}_{i,j} = (\mathbf{y}_{i,j} - \mathbf{b})/\mathbf{s}$	$h \cdot w \cdot \mathtt{sum}(\log \lvert\mathbf{s}\rvert)$
Invertible 1×1 convolution. $\mathbf{W} : [c \times c]$. See Section 3.2.	$\forall i,j : \mathbf{y}_{i,j} = \mathbf{W}\mathbf{x}_{i,j}$	$\forall i,j : \mathbf{x}_{i,j} = \mathbf{W}^{-1}\mathbf{y}_{i,j}$	$h \cdot w \cdot \log \lvert\det(\mathbf{W})\rvert$ or $h \cdot w \cdot \mathtt{sum}(\log \lvert\mathbf{s}\rvert)$ (see eq. (10))
Affine coupling layer. See Section 3.3 and (Dinh et al., 2014)	$\mathbf{x}_a, \mathbf{x}_b = \mathtt{split}(\mathbf{x})$ $(\log \mathbf{s}, \mathbf{t}) = \mathtt{NN}(\mathbf{x}_b)$ $\mathbf{s} = \exp(\log \mathbf{s})$ $\mathbf{y}_a = \mathbf{s} \odot \mathbf{x}_a + \mathbf{t}$ $\mathbf{y}_b = \mathbf{x}_b$ $\mathbf{y} = \mathtt{concat}(\mathbf{y}_a, \mathbf{y}_b)$	$\mathbf{y}_a, \mathbf{y}_b = \mathtt{split}(\mathbf{y})$ $(\log \mathbf{s}, \mathbf{t}) = \mathtt{NN}(\mathbf{y}_b)$ $\mathbf{s} = \exp(\log \mathbf{s})$ $\mathbf{x}_a = (\mathbf{y}_a - \mathbf{t})/\mathbf{s}$ $\mathbf{x}_b = \mathbf{y}_b$ $\mathbf{x} = \mathtt{concat}(\mathbf{x}_a, \mathbf{x}_b)$	$\mathtt{sum}(\log(\lvert\mathbf{s}\rvert))$

BLOW

BLOW

FLOW-BASED MODEL



BLOW

CRUCIAL CHANGES

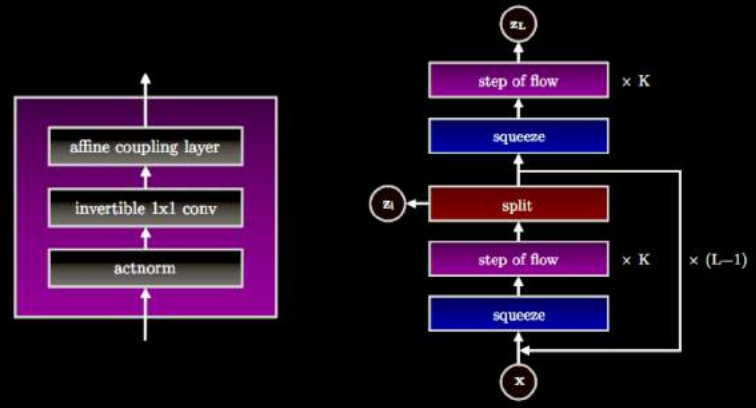
1. Single-scale structure
2. Many blocks
3. Forward-backward conversion
4. Hyperconditioning
5. Structure-wise shared embeddings
6. Data augmentation

BLOW

SINGLE-SCALE STRUCTURE

Multi-scale structure:

- Intermediary levels of representation
- Encourage gradient flow; facilitate training



(a) One step of our flow.

