

# What is text to speech?


Article • 09/24/2024

In this overview, you learn about the benefits and capabilities of the text to speech feature of the Speech service, which is part of Azure AI services.

Text to speech enables your applications, tools, or devices to convert text into human like synthesized speech. The text to speech capability is also known as speech synthesis. Use human like prebuilt neural voices out of the box, or create a custom neural voice that's unique to your product or brand. For a full list of supported voices, languages, and locales, see [Language and voice support for the Speech service](#).

## Core features

Text to speech includes the following features:

 Expand table

Feature	Summary	Demo
Prebuilt neural voice (called <i>Neural</i> on the <a href="#">pricing page</a> )	Highly natural out-of-the-box voices. Create an Azure subscription and Speech resource, and then use the <a href="#">Speech SDK</a> or visit the <a href="#">Speech Studio portal</a> and select prebuilt neural voices to get started. Check the <a href="#">pricing details</a> .	Check the <a href="#">Voice Gallery</a> and determine the right voice for your business needs.
Custom neural voice (called <i>Custom Neural</i> on the <a href="#">pricing page</a> )	Easy-to-use self-service for creating a natural brand voice, with limited access for responsible use. Create an Azure subscription and Speech resource (with the S0 tier), and <a href="#">apply</a> to use the custom voice feature. After you're granted access, visit the <a href="#">Speech Studio portal</a> and select <b>Custom voice</b> to get started. Check the <a href="#">pricing details</a> .	Check the <a href="#">voice samples</a> .

## More about neural text to speech features

Text to speech uses deep neural networks to make the voices of computers nearly indistinguishable from the recordings of people. With the clear articulation of words, neural text to speech significantly reduces listening fatigue when users interact with AI systems.

The patterns of stress and intonation in spoken language are called *prosody*. Traditional text to speech systems break down prosody into separate linguistic analysis and acoustic prediction steps governed by independent models. That can result in muffled, buzzy voice synthesis.

Here's more information about neural text to speech features in the Speech service, and how they overcome the limits of traditional text to speech systems:

- **Real-time speech synthesis:** Use the [Speech SDK](#) or [REST API](#) to convert text to speech by using [prebuilt neural voices](#) or [custom neural voices](#).
- **Asynchronous synthesis of long audio:** Use the [batch synthesis API](#) to asynchronously synthesize text to speech files longer than 10 minutes (for example, audio books or lectures). Unlike synthesis performed via the Speech SDK or Speech to text REST API, responses aren't returned in real-time. The expectation is that requests are sent asynchronously, responses are polled for, and synthesized audio is downloaded when the service makes it available.
- **Prebuilt neural voices:** Azure AI Speech uses deep neural networks to overcome the limits of traditional speech synthesis regarding stress and intonation in spoken language. Prosody prediction and voice synthesis happen simultaneously, which results in more fluid and natural-sounding outputs. Each prebuilt neural voice model is available at 24 kHz and high-fidelity 48 kHz. You can use neural voices to:
  - Make interactions with chatbots and voice assistants more natural and engaging.
  - Convert digital texts such as e-books into audiobooks.
  - Enhance in-car navigation systems.

For a full list of platform neural voices, see [Language and voice support for the Speech service](#).

- **Improve text to speech output with SSML:** Speech Synthesis Markup Language (SSML) is an XML-based markup language used to customize text to speech outputs. With SSML, you can adjust pitch, add pauses, improve pronunciation, change speaking rate, adjust volume, and attribute multiple voices to a single document.

You can use SSML to define your own lexicons or switch to different speaking styles. With the [multilingual voices](#), you can also adjust the speaking languages via SSML. To improve the voice output for your scenario, see [Improve synthesis with Speech Synthesis Markup Language](#) and [Speech synthesis with the Audio Content Creation tool](#).

- **Visemes:** [Visemes](#) are the key poses in observed speech, including the position of the lips, jaw, and tongue in producing a particular phoneme. Visemes have a strong correlation with voices and phonemes.

By using viseme events in Speech SDK, you can generate facial animation data. This data can be used to animate faces in lip-reading communication, education, entertainment, and customer service. Viseme is currently supported only for the en-us (US English) [neural voices](#).

#### ⓘ Note

We plan to retire the traditional/standard voices and non-neural custom voice in 2024. After that, we'll no longer support them.

If your applications, tools, or products are using any of the standard voices and custom voices, you must migrate to the neural version. For more information, see [Migrate to neural voices](#).

## Get started

To get started with text to speech, see the [quickstart](#). Text to speech is available via the [Speech SDK](#), the [REST API](#), and the [Speech CLI](#).

#### 💡 Tip

To convert text to speech with a no-code approach, try the [Audio Content Creation](#) tool in [Speech Studio](#) .

