Customize voice and sound with SSML

Article • 09/24/2024

You can use Speech Synthesis Markup Language (SSML) to specify the text to speech voice, language, name, style, and role for your speech output. You can also use multiple voices in a single SSML document, and adjust the emphasis, speaking rate, pitch, and volume. In addition, SSML features the ability to insert prerecorded audio, such as a sound effect or a musical note.

The article shows you how to use SSML elements to specify voice and sound. For more information about SSML syntax, see SSML document structure and events.

Use voice elements

At least one `voice` element must be specified within each SSML speak element. This element determines the voice that's used for text to speech.

You can include multiple `voice` elements in a single SSML document. Each `voice` element can specify a different voice. You can also use the same voice multiple times with different settings, such as when you change the silence duration between sentences.

The following table describes the usage of the `voice` element's attributes:

Expand table

Attribute	Description	Required or optional
`name`	The voice used for text to speech output. For a complete list of supported prebuilt voices, see Language support.	Required
`effect`	The audio effect processor that's used to optimize the quality of the synthesized speech output for specific scenarios on devices. For some scenarios in production environments, the auditory experience might be degraded due to the playback distortion on certain devices. For example, the synthesized speech from a car speaker might sound dull and muffled due to environmental factors such as speaker response, room reverberation, and background noise. The passenger might have to turn up the volume to hear more clearly. To avoid manual operations in such a	Optional

Attribute	Description	Required or optional
	scenario, the audio effect processor can make the sound clearer by compensating the distortion of playback. The following values are supported: • `eq_car` – Optimize the auditory experience when providing high-fidelity speech in cars, buses, and other enclosed automobiles. • `eq_telecomhp8k` – Optimize the auditory experience for narrowband speech in telecom or telephone scenarios. You should use a sampling rate of 8 kHz. If the sample rate isn't 8 kHz, the auditory quality of the output speech isn't optimized. If the value is missing or invalid, this attribute is ignored and no effect is applied.	

Voice examples

For information about the supported values for attributes of the `voice` element, see Use voice elements.

Single voice example

This example uses the `en-US-AvaMultilingualNeural` voice.

XML

```xml
<speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis" xml:lang="en-US">
    <voice name="en-US-AvaMultilingualNeural">
        This is the text that is spoken.
    </voice>
</speak>
```

Multiple voices example

Within the `speak` element, you can specify multiple voices for text to speech output. These voices can be in different languages. For each voice, the text must be wrapped in a `voice` element.

This example alternates between the `en-US-AvaMultilingualNeural` and `en-US-AndrewMultilingualNeural` voices. The neural multilingual voices can speak different languages based on the input text.

XML

```xml
<speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis" xml:lang="en-US">
    <voice name="en-US-AvaMultilingualNeural">
        Good morning!
    </voice>
    <voice name="en-US-AndrewMultilingualNeural">
        Good morning to you too Ava!
    </voice>
</speak>
```

Custom neural voice example

To use your custom neural voice, specify the model name as the voice name in SSML.

This example uses a custom voice named **my-custom-voice**.

XML

```xml
<speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis" xml:lang="en-US">
    <voice name="my-custom-voice">
        This is the text that is spoken.
    </voice>
</speak>
```

Audio effect example

You use the `effect` attribute to optimize the auditory experience for scenarios such as cars and telecommunications. The following SSML example uses the `effect` attribute with the configuration in car scenarios.

XML

```
<speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis" xml:lang="en-US">
    <voice name="en-US-AvaMultilingualNeural" effect="eq_car">
        This is the text that is spoken.
    </voice>
</speak>
```

Use speaking styles and roles

By default, neural voices have a neutral speaking style. You can adjust the speaking style, style degree, and role at the sentence level.

ⓘ Note

The Speech service supports styles, style degree, and roles for a subset of neural voices as described in the voice styles and roles documentation. To determine the supported styles and roles for each voice, you can also use the list voices API and the audio content creation web application.