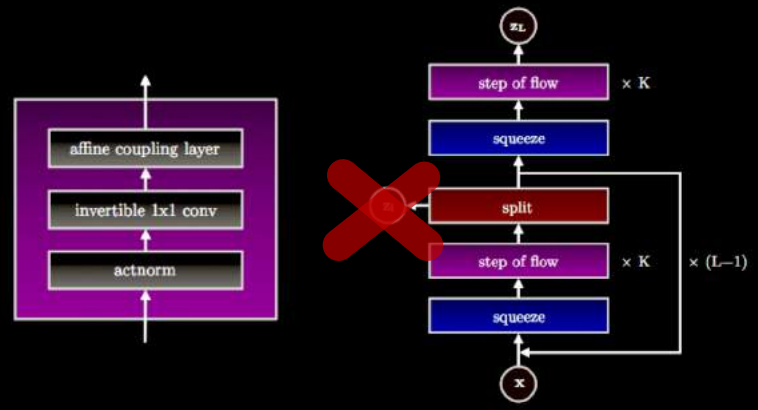
(b) Multi-scale architecture (Dinh et al., 2016).

BLOW

SINGLE-SCALE STRUCTURE

Multi-scale structure:

- Intermediary levels of representation
- Encourage gradient flow; facilitate training



(a) One step of our flow.

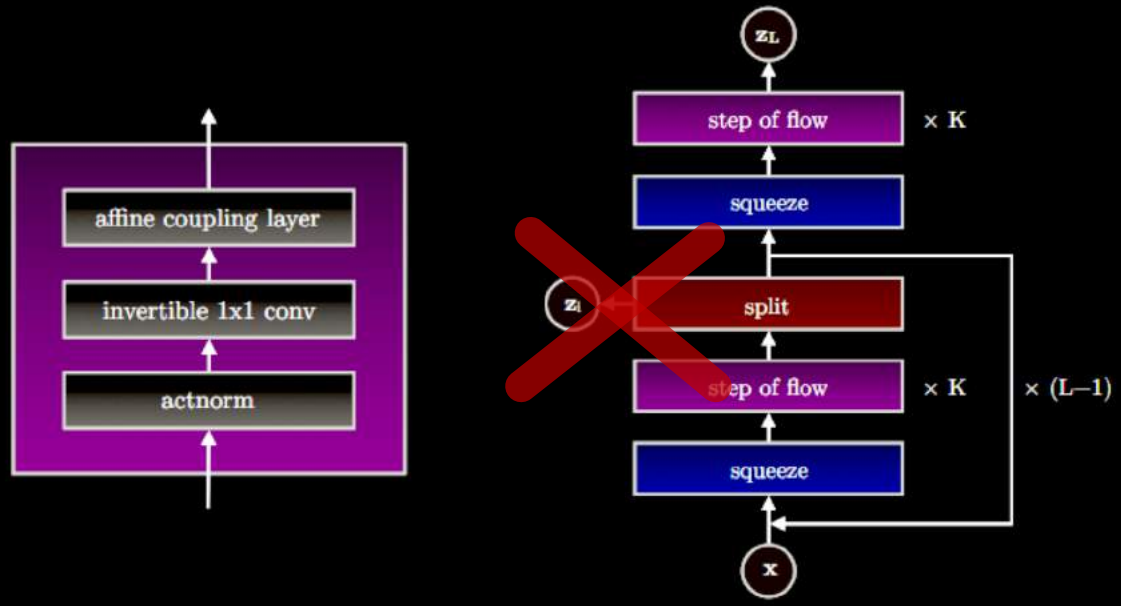
(b) Multi-scale architecture (Dinh et al., 2016).

However:

- Last level had the strongest presence of identity information
- Single-scale achieves better likelihoods
- Single-scale doesn't affect gradient flow

BLOW

SINGLE-SCALE STRUCTURE



(a) One step of our flow.

(b) Multi-scale architecture (Dinh et al., 2016).

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

BLOW
MANY BLOCKS

Images: at most 256x256 pixels

256 samples @ 16 kHz = 16 ms

Phoneme duration: between 50 and 180 ms

+ Phoneme transitions!

Need at least 3,500 samples!

BLOW
MANY BLOCKS

Images: at most 256x256 pixels

256 samples @ 16 kHz = 16 ms

Phoneme duration: between 50 and 180 ms

+ Phoneme transitions!

Need at least 3,500 samples!

How to adapt the receptive field?

