Google Research エンジニアが語る、Cloud Text-to-Speech の新機能実装の舞台裏

Heiga Zen

Google
Senior Staff Research Scientist

Text-to-Speech

テキスト音声合成は、任意の文章を音声に変換する技術です。
音声アシスタント・スマートスピーカー・音声翻訳・ナビゲーションなど、音声インタフェースの出力部として用いられます。

Text → Speech

Google Cloud Text-to-Speech

TTS used in Google products → Cloud customers

220+ voices across 40+ languages

Standard (124)	WaveNet (113)

Main Use Cases for Cloud Text-to-Speech

Call centers

Replaces the need to hire voice talents to record fixed voice prompts for IVR, & enables more conversational flows.

IoT & Mobile

Build smart devices that can respond back to user queries.

Media

Transform articles & books into audio that can be consumed like a podcast or audiobook.

Limitation:

Share voices across customers.

Motivation:

Enterprise customers want TTS voices representing their brands’ personas.

Problem:

Creating a unique voice from scratch for each IVR & mobile app is complex & expensive.

Solution:

Provides a custom Text-to-Speech model, using own voice recordings.

How:

Fine-tune multi-speaker Tacotron 2 model using provided recordings of a target speaker.

Technical Details

01

Tacotron 2 (2017)

- End-to-end neural TTS model
  - Encoder-Decoder-Attention
  - WaveNet vocoder
- Proposed in 2017
- Still one of the state-of-the-arts

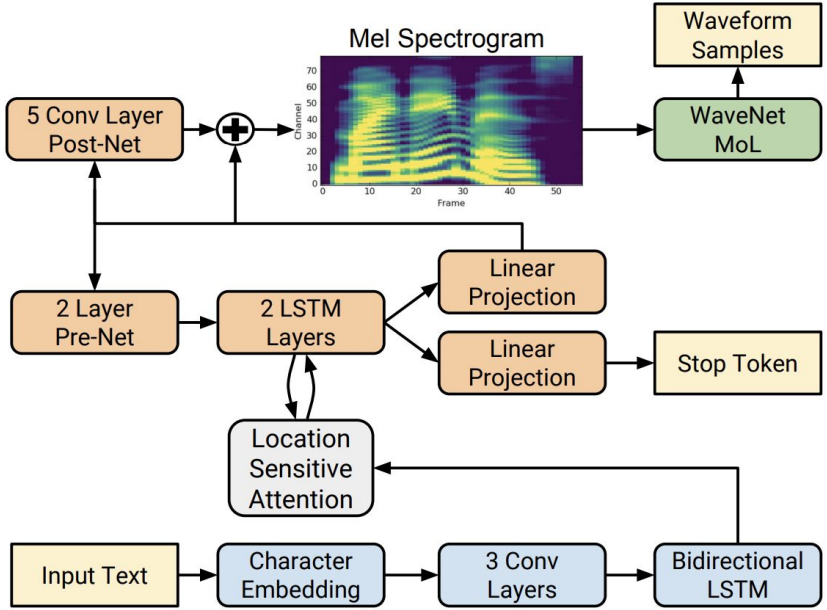


Fig. 1. Block diagram of the Tacotron 2 system architecture.

Tacotron 2 (2017)

System	MOS
Parametric	3.492 ± 0.096
Tacotron (Griffin-Lim)	4.001 ± 0.087
Concatenative	4.166 ± 0.091
WaveNet (Linguistic)	4.341 ± 0.051
Ground truth	4.582 ± 0.053
Tacotron 2 (this paper)	$\mathbf{4.526 \pm 0.066}$

Human Voice

Better than the above existing solutions, closer to human voice

Table 1. Mean Opinion Score (MOS) evaluations with 95% confidence intervals computed from the t-distribution for various systems.