## Sample code

Sample code for text to speech is available on GitHub. These samples cover text to speech conversion in most popular programming languages:

- [Text to speech samples \(SDK\)](#)
- [Text to speech samples \(REST\)](#)

# Custom neural voice

In addition to prebuilt neural voices, you can create custom neural voices that are unique to your product or brand. All it takes to get started is a handful of audio files and the associated transcriptions. For more information, see [Get started with custom neural voice](#).

## Pricing note

### Billable characters

When you use the text to speech feature, you're billed for each character that's converted to speech, including punctuation. Although the SSML document itself isn't billable, optional elements that are used to adjust how the text is converted to speech, like phonemes and pitch, are counted as billable characters. Here's a list of what's billable:

- Text passed to the text to speech feature in the SSML body of the request
- All markup within the text field of the request body in the SSML format, except for `< speak>` and `< voice>` tags
- Letters, punctuation, spaces, tabs, markup, and all white-space characters
- Every code point defined in Unicode

For detailed information, see [Speech service pricing](#).

#### Important

Each Chinese character is counted as two characters for billing, including kanji used in Japanese, hanja used in Korean, or hanzi used in other languages.

## Model training and hosting time for custom neural voice

Custom neural voice training and hosting are both calculated by hour and billed per second. For the billing unit price, see [Speech service pricing](#).

Custom neural voice (CNV) training time is measured by 'compute hour' (a unit to measure machine running time). Typically, when training a voice model, two computing tasks are running in parallel. So, the calculated compute hours are longer than the actual training time. On average, it takes less than one compute hour to train a CNV Lite voice; while for

CNV Pro, it usually takes 20 to 40 compute hours to train a single-style voice, and around 90 compute hours to train a multi-style voice. The CNV training time is billed with a cap of 96 compute hours. So in the case that a voice model is trained in 98 compute hours, you'll only be charged with 96 compute hours.

Custom neural voice (CNV) endpoint hosting is measured by the actual time (hour). The hosting time (hours) for each endpoint is calculated at 00:00 UTC every day for the previous 24 hours. For example, if the endpoint has been active for 24 hours on day one, it's billed for 24 hours at 00:00 UTC the second day. If the endpoint is newly created or suspended during the day, it's billed for its accumulated running time until 00:00 UTC the second day. If the endpoint isn't currently hosted, it isn't billed. In addition to the daily calculation at 00:00 UTC each day, the billing is also triggered immediately when an endpoint is deleted or suspended. For example, for an endpoint created at 08:00 UTC on December 1, the hosting hour will be calculated to 16 hours at 00:00 UTC on December 2 and 24 hours at 00:00 UTC on December 3. If the user suspends hosting the endpoint at 16:30 UTC on December 3, the duration (16.5 hours) from 00:00 to 16:30 UTC on December 3 will be calculated for billing.

## Personal voice

When you use the personal voice feature, you're billed for both profile storage and synthesis.

- **Profile storage:** After a personal voice profile is created, it will be billed until it is removed from the system. The billing unit is per voice per day. If voice storage lasts for a period of less than 24 hours, it will be billed as one full day.
- **Synthesis:** Billed per character. For details on billable characters, see the above [billable characters](#).

## Text to speech avatar

When using the text-to-speech avatar feature, charges will be incurred based on the length of video output and will be billed per second. However, for the real-time avatar, charges are based on the time when the avatar is active, regardless of whether it is speaking or remaining silent, and will also be billed per second. To optimize costs for real-time avatar usage, refer to the tips provided in the [sample code](#) (search "Use Local Video for Idle"). Avatar hosting is billed per second per endpoint. You can suspend your endpoint to save

costs. If you want to suspend your endpoint, you can delete it directly. To use it again, simply redeploy the endpoint.

