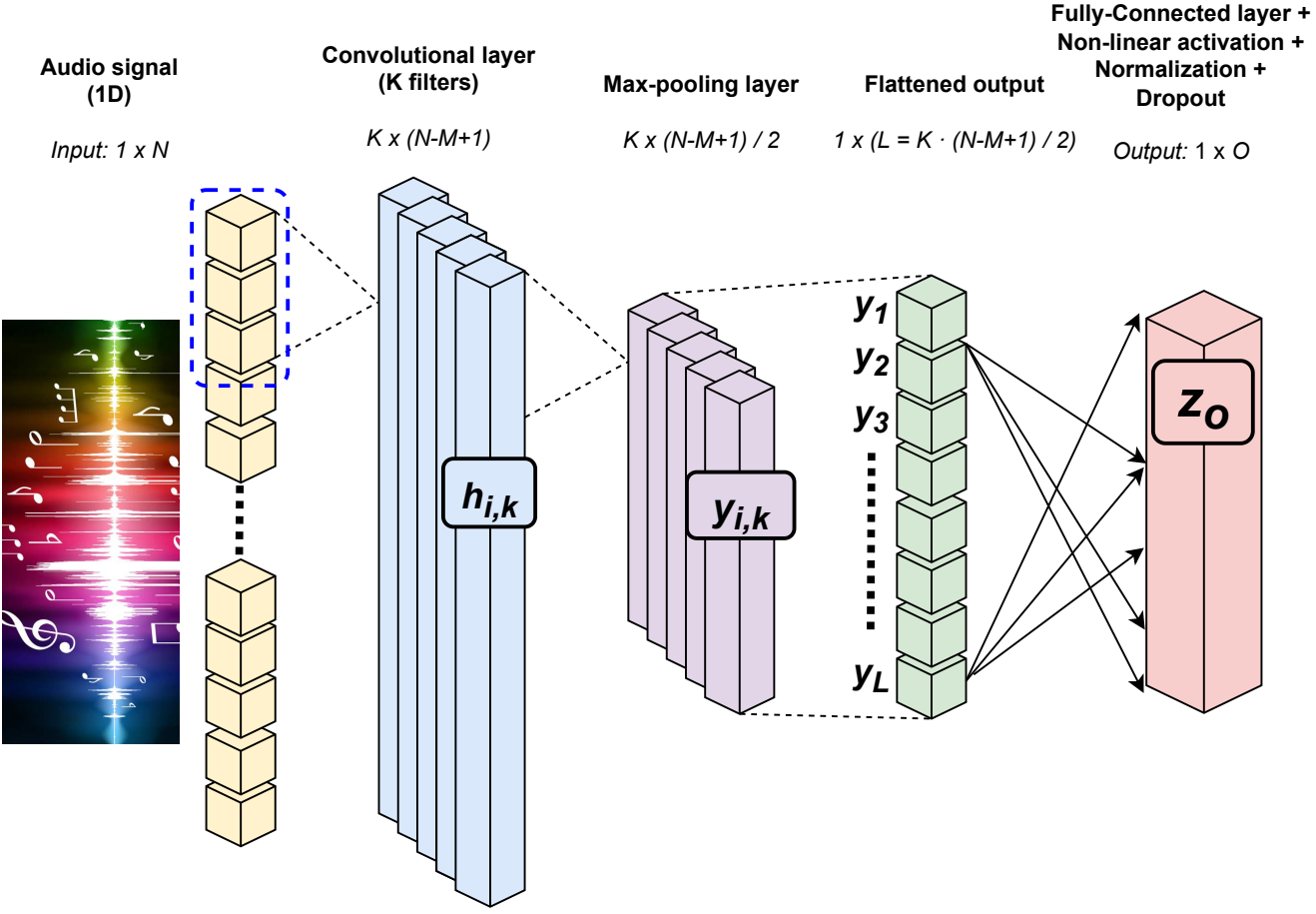
Audio signal
(1D)
Input: 1 x N
Convolutional layer
(K filters)
K x (N-M+1)
Max-pooling layer
K x (N-M+1) / 2
Flattened output
1 x (L = K · (N-M+1) / 2)
Fully-Connected layer +
Non-linear activation +
Normalization +
Dropout
Output: 1 x O
$h_{i,k}$
$y_{i,k}$
$y_1$
$y_2$
$y_3$
$y_L$
$z_O$