TABLE – SotA: TTS (Text – To- Speech).

ID	Features	Databases	Classifier	Methodology Details	Best Results	Link
1	• Mel-spectrograms • Phoneme-embeddings • Positional Encoding	• LJSpeech	• Feed-forward Transformer consists of Phoneme Embeddings + Positional Enmcodings, 6xFFT Block, Length Regulator with Duration Predictor (2xConv1D + 1xLinear Layer), 6xFFT Block, Linear Layer	• first train the autoregressive Transformer TTS model on 4 NVIDIA V100 GPUs • batch_size=16, Adam with β1=0.9, β1=0.98, $\varepsilon=10^{-9}$ • train the duration predictor -> feed the text and speech pairs in the training set to the model again to obtain the encoder-decoder attention alignments • source text sequence + generated mel-spectrograms with the autoregressive Transformer TTS model => the paired data for FastSpeech model training • train the FastSpeech model together with the duration predicton • output mel-spectrograms are transformed into audio samples using the pretrained WaveGlow	• MOS = 3.84 ± 0.08 • Latency = 0.18 ± 0.078s	ART 2019
2	• Raw audio waveform	• VCTK	• Causal Conv + Residual Blocks (Dilated Conv + Tanh x Dilated Conv + Sigmoid) + 2x Linear + Softmax	• Causal conv keeps the order of the samples • Softmax layer has 65,536 probabilities corresponding to 16-bit integer representation for a sample • Condition the model not in text, but speaker • Receptive field: 240ms	• MOS = 4.21 ± 0.081 (American English) • MOS = 4.08 ± 0.085 (Mandarin Chinse)	ART 2016
3	• 80-dimensional log-mel filter bank coefficients • Phoneme-embeddings	• LJ Speech • VCTK	• VAE (Posterior Encoder – residual blocks from WaveGlow + Prior Encore – transformer encoder + Decoder – HiFi GAN v1 Generator + Stohastic Duration Predictor - residual blocks with dilated and depth-separable convolutional layers)	• AdamW optimizer: β1 = 0.8, β2 = 0.99, weight decay λ = 0.01 • $Lr = 2 \times 10^{-4}$ • Batch_size=64 • 800k steps	• MOS = 4.43 (±0.06) LJSpeech • MOS = 4.38 (±0.06) VCTK	ART 2021
4	• Phoneme-embeddings • High-dimensional speech compressed representation.	• LJSpeech • A large-scale text corpus with 200 million sentences for phoneme pre-training	• Phoneme Encoder: 6-layer Feed-Forward Transformer (FFT) blocks+ Duration predictor with upsampling layer + Posterior Encoder (16-layer WaveNet) + Bidirectional Prior and Posterior Module + Waveform Decoder(Residual convolution blocks with upsampling).	• Pre-train phoneme embedding model. • Memory-based VAE to simplify the posterior. • Bidirectional prior/posterior modeling for enhancing prior and reducing the complexity of posterior. • Trained on 8 8 NVIDIA V100 GPUs • Optimizer: AdamW • lr=20e-42, γ=0.999875 • β1=0.8, β2=0.99	• MOS = 4.56 ± 0.13	ART 2022
5	• Phoneme-embeddings • Mel-spectrograms • Scalar Quantization Codec	• Multilingual LibriSpeech (MLS) • WenetSpeech	• Text Encoder(T5 arihitecture) + Speaker Encoder + SQ-Codec(consists of an encoder, decoder, and scalar quantization for compact latent representations) + Transformer Diffusion Model + Dense Layer	• lr=10e-41 for transformer diffusion, with a cosine scheduler and warmup of 1k steps. • optimizer: Adam optimizer used for both SQ-Codec and transformer diffusion model. • Diffusion steps=25 • Sentence duration predictors are used to control the length of generated speech.	• MOS = 4.06 ±.052	ART 2024

6	• speaker latent vectors from the reference mel spectrogram. • local frame-level features.	• LibriTTS • AiShell3	• FastSpeech 2 Integration(Encoder - phoneme encoder integrated with TSCM to incorporate speaker control + Variance Adaptor - adds duration, pitch, and energy information + Mel-spectrogram Decoder - utilizes TSCM for speaker-specific adaptations) + VITS Integration(Text Encoder, Duration Predictor, Generator: Integrated with TSCM to improve speaker control).	• batch_size=50 • Training Steps=250,000 • Optimizer: Adam • Pretrained HiFi-GAN vocoder used to convert mel spectrograms to audio.	• TSCM-FastSpeech 2: MOS = 3.81 • TSCM-VITS: MOS = 4.44	ART 2024
