A. Detailed model architecture

Video Encoder	Input size	$D_w = D_h = 128, D_c = 1$
	Max Pooling per layer	(F, T, T, T, T)
	Number of output channels per layer	(64, 128, 256, 512, 512)
	Stride per layer	(2, 1, 1, 1, 1)
	Kernel size (for all layers)	(3, 3, 3)
	Activation (for all layers)	ReLU
	Normalization (for all layers)	Group Norm
Text Encoder	Input size	$D_e = 512$
	Conv layers × 3	2048 5×1 kernel with 1×1 stride
	Activation (for all Conv layers)	ReLU
	Bi-LSTM	1024-dim per direction
	Normalization (for all Conv layers)	Batch Norm
Multi Source Attention	Attention input size	$D_m = 2048$
	GMM attention (per source)	128-dim context
	Linear Projection	Fully connected layer
Decoder	PreNet	2 fully connected layers with 256 neurons and ReLU act.
	LSTM × 2	1024-dim
	Bi-LSTM	1024-dim per direction
	PostNet	5 conv layers with 512 5×1 kernel with 1×1 stride and TanH act.
	Normalization (for all Decoder layers)	Batch Norm
	Teacher forcing prob	1.0

B. Training hyperparameters

Training	learning rate	0.0003
	learning rate scheduler type	Linear Rampup with Exponential Decay
	scheduler decay start	40k steps
	scheduler decay end	300k steps
	scheduler warm-up	400 steps
	batch size	512
Optimizer	optimizer details	Adam with $\beta_1 = 0.9, \beta_2 = 0.999$
Regularization	L2 regularization factor	1e-06

C. Word error rate discussion

As explained in Sec. 4.4, our VoxCeleb2 transcripts are automatically generated and thus contain transcription errors. As a result one can expect the WER for models trained on this data to be non-zero. In order to validate this hypothesis, that the result of such noisy data leads to a non-zero WER, we trained a version of the our model that accepts only text as input (without silent video), denoted as TTS-OUR. TTS-OUR was trained twice, once on the LibriTTS [64] dataset, and a second time when using our in-the-wild LSVSR dataset. When looking at Table 6 it is clear that when trained on LibriTTS this model achieves a low WER of 7%, while the same model when trained on in-the-wild dataset get a WER of 27%. This suggests that a WER in the region of $[20\%, 30\%]$ should be expected when using LSVSR.

That being said, we believe reporting WER is valuable as a sanity check for noisy datasets, specially when trying to capture more than just the words.

Training data	WER
LIBRITTS	7%
LSVSR	27%

Table 6. Comparison of WER on the VoxCeleb2 test set for our text only TTS model (TTS-OUR) when trained on different datasets.