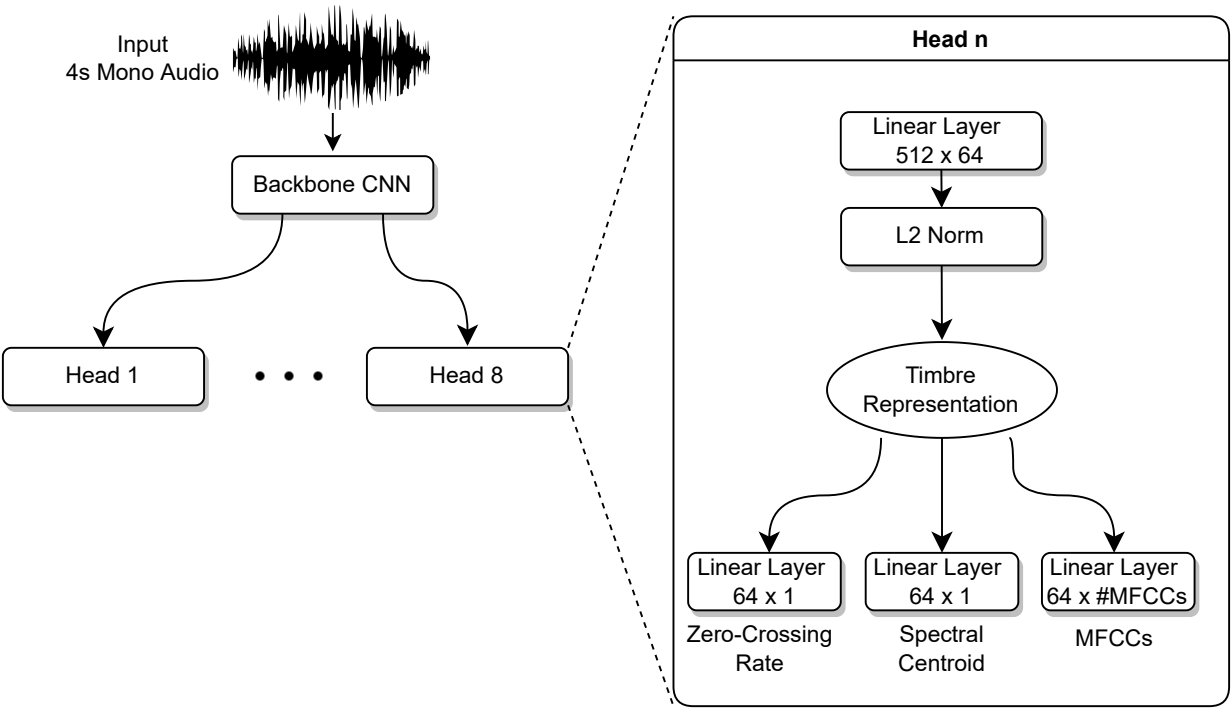
Input
4s Mono Audio
Backbone CNN
Head 1
Head 8
Head n
Linear Layer
512 x 64
L2 Norm
Timbre
Representation
Linear Layer
64 x 1
Zero-Crossing
Rate
Linear Layer
64 x 1
Spectral
Centroid
Linear Layer
64 x #MFCCs
MFCCs