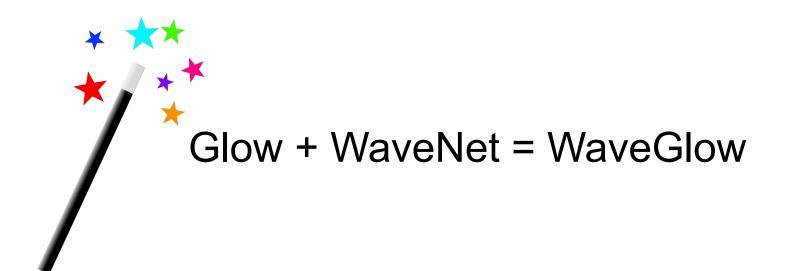
# **WAVEGLOW: A** FLOW-BASED GENERATIVE NETWORK FOR SPEECH **SYNTHESIS**

Ryan Prenger, Rafael Valle, Bryan Catanzaro

**NVIDIA** Corporation

Reviewer: Mark Baushenko







### **WAVEGLOW**

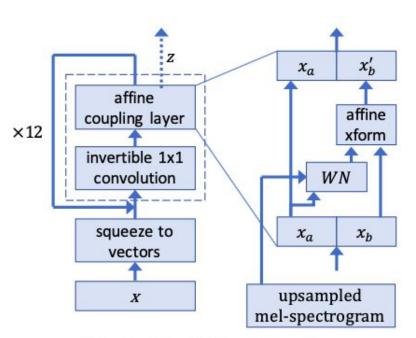


Fig. 1: WaveGlow network

## **Affine Coupling Layer**

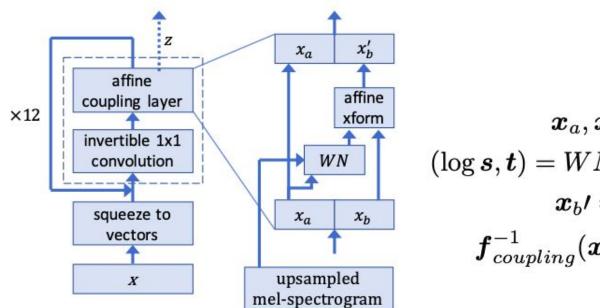
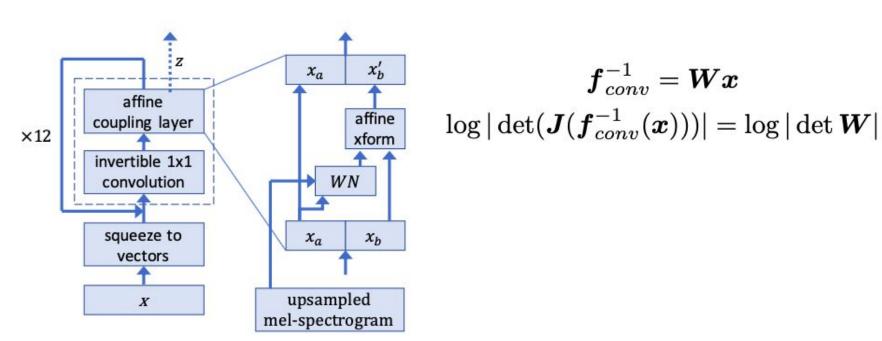


Fig. 1: WaveGlow network

$$egin{aligned} oldsymbol{x}_a, oldsymbol{x}_b &= split(oldsymbol{x}) \ (\log oldsymbol{s}, oldsymbol{t}) &= WN(oldsymbol{x}_a, mel\text{-}spectrogram) \ oldsymbol{x}_b\prime &= oldsymbol{s}\odot oldsymbol{x}_b + oldsymbol{t} \ oldsymbol{f}_{coupling}^{-1}(oldsymbol{x}) &= concat(oldsymbol{x}_a, oldsymbol{x}_b\prime) \end{aligned}$$

#### 1x1 Invertible Convolution



#### 1x1 Invertible Convolution

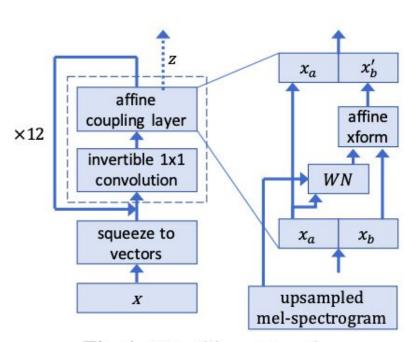


Fig. 1: WaveGlow network

$$egin{aligned} oldsymbol{f}_{conv}^{-1} &= oldsymbol{W} oldsymbol{x} \ \log |\det oldsymbol{J}(oldsymbol{f}_{conv}^{-1}(oldsymbol{x}))| = \log |\det oldsymbol{W}| \ \log p_{ heta}(oldsymbol{x}) &= -rac{oldsymbol{z}(oldsymbol{x})^T oldsymbol{z}(oldsymbol{x})}{2\sigma^2} \ &+ \sum_{j=0}^{\# coupling} \log oldsymbol{s}_j(oldsymbol{x}, \textit{mel-spectrogram}) \ &+ \sum_{k=0}^{\# conv} \log \det |oldsymbol{W}_k| \end{aligned}$$

# **Audio quality comparison**

Model	<b>Mean Opinion Score (MOS)</b>
Griffin-Lim	$3.823 \pm 0.1349$
WaveNet	$3.885 \pm 0.1238$
WaveGlow	$3.961 \pm 0.1343$
<b>Ground Truth</b>	$4.274 \pm 0.1340$

# Thanks for attention