Cloud Custom Voice Blocks

Text normalization

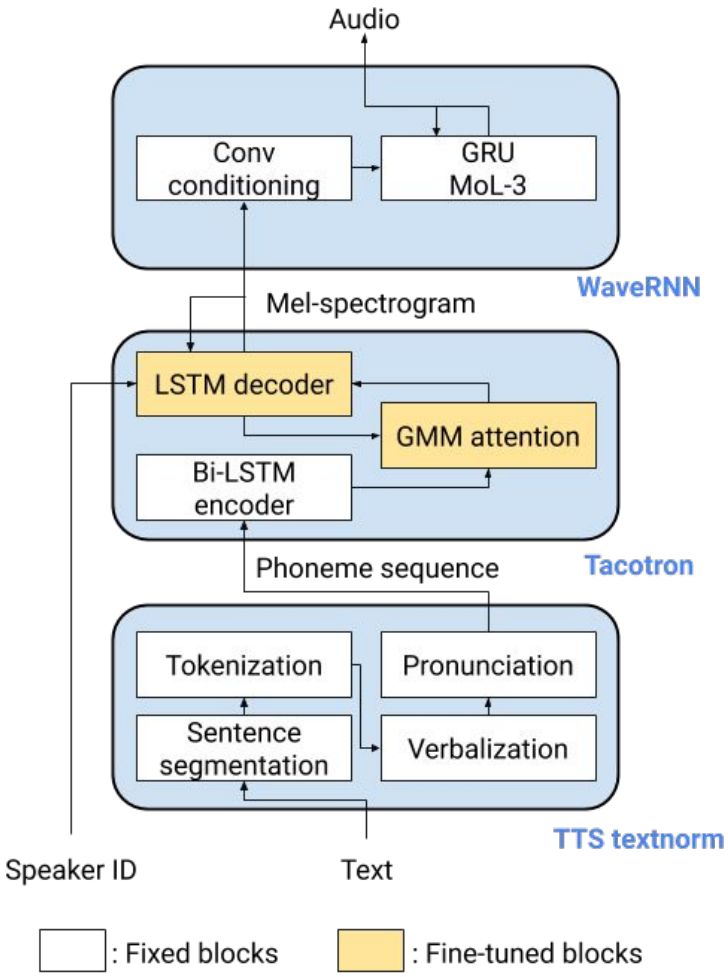
- Raw text → Phoneme + punctuation
- Same as Cloud TTS

Tacotron

- Phoneme + punctuation → Spectrogram
- Fine-tuning decoder

WaveRNN

- Spectrogram → Audio signals
- Universal neural vocoder



Tacotron for Custom Voice

Model

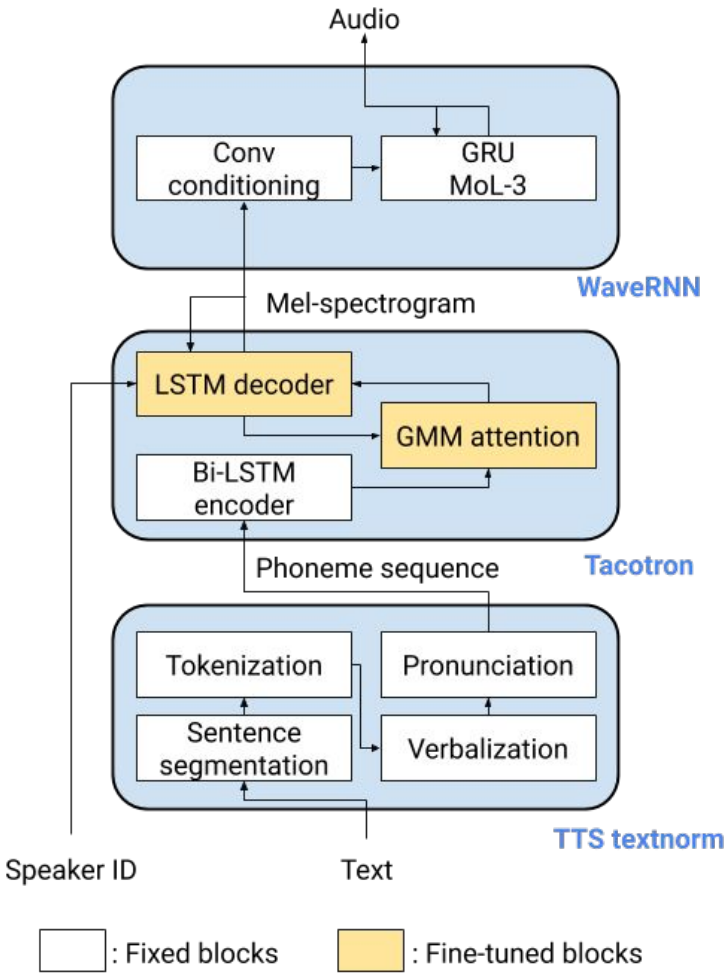
- GMM attention for better robustness
- Predict multiple frames for faster inference

Data

- Consistent data preprocessing
- ~430 hours , ~1.1k speakers

Fine-tuning

- Fine-tune decoder, attention, speaker emb.
- Include base training data for better stability