7	• Mel-spectrograms • Spectrogram Tokens (Vector Quantized) • Phoneme Embeddings	• LJSpeech	• Spectrogram VQ Model(consists of an encoder, decoder, and discrete codebook) + Text Encoder(EfficientSpeech encoder with 2 transformer blocks) + Discrete Diffusion Model (12-layer transformer with 8 heads) with Contrastive Learning.	• Compresses mel-spectrograms into discrete tokens to reduce computational costs. • Diffusion Model Training: Uses a forward and reverse diffusion process with a Mask-and-Replace strategy. • Contrastive Learning: Text-wise Contrastive Learning Loss added to improve text-spectrogram alignment. • Inference: Generated mel-spectrograms are converted into audio using the Griffin-Lim algorithm.	• MOS: 3.64 ± 0.05 • mRTF (Real-Time Factor): 73.9 (GPU), 17.6 (CPU)	ART 2023
8	• Mel-spectrograms • Phoneme Representations • Dynamic Quantized Representation	• LJSpeech	• Sequential Autoencoder (Encoder - convolution blocks and LSTMs + Dynamic Codebook Module + Decoder: Tacotron 2-based)	• Dynamic Quantized Representation Learning: Quantization with a dynamic codebook that expands based on unpaired data using pseudo-labels generated by a pre-trained ASR. • Train with a mix of 120 minutes of paired data and 600 minutes of unpaired data. • batch_Size=64 • Optimizer: Adam (β1=0.9, β2=0.999, , lr=10e−3). • Use Connectionist Temporal Classification (CTC) for recognition loss.	• MOS: 3.12 ± 0.32 with mixed data (120 min paired + 600 min unpaired).	ART 2024
9	• Mel-spectrograms • Phoneme embeddings • Pitch, duration, and energy features	• LJSpeech • VCTK • LibriTTS	• Phoneme Encoder(Transformer blocks) + Variance Adaptor + CM-Decoder (non-causal WaveNet-like structure) + Vocoder (HiFi-GAN).	• CM-TTS employs a consistency model-based approach for real-time mel-spectrogram generation. • Utilizes weighted samplers to improve model training by incorporating dynamic probabilities. • Model trained for 300K steps with exponential learning rate decay and a batch size of 32.	• MOS: 3.9618 ± 0.0186 • Latency: Real-time capability with fewer synthesis steps (1, 2, 4 steps tested).	ART 2024
10	• Prosody features (pitch, duration) • Phoneme embeddings	• LibriSpeech: • MLS corpus	• Phoneme Encoder + Codec Decoder (SoundStream) + VALL-E (12-layer transformers)	• Prosody Tokens are predicted using a Chain-of-Thought (CoT) prompting technique, stabilizing pitch and duration before speech token prediction • Trained on 8 NVIDIA V100 and 16 AMD MI200 GPUs with Adam optimizer • Utilized nucleus sampling for phoneme, pitch, and duration prediction • Tested multiple window sizes for duration-guided masking, with optimal size being 1 for WER improvements • Evaluation on hard sentences to assess robustness, with error types classified as mispronunciation, omission, repetition, and hallucination	• Word Error Rate (WER): 2.5% (RALL-E), 5.6% (VALL-E) on standard LibriSpeech test-clean set • MOS:nUTMOS = 4.00 for RALL-E,	ART 2024
11	• Mel-spectrograms • Vector Quantized Variational codes	• XTTS • LibriTTS-R • Common Voice	• VQ-VAE (13M parameters) encodes mel-spectrograms to 1024 codebook vectors + GPT-2 encoder (443M parameters) predicts VQ-VAE audio codes from text input + Conditioning Encoder (6 layers) for generating speaker embeddings, producing 32 embeddings per audio + HiFi-GAN based decoder (26M parameters) reconstructs audio from latent vectors	• The model was trained on multilingual datasets using a language batch balancer • XTTS was trained for approximately 2.5M steps on 4 NVIDIA A100 GPUs (80GB) • AdamW optimizer, betas 0.9 and 0.96, with MultiStepLR learning rate scheduler • XTTS improves speaker cloning capability by conditioning the encoder on multiple embeddings, rather than a single embedding	• English Evaluation: CER: 0.5425 UTMOS: 4.007 ± 0.25 SECS: 0.6423	ART 2024