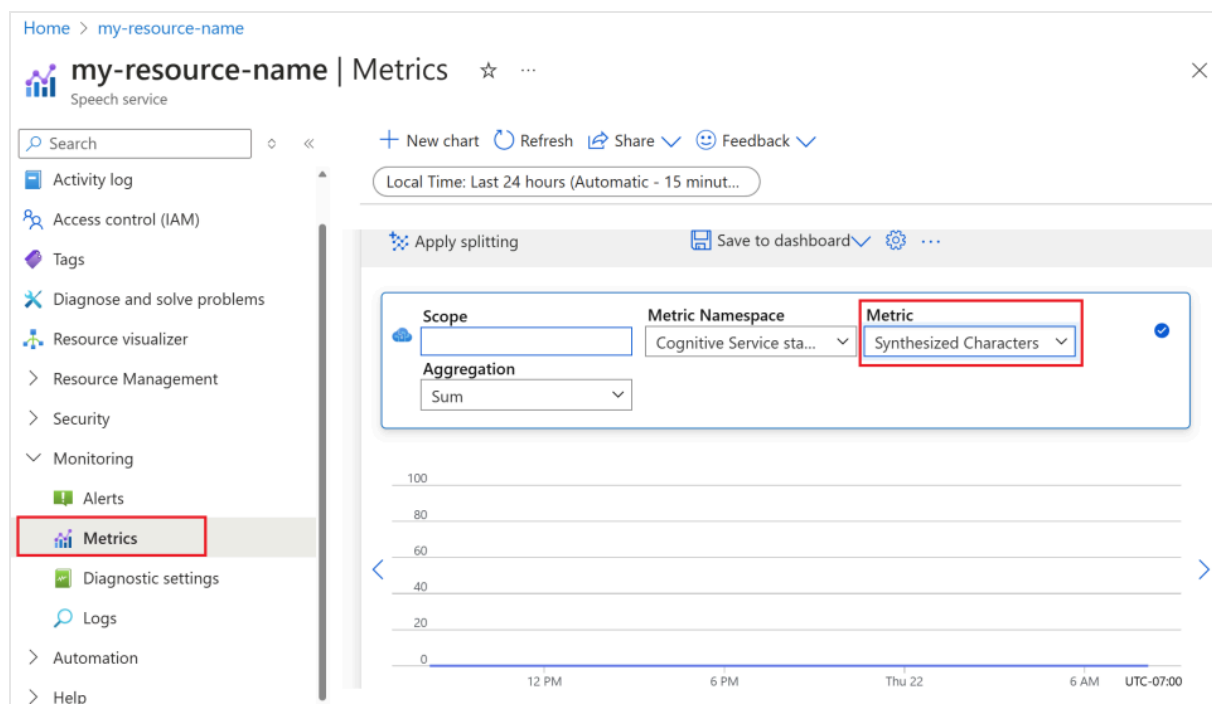
## Monitor Azure text to speech metrics

Monitoring key metrics associated with text to speech services is crucial for managing resource usage and controlling costs. This section will guide you on how to find usage information in the Azure portal and provide detailed definitions of the key metrics. For more details on Azure monitor metrics, refer to [Azure Monitor Metrics overview](#).

### How to find usage information in the Azure portal

To effectively manage your Azure resources, it's essential to access and review usage information regularly. Here's how to find the usage information:

1. Go to the [Azure portal](#) and sign in with your Azure account.
2. Navigate to **Resources** and select your resource you wish to monitor.
3. Select **Metrics** under **Monitoring** from the left-hand menu.



4. Customize metric views.

You can filter data by resource type, metric type, time range, and other parameters to create custom views that align with your monitoring needs. Additionally, you can save


the metric view to dashboards by selecting **Save to dashboard** for easy access to frequently used metrics.

5. Set up alerts.

To manage usage more effectively, set up alerts by navigating to the **Alerts** tab under **Monitoring** from the left-hand menu. Alerts can notify you when your usage reaches specific thresholds, helping to prevent unexpected costs.

# Definition of metrics

Below is a table summarizing the key metrics for Azure text to speech services.

 **Expand table**

Metric name	Description
Synthesized Characters	Tracks the number of characters converted into speech, including prebuilt neural voice and custom neural voice. For details on billable characters, see <a href="#">Billable characters</a> .
Video Seconds Synthesized	Measures the total duration of video synthesized, including batch avatar synthesis, real-time avatar synthesis, and custom avatar synthesis.
Avatar Model Hosting Seconds	Tracks the total time in seconds that your custom avatar model is hosted.
Voice Model Hosting Hours	Tracks the total time in hours that your custom neural voice model is hosted.
Voice Model Training Minutes	Measures the total time in minutes for training your custom neural voice model.

# Reference docs

- [Speech SDK](#)
- [REST API: Text to speech](#)

# Responsible AI

An AI system includes not only the technology, but also the people who use it, the people who are affected by it, and the environment in which it's deployed. Read the transparency notes to learn about responsible AI use and deployment in your systems.

- [Transparency note and use cases for custom neural voice](#)
- [Characteristics and limitations for using custom neural voice](#)
- [Limited access to custom neural voice](#)
- [Guidelines for responsible deployment of synthetic voice technology](#)
- [Disclosure for voice talent](#)
- [Disclosure design guidelines](#)
- [Disclosure design patterns](#)
- [Code of Conduct for Text to speech integrations](#)
- [Data, privacy, and security for custom neural voice](#)

## Next steps

- [Text to speech quickstart](#)
- [Get the Speech SDK](#)

---

## Feedback

Was this page helpful?

 Yes

 No

[Provide product feedback](#) | [Get help at Microsoft Q&A](#)