A. Table of hyper-parameters

Table 6. Model hyper-parameters used in the experiments. ("$\times n$": n layers)

	Conversational (UC)	CommonVoice 11 (CV11)	CommonVoice 11 (CVE)
Input & Output			
Sample rate (Hz)	16,000	16,000	16,000
Mel channels	128	128	128
Mel lower band (Hz)	20	20	20
Mel upper band (Hz)	8,000	8,000	8,000
Frame size (ms)	50.0	50.0	50.0
Frame step (ms)	12.5	12.5	12.5
SpecAugment			
Freq blocks	2	2	2
Time blocks	10	10	10
Freq block max length ratio	0.33	0.33	0.33
Time block max length ratio	0.05	0.05	0.05
Encoder			
Conformer dims	144×16	144×16	144×16
Attention heads	4	4	4
Conv kernal size	32	32	32
Subsample factor	4	4	4
Attention (source & target)			
Output & Hidden dim	512	512	512
Attention heads	8	8	8
Dropout prob	0.2	0.2	0.2
Decoder (source & target)			
Transformer (dim $\times$ layers)	512×4	512×4	512×4
Hidden dims	512×4	512×4	512×4
Dropout prob	0.3	0.3	0.3
Phoneme embedding dim	256	256	256
Label smoothing uncertainty	0.1	0.1	0.1
Loss weight	1.0	1.0	1.0
Duration predictor (source & target)			
Bi-LSTM (dim $\times$ layers)	128×2	128×2	128×2
Loss weight	1.0	1.0	10.0
Synthesizer (source & target)			
LSTM dims	$1{,}024 \times 2$	$1{,}024 \times 2$	$1{,}024 \times 2$
LSTM zoneout prob	0.1	0.1	0.1
Pre-net dims	128×2	128×2	128×2
Pre-net dropout prob	0.5	0.5	0.5
Post-net (kernel, channels) $\times$ layers	$(5, 512) \times 4 + (5, 128)$	$(5, 512) \times 4 + (5, 128)$	$(5, 512) \times 4 + (5, 128)$
Loss weight	1.0	1.0	1.0
WaveFit vocoder			
Iterations	5	5	5
UBlock upsampling factors	[5, 5, 2, 2, 2]	[5, 5, 2, 2, 2]	[5, 5, 2, 2, 2]
STFT loss resolutions	3	3	3
Hann win size, frame shift, FFT size res 1	[160, 32, 512]	[160, 32, 512]	[160, 32, 512]
Hann win size, frame shift, FFT size res 2	[400, 80, 1024]	[400, 80, 1024]	[400, 80, 1024]
Hann win size, frame shift, FFT size res 3	[800, 160, 2048]	[800, 160, 2048]	[800, 160, 2048]
Multi-period discriminator	Kong et al. (2020)	Kong et al. (2020)	Kong et al. (2020)
Multi-period discriminator loss weight	1.0	1.0	1.0
Training			
Optimizer	Adam (Kingma & Ba, 2014)	Adam (Kingma & Ba, 2014)	Adam (Kingma & Ba, 2014)
Learning rate schedule	Vaswani et al. (2017)	Vaswani et al. (2017)	Vaswani et al. (2017)
Learning rate (peak)	1.3×10^{-3}	1.3×10^{-3}	1.3×10^{-3}
Warm-up steps	20K	40K	20K
Batch size	512	512	512
L^2 regularization weight	10^{-6}	10^{-6}	10^{-6}
MUSE loss weight	100000.0	1000.0	1.0