More aggressive squeezing, more layers, …

More blocks with less layers (8x12) and same squeezing

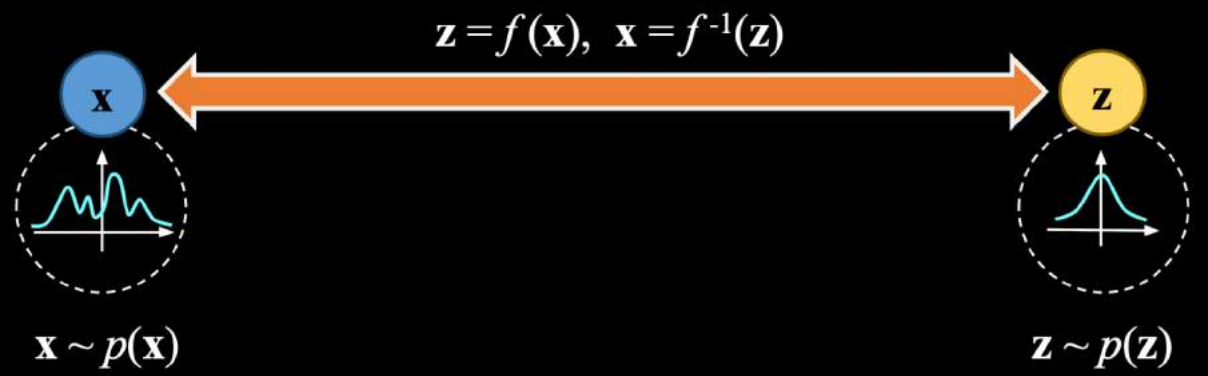
12.5k samples @ 16 kHz = 781 ms

BLOW

FORWARD-BACKWARD CONVERSION

Default strategy: Operate in the latent space ($\mathbf{z}$)

- Progressive changes/interpolations
- Few/zero-shot learning



BLOW

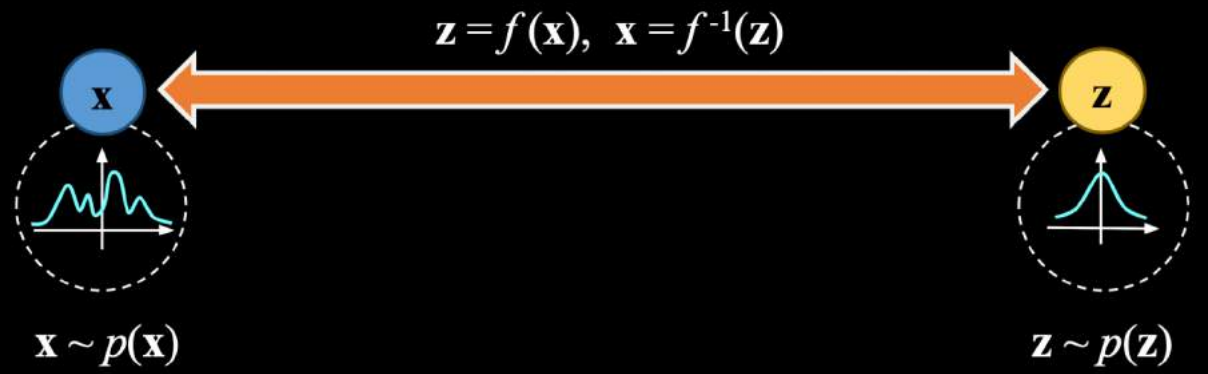
FORWARD-BACKWARD CONVERSION

Default strategy: Operate in the latent space ($\mathbf{z}$)

- Progressive changes/interpolations
- Few/zero-shot learning

However:

- Noisy
- Not powerful enough



BLOW

FORWARD-BACKWARD CONVERSION

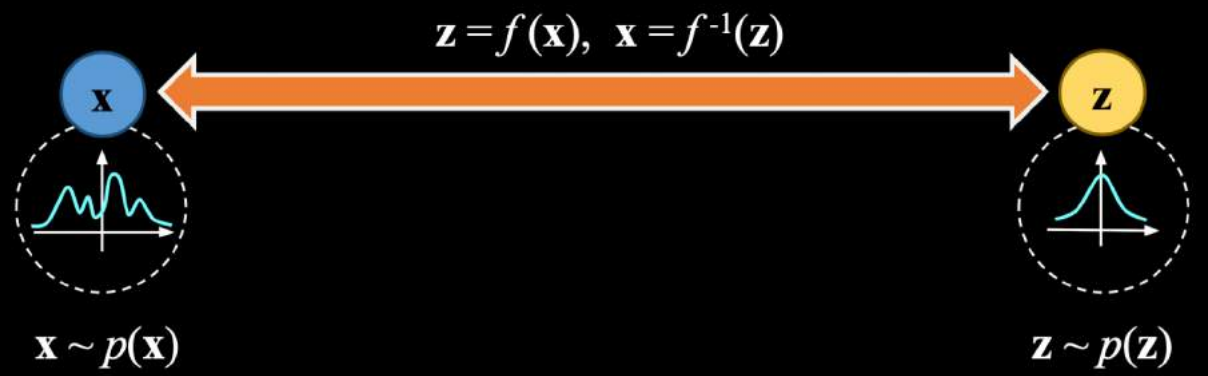
Default strategy: Operate in the latent space ($\mathbf{z}$)

- Progressive changes/interpolations
- Few/zero-shot learning

However:

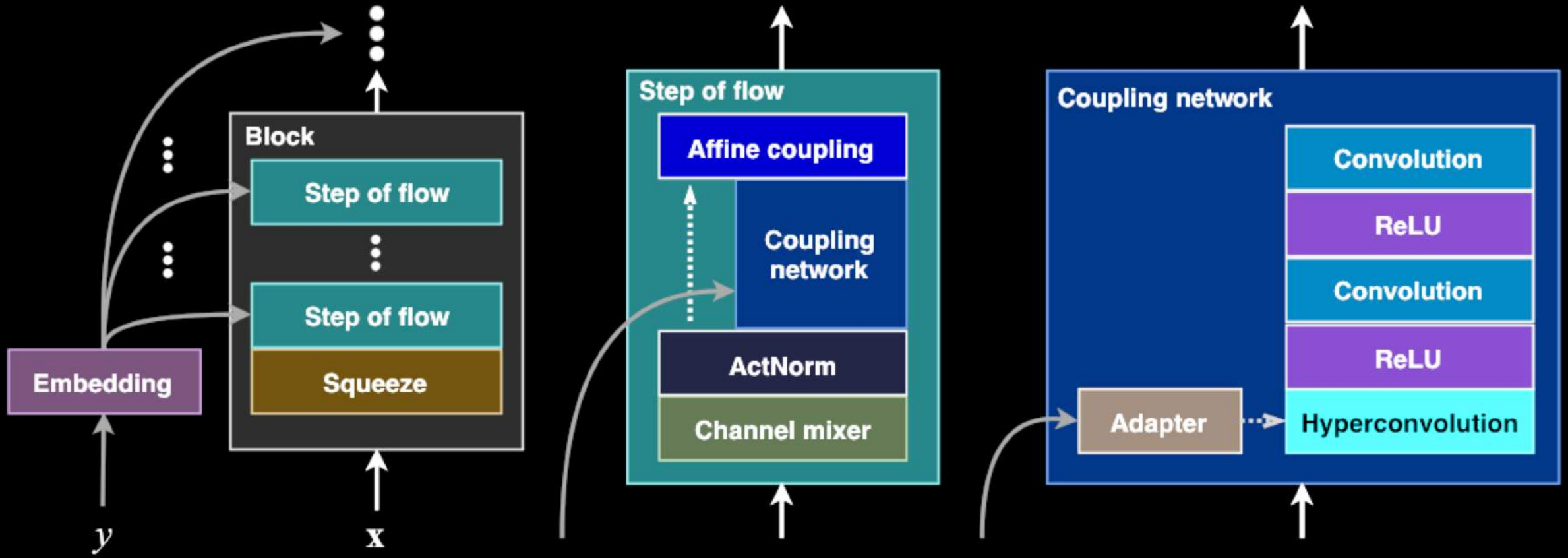
- Noisy
- Not powerful enough

Identity-neutral $\mathbf{z}$!



