

	Speaker adaptation		Speaker encoding	
Approaches	Embedding-only	Whole-model	Without fine-tuning	With fine-tuning
Pre-training	Multi-speaker generative model			
Data	Text and audio		Audio	
Cloning time	~ 8 hours	$\sim 0.5 - 5$ mins	$\sim 1.5 - 3.5$ secs	$\sim 1.5 - 3.5$ secs
Inference time	$\sim 0.4 - 0.6$ secs			
Parameters per speaker	128	~ 25 million	512	512