12	• Phoneme monotonic alignment • Merged codec with reduced sampling rate • Phoneme prediction during training	• LibriSpeech	• Encoder-decoder (convolution-based encoder) + Residual Vector Quantizer module (8-layer) + • Transformer-based architecture (12-layer) + Vocoder	• Two-stage training: autoregressive model predicts acoustic tokens from phonemes and aligned phoneme sequences, and NAR model iteratively generates tokens for higher layers. • The merged codec reduces the number of autoregressive steps by downsampling in the first layer without retraining the codec. • Monotonic alignment ensures that phoneme and acoustic tokens align, improving robustness by preventing repetition or skipping. • Experiments used 3-second acoustic prompts and phoneme sequences for zero-shot TTS tasks like speech continuation and cross-sentence synthesis.	• QMOS = 4.02, SMOS = 3.89 • Latency: achieved a generation time of 3.67s for 10s of speech	**ART** 2024
13	• Mel-spectrograms • Time-Invariant and Time-Variant Style Representations	• VCTK • Emotional Speech Dataset.	• Text Encoder(8 layers of Transformer encoders with relative positional embedding and adaptive layer normalization) + Aligner(Convolution-based Duration Predictor) + Diffusion Decoder (convolution blocks and DiT blocks) + Time-Invariant and Time-Variant Encoders (uses cross-attention).	• Batch_size=32 • Optimizer: Adam (lr=10e-4) • 1000 epochs for VCTK, 1500 epochs for ESD • Diffusion Process: Incorporates Gaussian noise into input data and iteratively refines to generate mel-spectrograms.	• VCTK: Seen Speakers: 3.75 (MOS-N), 3.88 (MOS-S) Unseen Speakers: 3.76 (MOS-N), 3.81 (MOS-S) • ESD: Seen Speakers: 3.73 (MOS-N), 3.84 (MOS-S) Unseen Speakers: 3.57 (MOS-N), 3.52 (MOS-S)	**ART** 2024
14	• Mel-spectrograms • Phoneme-embeddings • Scalar Quantization Codec	• Multilingual LibriSpeech • WenetSpeech	• Text Encoder (ByT5 model with 2 transformer blocks) + Speaker Encoder (Pre-trained FACodec) + SQ-Codec (Encoder, decoder, and scalar quantization) + Transformer Diffusion Model (flow-based scalar latent transformer diffusion with 12 layers and 8 heads).	• Batch_size =32 • Optimizer: Adam (lr=10e−4) • 400,000 training steps • Diffusion steps: 25 • Sentence Duration Prediction:4 strategies explored: ByT5-based, ChatGPT-based, FS2-based, and AR-based duration predictors. • Generated mel-spectrograms converted to audio using SQ-Codec decoder.	• MOS: 4.28 ± 0.12	ART 2024
15	• Mel-spectrograms • Style Embeddings (text and audio) • Speaker Embeddings	• Emotional Speech Dataset	• General Style Fusion Encoder (CLIP-based text encoder and audio encoder) + Hierarchical Conformer Two-Branch Style Control Module (fuses style and speaker control embeddings into the VITS-based TTS architecture for optimal control of both speaker and style) + Backbone (VITS).	• 1,000,000 training steps • Multimodal Input Processing: combines style prompts and audio references to control style and speaker embeddings. • Gradient Reversal Layer: used to disentangle speaker and style information. • HiFi-GAN is used as a vocoder, with Speech Super Resolution upsampling for enhanced quality.	• Speaker-MOS: 4.19 • Emotion-MOS: 4.28	ART 2024
16	• Phoneme embeddings • Pitch • Duration • Energy	• BEAT2	• Phoneme Encoder (Causal Transformer Encoder) + Rhythmic Predictors (Separate CNN-based predictors for pitch, duration, and energy) + Shared Rhythm Predictors + Speech Decoder (1D dilated convolutions) + Gesture Decoder (Pretrained VQ-VAE for gesture reconstruction using semantic and rhythmic latent features) + Neural Architecture Search	• Joint Generation: the model jointly generates speech and gestures by sharing intermediate rhythmic features (pitch, duration, and energy). • Causal Network: redesigned to avoid dependencies on future inputs for real-time applications. • Neural Architecture Search: used to fine-tune model components (e.g., layers and convolutional kernels) for performance improvement. • Training on NVIDIA A100 GPUs, leveraging reinforcement learning-based NAS to optimize network hyperparameters.	• MOS: 3.93 (speech quality) • Latency: 0.17 seconds per second (speech and gesture generation on NVIDIA 3090)	**ART** 2024