BLOW

FORWARD-BACKWARD CONVERSION



BLOW

HYPERCONDITIONING

Straightforward strategy: Conditioning in the coupling network

- No invertibility constrains
- Where most of the transformation takes place

BLOW

HYPERCONDITIONING

Straightforward strategy: Conditioning in the coupling network

- No invertibility constrains
- Where most of the transformation takes place

However:

- Concatenation was ignored
- Addition not powerful enough

BLOW

HYPERCONDITIONING

Straightforward strategy: Conditioning in the coupling network

- No invertibility constrains
- Where most of the transformation takes place

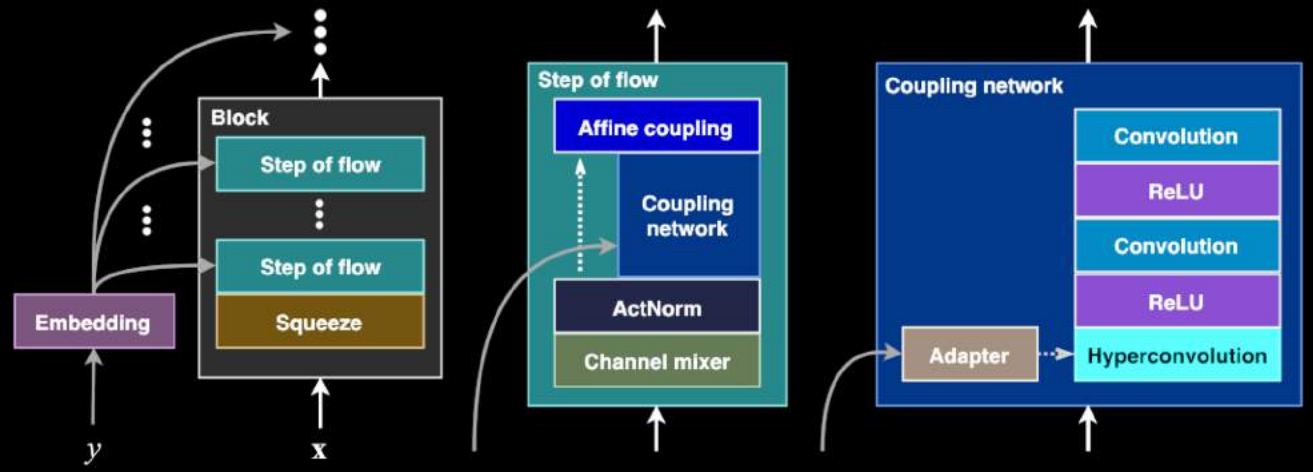
However:

- Concatenation was ignored
- Addition not powerful enough
- Let's condition the weights of the convolutions (hyper nets)

BLOW
HYPERCONDITIONING

$$\mathbf{h}^{(i)} = \mathbf{W}_y^{(i)} * \mathbf{H} + b_y^{(i)}$$

$$\mathcal{K}_y = \left\{ \left(\mathbf{W}_y^{(1)}, b_y^{(1)} \right) \dots \left(\mathbf{W}_y^{(n)}, b_y^{(n)} \right) \right\} = g\left(\mathbf{e}_y\right)$$

