In-House Text to Speech

AI Transparency Technical Note

April 2026

webex by CISCO

Contents

At Cisco, we appreciate that Artificial Intelligence (AI) can be leveraged to power an inclusive future for all. We also recognize that by applying this technology, we have a responsibility to mitigate potential harm. That is why Cisco adheres to our Responsible AI Framework (the "Framework"), which is based on six principles of Transparency, Fairness, Accountability, Privacy, Security and Reliability. Cisco translates these principles into product development requirements, which ultimately form part of the product development lifecycle alongside our Security by Design, Privacy by Design, and Human Rights by Design processes.

Accordingly, Webex by Cisco (Webex) features that leverage AI are built with transparency, fairness, accountability, privacy, security, and reliability at their core. Each feature powered by AI undergoes an AI Impact (AII) Assessment – a best-in-class review of how the technical underpinnings of the functionality measure against the Framework Principles.

Cisco Text-to-Speech was built with the Framework Principles at the center of how we deliver AI-powered technology. This Technical Note describes more information about the feature and how Cisco responsibly leverages AI deliver the functionality.

Feature Overview

Webex Calling offers an optional feature that utilizes an announcement repository to save announcements for use across features like Auto Attendant, Call Queue, Customer Assist, and Music on Hold. Customers have the flexibility to choose between populating that repository with static audio generated via Cisco's Text-to-Speech (TTS) model, or by manually recording and uploading audio files into the announcement repository. TTS synthesizes speech from input text which produces natural sounding voices with minimal distortion or artifacts and eliminates the need for manual recording. Currently the text-to-speech model is only available in English, with other languages coming soon.

Model Overview

Introduction

Cisco Text-To-Speech is a Cisco-built machine learning model that generates audio from text. Text-to-Speech models are often used in interactive voice response systems and autonomous voice agents. With IVR systems, TTS generates dynamic audio prompts by synthesizing speech during a live call using a text input instead of a pre-recorded static audio prompt. For autonomous voice agents where conversational AI solutions rely on open-ended conversations, the responses generated from Natural Language Understanding (NLU) can be dynamic; therefore, require TTS to generate audio responses for the agent.

Model Architecture

Cisco's Text-to-Speech model is based on the latest text-to-speech research and trained on synthetic, high-quality data carefully to output natural speech. The model first phoneticizes the input text (i.e., converts the text into a sequence pf phonemes) and then synthesizes the obtained phoneme sequence. This two-step process helps to ensure that the resulting audio file contains speech corresponding exactly to the input text.

The figure below shows the data flow for Cisco's Text-to-Speech.