WaveRNN for Custom Voice

Model

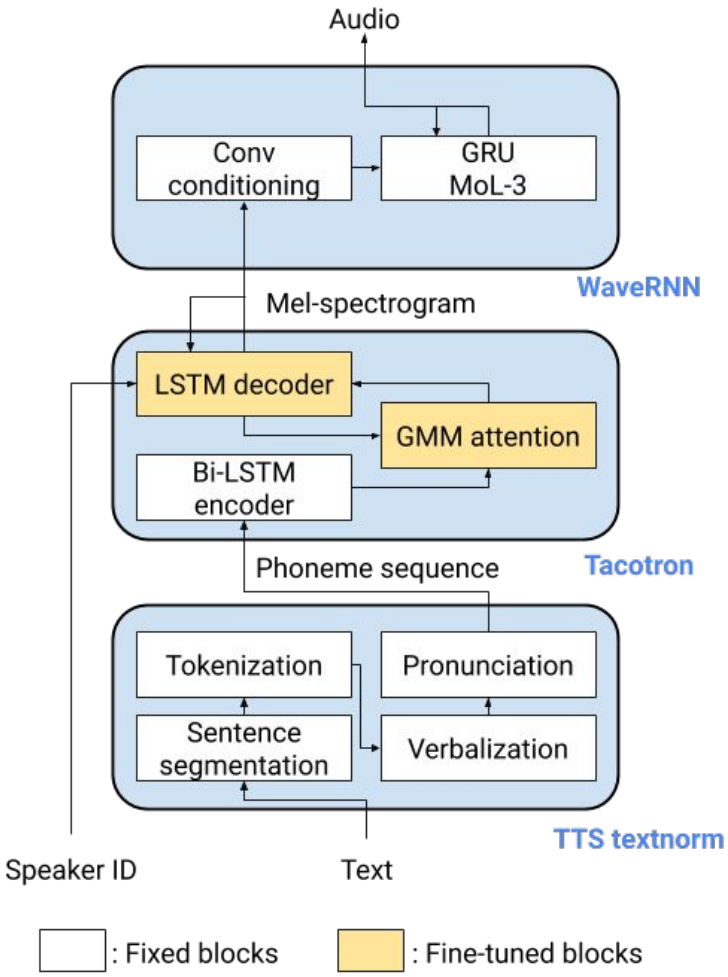
- Mixture of 3 Logistic distribution
- 24kHz, 16-bit sampling

Data

- Same as Tacotron (430h, ~1.1k speakers)
- Predicted spectra (teacher forcing) as input

Universal vocoder

- No fine-tuning / retraining
- One WaveRNN to serve multiple customers



Custom Voice Fine-tuning Evaluation

Fine-tune data : base data ratio (%)	Fine-tune steps	Naturalness MOS	Similarity MOS	Robustness failure (%) on short eval set	Robustness failure (%) on long eval set
100 : 0	2,000	**4.41 ± 0.04**	3.31 ± 0.07	0.0012	0.012
50 : 50	6,000	**4.41 ± 0.04**	3.37 ± 0.08	0.0006	0.007
10 : 90	40,000	**4.42 ± 0.04**	3.23 ± 0.08	**0.0003**	**0.005**

Highly natural synthetic speech w/ less failures

Working Flow

02

Customer Onboarding Process

1. Customer requests access to the Custom Voice feature via **this form**

   Only by customers w/ approved use cases as Google AI Principles

2. Customer hires a voice talent & records 30-min audio in a studio
3. Google trains a custom TTS model
4. Google sends samples of synthetic speech to Customer
5. Customer use API to generate samples for evaluation

Data Preparation (Customer Side)

Google provides recordings scripts (~30 mins)

- Full phoneme coverage (min 5 instances per phoneme)
- Easy to read
- 5,000 words + 500 word buffer
- Proofread before delivery

Customer records the scripts in a studio

- 30 min studio-quality recordings
- Provide guidance & specifications (e.g., SNR, loudness, microphone)

