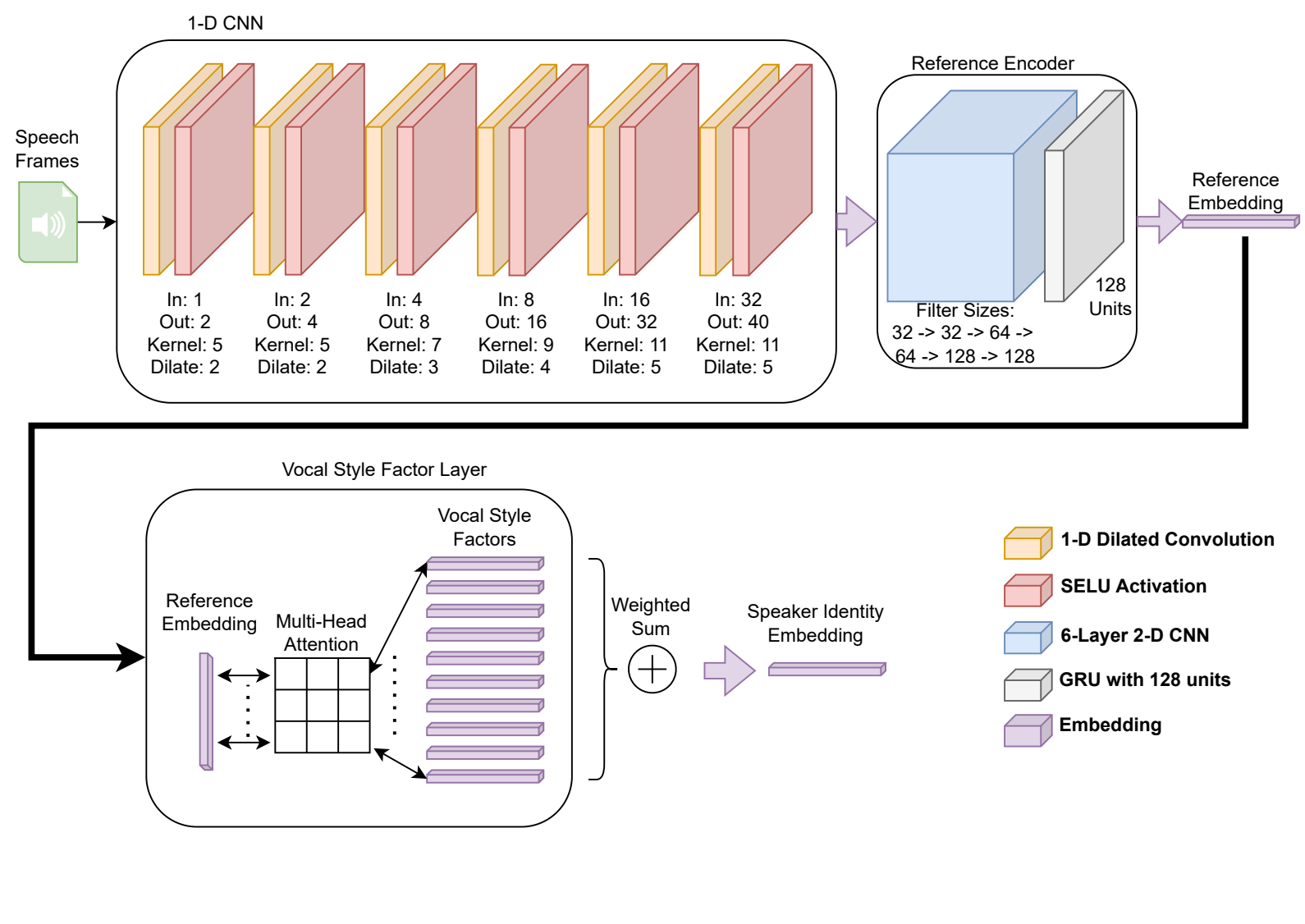
1-D CNN
Speech Frames
In: 1
Out: 2
Kernel: 5
Dilate: 2
In: 2
Out: 4
Kernel: 5
Dilate: 2
In: 4
Out: 8
Kernel: 7
Dilate: 3
In: 8
Out: 16
Kernel: 9
Dilate: 4
In: 16
Out: 32
Kernel: 11
Dilate: 5
In: 32
Out: 40
Kernel: 11
Dilate: 5
Reference Encoder
Filter Sizes:
32 -> 32 -> 64 ->
64 -> 128 -> 128
128
Units
Reference Embedding
Vocal Style Factor Layer
Reference Embedding
Multi-Head Attention
Vocal Style Factors
Weighted Sum
Speaker Identity Embedding
1-D Dilated Convolution
SELU Activation
6-Layer 2-D CNN
GRU with 128 units
Embedding