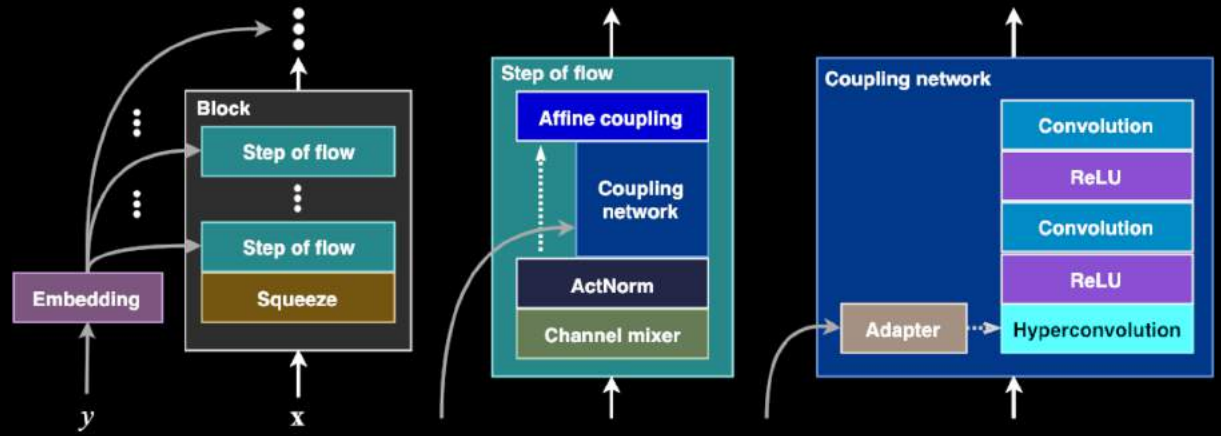


BLOW

STRUCTURE-WISE SHARED EMBEDDINGS

Straightforward strategy:

One independent embedding per coupling network

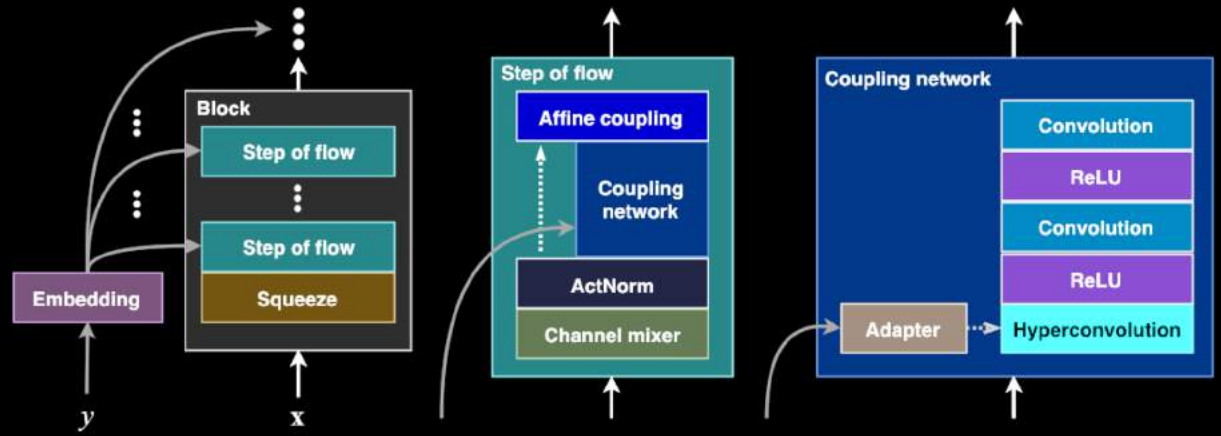


BLOW

STRUCTURE-WISE SHARED EMBEDDINGS

Straightforward strategy:

One independent embedding per coupling network



However:

- Independent conditioning not focusing on the essence (the speaker identity) … Too much freedom?
- Inspired by StyleGAN, we employ a single embedding vector for all layers

BLOW

DATA AUGMENTATION

Time domain augmentations:

1. Temporal jitter (max. ±1 frame)
2. Random pre-/de-emphasis (max. ±0.25)
3. Random amplitude scaling
4. Random flip

EXPERIMENTAL SETUP

EXPERIMENTAL SETUP

EVALUATION

- Data: VCTK
- Compared approaches:
  - Flow-based: Glow + WaveGlow
  - State-of-the-art: StarGAN-VC + VQ-VAE
- Scores:
  - Objective: Likelihood + Spoofing
  - Subjective (VC-Challenge): Naturalness + Similarity