17	• Mel-spectrograms • Prosody (pitch, duration) • Phoneme embeddings • Style diffusion modeling (latent random variable for speech style)	• LJSpeech • VCTK • LibriTTS	• Text Encoder (Causal Transformer for phoneme representation) + Prosodic Text Encoder (BERT-based) +Style Encoder + Duration and Prosody Predictors + Waveform Decoder (iSTFTNet or HifiGAN) + SLM Discriminators (WavLM-based adversarial training, with convolutional head)	• Training Strategy: two-stage process; first pre-train acoustic modules (100 epochs on LJSpeech), then jointly optimize all components with differentiable duration modeling. • Adversarial Training: utilizes large pre-trained Speech Language Models (SLMs) like WavLM as discriminators for human-like quality synthesis. • End-to-End Training: joint optimization of text encoder, style encoder, prosody predictor, and waveform decoder for direct waveform generation. • Diffusion-Based Sampling: style diffusion model samples a latent style vector conditioned on text, enabling diverse and expressive speech generation without reference audio. • Zero-Shot Speaker Adaptation: fine-tuned on LibriTTS for zero-shot speaker adaptation using only 3-second reference clips.	• MOS: 3.83 (LJSpeech, surpasses ground truth with CMOS +0.28) 4.15 (LibriTTS, zero-shot) 4.03 (Similarity score for zero-shot speaker adaptation on LibriTTS) • Latency: 0.0185 Real-Time Factor	ART 2024
18	• Mel-spectrograms • Latent space control for speech variation (pitch, tone, speech rate, cadence, accent)	• LJSpeech • LibriTTS	• Text Encoder (modified Tacotron encoder -instance-norm replaces batch-norm) + Affine Coupling Layer + Latent Space Control (invertible mapping between mel-spectrograms and latent space, modeled using a Gaussian Mixture Model) + Waveform Decoder (WaveGlow)	• Maximize the likelihood of training data to train the autoregressive flow model. • Train the model on the LSH dataset for 1,000 epochs and fine-tune for 500 epochs on LibriTTS. • Models trained on NVIDIA DGX-1 with 8 GPUs. • Variability Control: Adjust the amount of variation in speech output by sampling from a Gaussian prior with different variances (σ^2 = 0.0, 0.5, 1.0). • Posterior Sampling: Style transfer between seen and unseen speakers by sampling from a posterior distribution conditioned on prior evidence (e.g., expressive vs. monotonic speech styles). • Interpolation: Smooth interpolation between samples and speakers by manipulating latent space (z-space).	• MOS: 3.665 ± 0.1634 (Flowtron) 3.521 ± 0.1721 (Tacotron 2) 4.274 ± 0.1340 (real data)	ART 2020
19	• Mel-spectrograms • Hierarchical latent variables • Phoneme embeddings • Positional encoding (sinusoidal)	• LJSpeech • Multi-speaker internal Mandarin Chinese corpus (55 hours, 7 female speakers)	• Very Deep Variational Autoencoder(residual blocks containing 4 convolution layers with GELU activation) + Residual Attention Mechanism + Text Encoder (Feature-wise Linear Modulation + 4 convolution layers and positional encodings) + Speaking Speed Predictor	• Batch size of 32, trained on 2 NVIDIA V100 GPUs for 90k iterations using the Adam optimizer (β1=0.9, β2=0.999), learning rate maxing at 1.5e-4 with 10k warm-up steps. • Combined loss including reconstruction loss, Kullback-Leibler divergence, and speaking speed prediction loss. • Inference is 16x faster than Tacotron 2 due to the non-autoregressive nature of the model.	• MOS on LJSpeech: 3.88 ± 0.20 • MOS on the multi-speaker Mandarin dataset: 4.49 ± 0.11 • VARA-TTS inference speed: 32.01ms	

