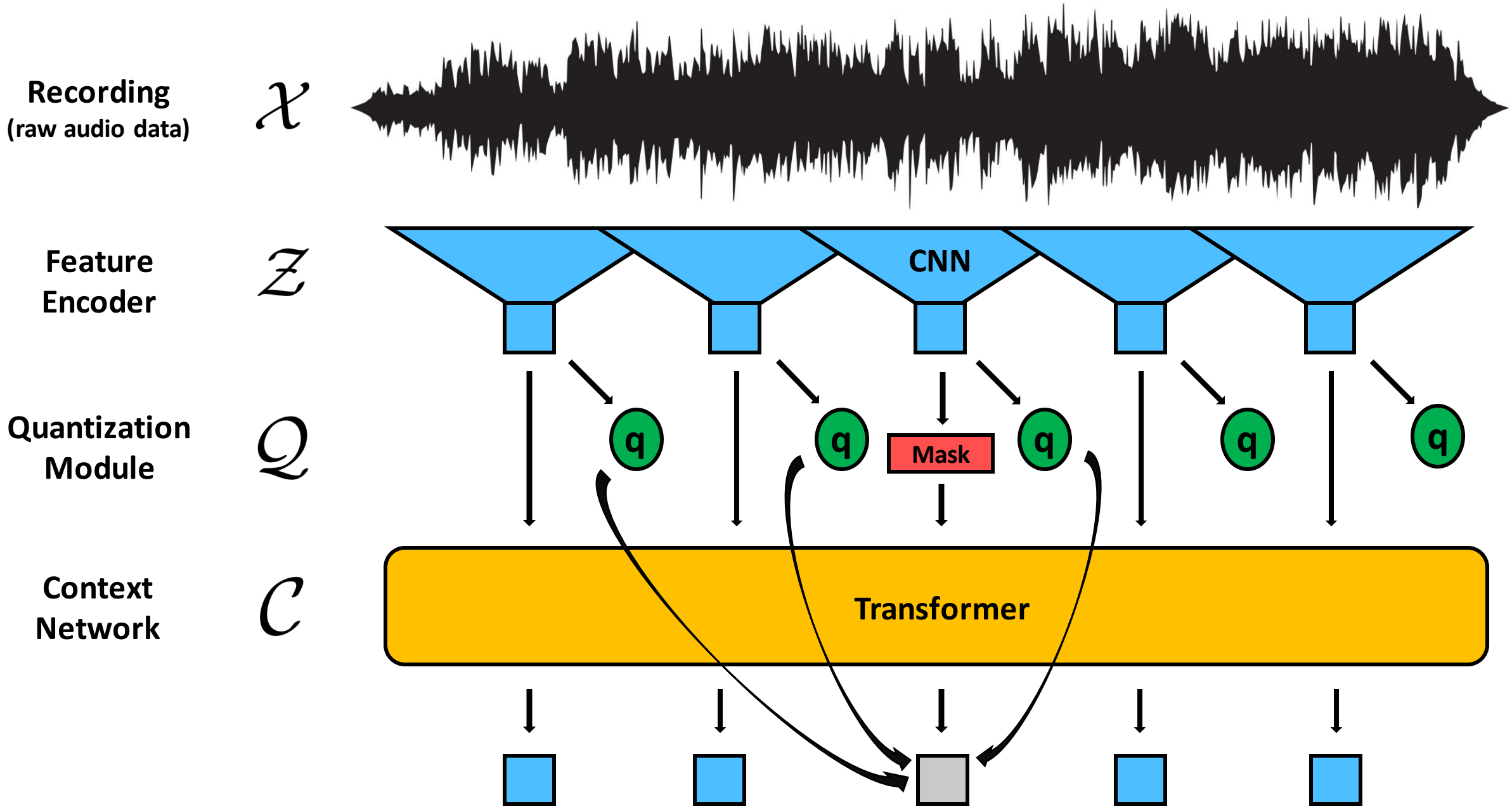
Recording
(raw audio data)
$\mathcal{X}$
Feature
Encoder
$\mathcal{Z}$
CNN
Quantization
Module
$\mathcal{Q}$
q
q
Mask
q
q
q
Context
Network
$\mathcal{C}$
Transformer