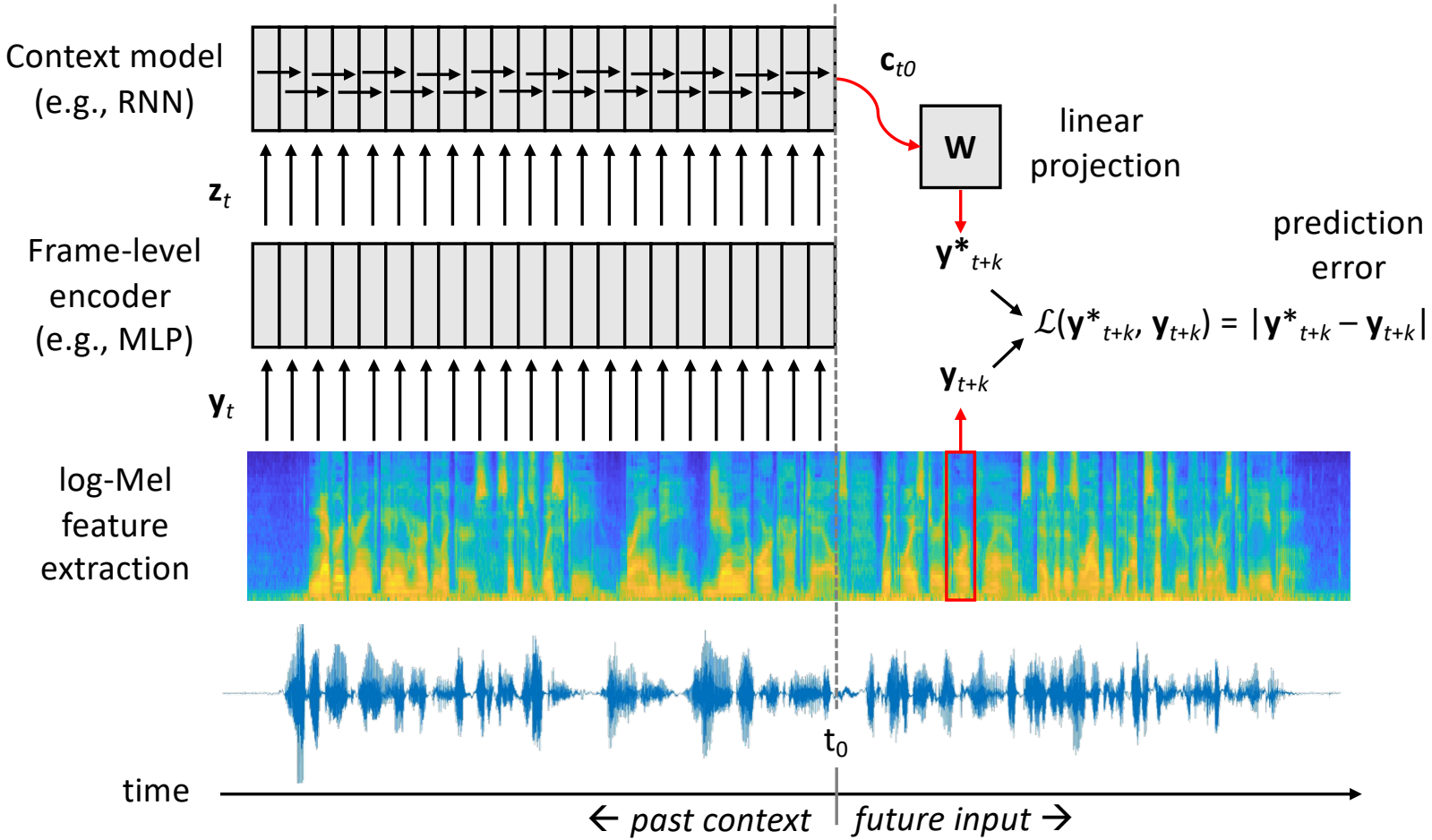
Context model
(e.g., RNN)
$\mathbf{c}_{t0}$
W
linear
projection
$\mathbf{z}_t$
Frame-level
encoder
(e.g., MLP)
$\mathbf{y}^*_{t+k}$
prediction
error
$\mathcal{L}(\mathbf{y}^*_{t+k}, \mathbf{y}_{t+k}) = |\mathbf{y}^*_{t+k} - \mathbf{y}_{t+k}|$
$\mathbf{y}_t$
$\mathbf{y}_{t+k}$
log-Mel
feature
extraction
t_0
time
← past context
future input →