20	• Mel-spectrograms (80-dimensional) • Speaker embeddings • Phoneme-level input for text processing	• VCTK • LibriTTS	• Speaker Encoder (ECAPA-TDNN) + Squeeze-and-Excitation blocks for channel interdependencies + Res2Net blocks with skip connections for feature aggregation + Improved pooling with channel/context-dependent frame attention + Acoustic Model (FastSpeech 2, non-autoregressive model) + Transformer-based encoder and decoder + Variant adaptor with duration, pitch, and energy predictors + Postnet (Conv1D blocks) added after the decoder for fine-tuning.+ Vocoder (HiFi-GAN)	• Data downsampled to 22,050 Hz for experiments • Pretrained speaker encoder on VoxCeleb1 and VoxCeleb2 datasets • Acoustic model trained for 400k steps with a batch size of 16 using an RTX 3090 GPU • Ground-truth phoneme durations were obtained via Montreal Force Aligner • Testing includes seen speakers (from VCTK) and unseen speakers (from LibriTTS and VCTK). • Average speaker embedding calculated for each speaker to improve stability and speaker similarity.	• MOS-N: Ground truth: ~4.19 ECAPA-TDNN: 3.62 (seen), 3.47 (unseen LibriTTS) x-vector: 3.51 (seen), 3.38 (unseen LibriTTS)	ART 2022
21	• Phoneme embeddings (raw phonetic input) • Explicit modeling of prosody (pitch, duration) • BERT-based contextualized word embeddings • Hybrid grapheme-to-phoneme conversion with punctuation modeling	• French language dataset from Blizzard Challenge 2023 • Speaker 1 (NEB): 50 hours of aligned, high-quality audiobook data • Speaker 2 (AD): 2 hours of aligned data	• Custom network for prosody prediction (handles duration and pitch) + CamemBERT for contextualized word embeddings + HiFi-GAN for vocoding + 3 parallel stacks of BiLSTMs for duration, pitch, and vocoder conditioning + Shared sub-network with 3 convolutional layers + BiLSTM for phonetic embeddings + HiFi-GAN conditioned on the predicted duration and pitch embeddings + Grapheme-to-Phoneme Module (Stack of 3 convolutional layers followed by BiLSTMs)	• Training the phonemizer with a split dataset (90% train, 10% validation) • Early stopping with sentence accuracy rate • Fine-tuned CamemBERT along with the custom prosody network and HiFiGAN • BERT optimized with a fixed learning rate of 10e−6, 1M steps • Phonetic alignment and pitch annotations used for forced alignment • Two layers of non-uniform upsampling to match phonemes with BERT embeddings and duration predictions • Training performed on an NVIDIA RTX 3090 with a batch size of 16 over 3 weeks	• MOS for Speaker NEB: Speech Experts: 4.0 (±1.48) Non-speech experts: 4.3 (±0.74) • MOS for Speaker AD: Speech Experts: 3.3 (±1.00) Non-speech experts: 4.1 (±0.83)	ART 2023
22	• Self-supervised speech representations • Mel-spectrograms • Phoneme embeddings	• VCTK • LibriTTS	• Hierarchical Conditional Variational Autoencoder: Self-supervised speech representations from XLS-R (12th layer for linguistic information) + Text Encoder (Transformer with relative positional encoding) + Linguistic Encoder (Bi-directional WaveNet) + Acoustic Encoder (Non-causal WaveNet residual blocks for acoustic latent variables) + HiFi-GAN vocoder + Flow-based Stochastic Duration Predictor + Monotonic Alignment Search	• Training with AdamW optimizer (β1 = 0.8, β2 = 0.99, weight decay = 0.01) • HierSpeech trained on 4 NVIDIA A100 GPUs for 600k steps, batch size 256 • Untranscribed speech training (HierSpeech-U): Speaker adaptation without text transcripts using a style encoder • Evaluation using fine-tuned wav2vec 2.0 for phoneme and word error rates (PER, WER)	• VCTK: MOS = 4.04 (N), 3.22 (S) • LibriTTS: MOS = 3.98 (N), 3.26 (S) • Untranscribed Speech: MOS (naturalness): 4.08	ART 2022