Table 2: The detailed structure of the 1D Inception module and voice encoder. In the descriptions, Conv x/y denotes 1D convolution with kernel size of x and stride length of y

1D Inception Module		
Component	**Activation**	**Dimension**
Input	-	$d_i \times t_i$
Conv 2/2	BN + ReLU	$d \times t_o$
Conv 3/2	BN + ReLU	$d \times t_o$
Conv 5/2	BN + ReLU	$d \times t_o$
Conv 7/2	BN + ReLU	$d \times t_o$
Concat.	-	$4d \times t_o$

Voice Encoder	
Layer	**Dimension**
Input	$40 \times t_0$
Inception 1	$256 \times t_1$
Inception 2	$384 \times t_2$
Inception 3	$576 \times t_3$
Inception 4	$864 \times t_4$
Inception 5	$512 \times t_5$
Time AvePool	512×1