Packaging, Evaluation & User Acceptance Tests

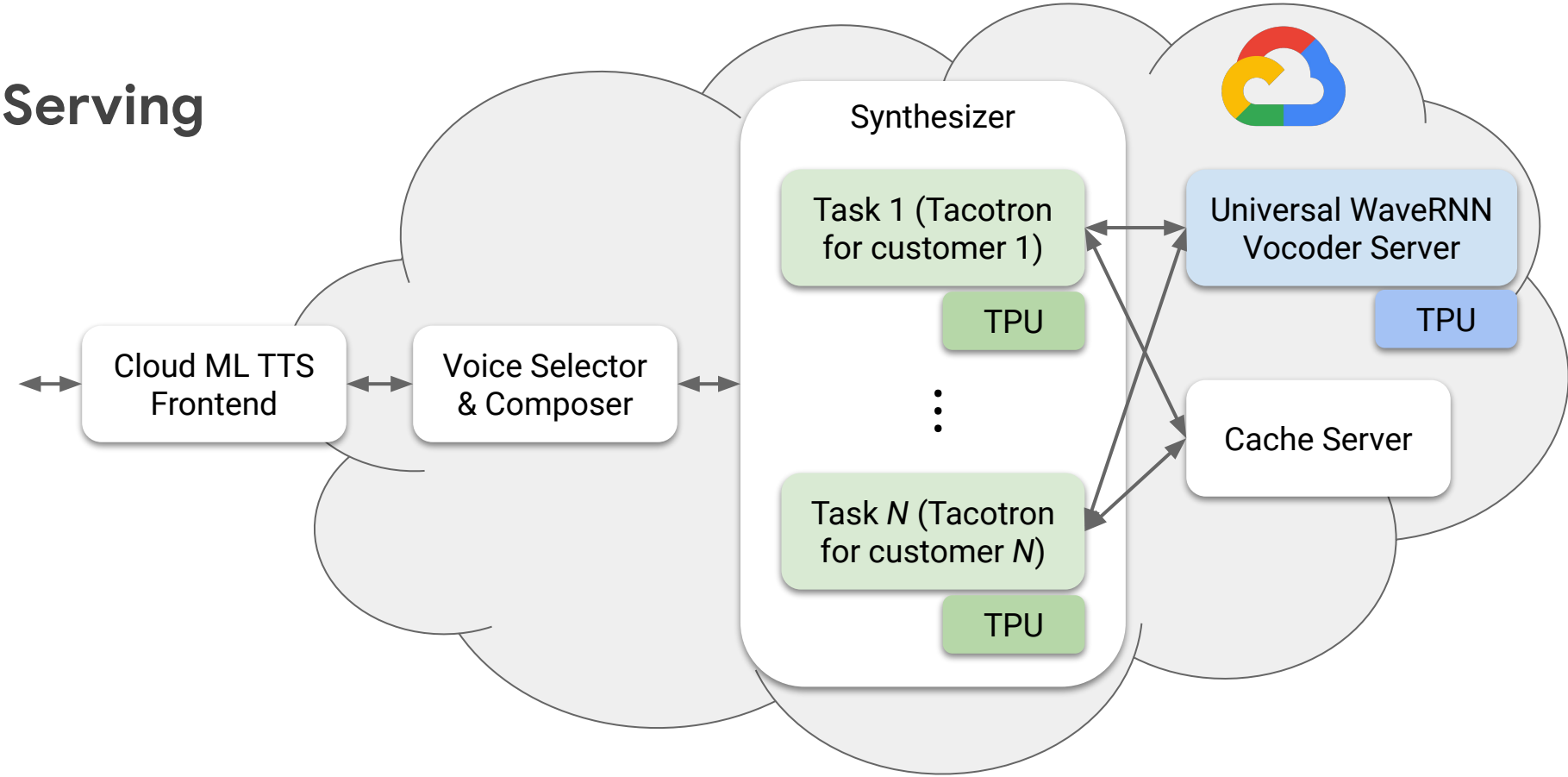
Google side

- Package fine-tuned model & proprietary text analysis front-end
- Run basic sanity checks (loudness, intelligibility, etc.)
- Run large-scale robustness eval
- Run subjective evaluation

Customer side

- Check off-line samples by customer
- Conducts user acceptance testing (UAT) process

Serving



Overall latency: ~500ms (w/o cache, depends on sentence length)

Custom Voice Examples

Female - Original Voice

Male - Original Voice

Female - Synthetic voice Example 1

Male - Synthetic voice Example 1

Female - Synthetic voice Example 2

Male - Synthetic voice Example 2

Conclusions

03

Cloud Text-to-Speech - Custom Voice (Beta)

- Build customer-specific voice from customer’s recordings
- Naturally sounding synthetic speech
- Serve voices on Google Cloud
- Access it via Cloud TTS API

Strong interests from many enterprise customers in various countries

Future Work

- Better synthetic speech (naturalness & similarity)
- Support expressiveness & styles
- More locales
- Lower latency

Custom Voice Site

https://cloud.google.com/text-to-speech/custom-voice/docs

Thank you