RESULTS

RESULTS

ABLATION STUDY

RESULTS

ABLATION STUDY

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%)

RESULTS

ABLATION STUDY

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%)

RESULTS

ABLATION STUDY

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%)

RESULTS

ABLATION STUDY

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%)

RESULTS

ABLATION STUDY

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%)

RESULTS

OBJECTIVE AND SUBJECTIVE SCORES

RESULTS

OBJECTIVE AND SUBJECTIVE SCORES

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**

RESULTS

OBJECTIVE AND SUBJECTIVE SCORES

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**

RESULTS

OBJECTIVE AND SUBJECTIVE SCORES

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**

RESULTS

OBJECTIVE AND SUBJECTIVE SCORES

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**

RESULTS

AMOUNT OF DATA + TARGET/SOURCE PREF.

RESULTS

AMOUNT OF DATA + TARGET/SOURCE PREF.

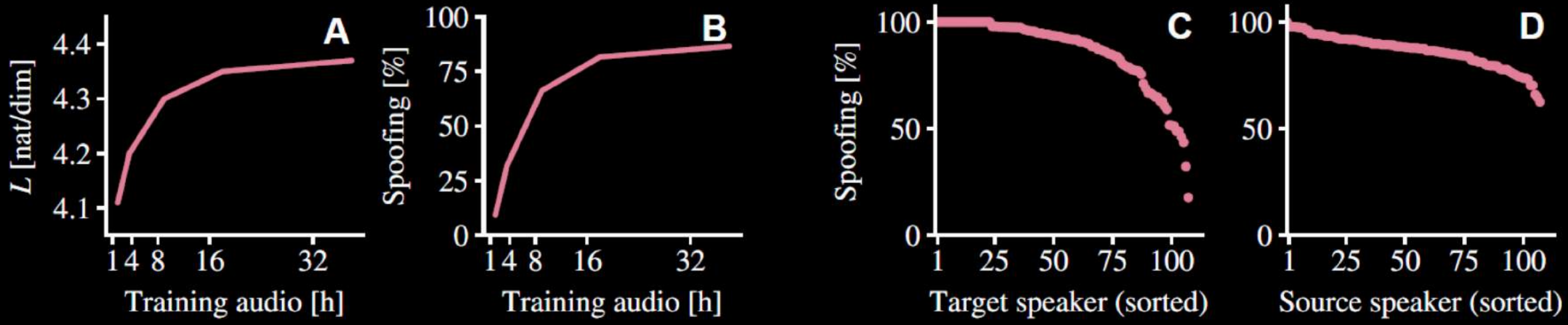


Figure 2: Objective scores with respect to amount of training (A–B) and target/source speaker (C–D).

RESULTS

CONVERSION EXAMPLES

- Source
- Target
- Conversion

- Source
- Target
- Conversion

RESULTS

CONVERSION EXAMPLES

Source

Target

Conversion

Source

Target

Conversion

More examples: https://blowconversions.github.io

PyTorch code: https://github.com/joansj/blow

Paper: https://arxiv.org/abs/1906.00794

CONCLUSION

CONCLUSION

SUMMARY

- Second flow-based generative model for raw audio synthesis

CONCLUSION

SUMMARY

- Second flow-based generative model for raw audio synthesis
- Contributions beyond standard practice:
  1. Single-scale structure
  2. Many blocks
  3. Forward-backward conversion
  4. Hyperconditioning
  5. Structure-wise shared embeddings
  6. Data augmentation

CONCLUSION

SUMMARY

- Second flow-based generative model for raw audio synthesis
- Contributions beyond standard practice:
  1. Single-scale structure
  2. Many blocks
  3. Forward-backward conversion
  4. Hyperconditioning
  5. Structure-wise shared embeddings
  6. Data augmentation
- Very promising results in non-parallel, raw-audio voice conversion