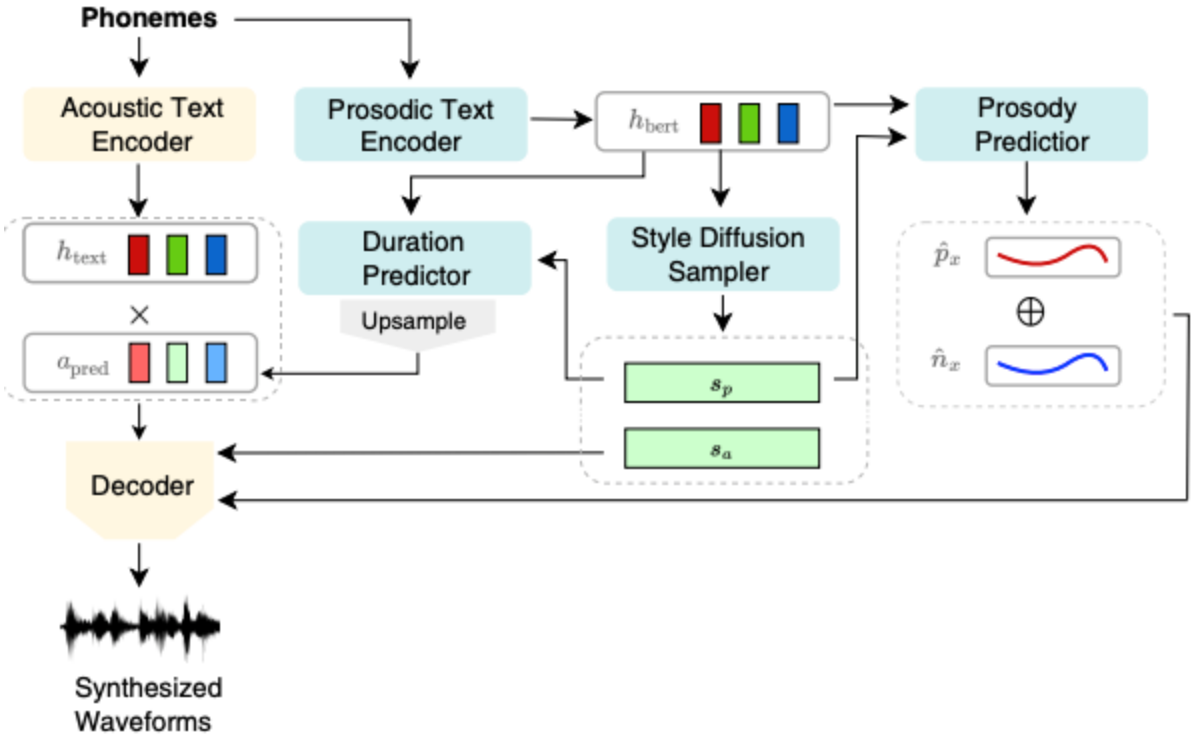

Figure 1: Text-to-Speech data flow

Usage Guidelines

Cisco's Text-to-Speech model and derived feature(s) cannot be used outside of Webex products.

Model Inputs and Outputs

Typical inputs to the Text-to-Speech model are the given text, and the Speech Synthesis Markup Language (SSML) tags for the desired speech output. The model allows users to choose between male and female voices and SSML tags like <prosody>, <say-as>, <emphasis>, and <break> to customize the audio output. Typical outputs of the model are generated natural sounding audio.

Data Sources for Training and Evaluation

The Text-to-Speech model is trained and evaluated on audio and text data. Sources of the data include off-the-shelf datasets, purchased custom recording data, and synthetically generated data carefully crafted to output natural speech. Cisco does not use customer data to train or fine tune its TTS model.

Model Evaluation and Performance

Cisco's TTS model is rigorously assessed by human and automated tests that evaluate and improve the performance, quality, and reliability of the model. The evaluation of the synthesized speech focuses on how natural and close to human speech they sound, the prosody and intonation, as well as the artifacts in the voice. Humans are involved with review, testing, and quality assurance including internal beta channels and external crowdsourcing evaluations.

Safety and Ethical Considerations

Cisco assessed the Text-to-Speech model and its use of AI against the Framework Principles by undertaking an AI Impact (AII) Assessment. Through that process, Cisco analyzed the use of the TTS model against safety and ethical considerations, including whether and to what extent the model mitigates bias, integrates fairness, and whether and to what extent it may compromise human rights.

Cisco evaluated the quality of the generated output through internal Beta channels and external crowdsourcing, while also performing thorough internal human assessments and Automatic Speech Recognition (ASR) tests. Please contact Cisco if you have concerns or feedback about your experience with the Text-to-Speech feature.

Fairness

Cisco continuously evaluates datasets and labeling techniques. As a result, the TTS model is expected to perform similarly across Webex products, if used, and for the intended purpose and identified user base, regardless of demographics.

Privacy and Security

Cisco addresses processing of and security around personal data, including retention periods for that data, in its Offer Disclosures, found on the Cisco Trust Portal. Cisco does not retain the input data or the output data from its TTS model.

Updates and Maintenance

Cisco routinely monitors and rebalances the training dataset that powers the Text-to-Speech model to address any issues that are encountered. The improvements Cisco makes based on those identified cases, including objective and subjective scores that measure the accuracy and performance of the model, are well documents.

Ongoing Webex updates and product release documents include product updates and any improvements in user experience. Cisco may also use customer email notifications and Help Center documentation to inform customers of product changes.

License

The Text-to-Speech model is proprietary and not publicly available. Cisco uses it only to deliver this feature in the context of Webex applications, Cisco devices, or with Webex SDKs.

References

RAI Principles

RAI Framework

Cisco Trust Portal