	blocks	flows	layers	kernel size	width
WaveNet (toy)	3	-	10	3	256
WaveNet (musdb18)	3	-	10	3	256
FloWaveNet (toy)	4	6	10	3	32
FloWaveNet (musdb18)	8	6	10	3	48

Table 2: The hyperparameters for the FloWaveNet and WaveNet models. In case of the WaveNet blocks refers to the blocks as described in the original WaveNet architecture [14] while in the FloWaveNet the layers refer to the layers of the WaveNet in the coupling layers.

5 Appendix

5.1 Source separation with SGLD

For better understanding of the source separation approach we had in mind using the generative models as prior we give the implementation in Algorithm 1.

Algorithm 1 The Langevin sampling procedure for source separation is fairly straight forward. For a fixed number of steps T we sample we take a step into the direction of the gradient under the priors and the gradient of the mixing constraint while adding Gaussian noise ϵ_t.

```
for $t = 1 \dots T$ do
    for $k = 1 \dots N$ do
        $\epsilon_t \sim \mathcal{N}(0, \mathbf{1})$
        $\Delta \boldsymbol{s}_k^t \leftarrow \boldsymbol{s}^t + \eta \cdot \nabla \log p(\boldsymbol{s}^t) + 2\sqrt{\eta}\epsilon_t$
    end for
    for $k = 1 \dots N$ do
        $\boldsymbol{s}_k^{t+1} \leftarrow \Delta \boldsymbol{s}_k^t - \frac{\eta}{\sigma^2} \cdot [\boldsymbol{m} - \frac{1}{N} \sum_i^N \boldsymbol{s}_i^t]$
    end for
end for
```

5.2 Model and training details

We construct the flow models closely following the architecture of FloWaveNet [5] which we show in Figure 5. It combines the affine coupling layer proposed in RealNVP [1] with the Activation Normalization proposed in Glow [7] but does not learn the channel mixing function as in Glow and apply the fixed checkerboard masking over the channel dimension.

The WaveNets are constructed as described in the original WaveNet work [14]. As in the original work the outputs of the model at each time-point are modeled with a multinomial distribution with a size of 256 and therefore uses a cross-entropy loss for optimization. The quantization of the wave data is done with standard μ-law encoding.

The hyperparameters for all for model architectures are listed in Table 2.

The models are trained with the Adam optimizer [6]. As all models are fully convolutional the input size is in no way regimented by the architecture, only in so far that we are avoiding padding in the lower layers nevertheless we fix the size of all frames to $2^{14} = 16384$. The initial learning rate is set to $1e-4$ and decreased with $\gamma = 0.6$ in a fixed five-step decrease schedule. The toy model is trained with a batch size of 5 and the musdb18 model with a batch size of 2. We train the two unconditional flows and the WaveNets are trained for each 150.000 steps. The fine-tuning with the added noise is each trained until convergence which in practice was achieved in 20.000 to 40.000 steps.

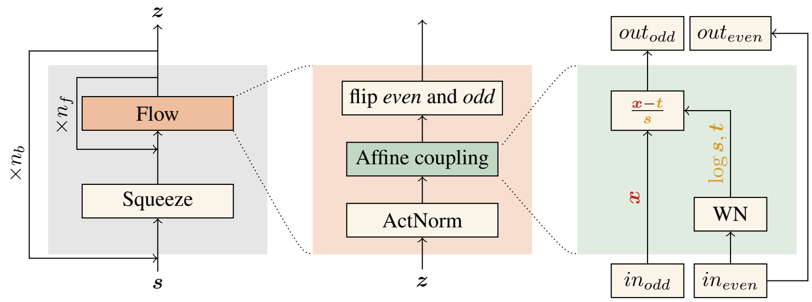

Figure 5: The building blocks for the FloWaveNet model. The model consists of n_b blocks (left). Each block consists of n_f flows (middle). In each flow we apply activation normalization, followed by the affine coupling (right), after which the binary mask for the even/odd mapping is inverted. The affine coupling layer uses a WaveNet with the *even* set as the input to output scaling $\log s$ and translation t with which the *odd* set is transformed. The squeeze operator, *squeezes* the time-dimension into the channel dimension doubling the number of channels.