The following table describes the usage of the `mstts:express-as` element's attributes:

Expand table

Attribute	Description	Required or optional
`style`	The voice-specific speaking style. You can express emotions like cheerfulness, empathy, and calmness. You can also optimize the voice for different scenarios like customer service, newscast, and voice assistant. If the style value is missing or invalid, the entire `mstts:express-as` element is ignored and the service uses the default neutral speech. For custom neural voice styles, see the custom neural voice style example.	Required
`styledegree`	The intensity of the speaking style. You can specify a stronger or softer style to make the speech more expressive or subdued. The range of accepted values are: `0.01` to `2` inclusive. The default value is `1`, which means the predefined style intensity. The minimum unit is `0.01`, which results in a slight tendency for the target style. A value of `2` results in a	Optional

Attribute	Description	Required or optional
	doubling of the default style intensity. If the style degree is missing or isn't supported for your voice, this attribute is ignored.	
`role`	The speaking role-play. The voice can imitate a different age and gender, but the voice name isn't changed. For example, a male voice can raise the pitch and change the intonation to imitate a female voice, but the voice name isn't changed. If the role is missing or isn't supported for your voice, this attribute is ignored.	Optional

The following table describes each supported `style` attribute:

Expand table

Style	Description
`style="advertisement_upbeat"`	Expresses an excited and high-energy tone for promoting a product or service.
`style="affectionate"`	Expresses a warm and affectionate tone, with higher pitch and vocal energy. The speaker is in a state of attracting the attention of the listener. The personality of the speaker is often endearing in nature.
`style="angry"`	Expresses an angry and annoyed tone.
`style="assistant"`	Expresses a warm and relaxed tone for digital assistants.
`style="calm"`	Expresses a cool, collected, and composed attitude when speaking. Tone, pitch, and prosody are more uniform compared to other types of speech.
`style="chat"`	Expresses a casual and relaxed tone.
`style="cheerful"`	Expresses a positive and happy tone.
`style="customerservice"`	Expresses a friendly and helpful tone for customer support.
`style="depressed"`	Expresses a melancholic and despondent tone with lower pitch and energy.
`style="disgruntled"`	Expresses a disdainful and complaining tone. Speech of this emotion displays displeasure and contempt.

Style	Description
`style="documentary-narration"`	Narrates documentaries in a relaxed, interested, and informative style suitable for documentaries, expert commentary, and similar content.
`style="embarrassed"`	Expresses an uncertain and hesitant tone when the speaker is feeling uncomfortable.
`style="empathetic"`	Expresses a sense of caring and understanding.
`style="envious"`	Expresses a tone of admiration when you desire something that someone else has.
`style="excited"`	Expresses an upbeat and hopeful tone. It sounds like something great is happening and the speaker is happy about it.
`style="fearful"`	Expresses a scared and nervous tone, with higher pitch, higher vocal energy, and faster rate. The speaker is in a state of tension and unease.
`style="friendly"`	Expresses a pleasant, inviting, and warm tone. It sounds sincere and caring.
`style="gentle"`	Expresses a mild, polite, and pleasant tone, with lower pitch and vocal energy.
`style="hopeful"`	Expresses a warm and yearning tone. It sounds like something good will happen to the speaker.
`style="lyrical"`	Expresses emotions in a melodic and sentimental way.
`style="narration-professional"`	Expresses a professional, objective tone for content reading.
`style="narration-relaxed"`	Expresses a soothing and melodious tone for content reading.
`style="newscast"`	Expresses a formal and professional tone for narrating news.
`style="newscast-casual"`	Expresses a versatile and casual tone for general news delivery.
`style="newscast-formal"`	Expresses a formal, confident, and authoritative tone for news delivery.
`style="poetry-reading"`	Expresses an emotional and rhythmic tone while reading a poem.
`style="sad"`	Expresses a sorrowful tone.

Style	Description
`style="serious"`	Expresses a strict and commanding tone. Speaker often sounds stiffer and much less relaxed with firm cadence.
`style="shouting"`	Expresses a tone that sounds as if the voice is distant or in another location and making an effort to be clearly heard.
`style="sports_commentary"`	Expresses a relaxed and interested tone for broadcasting a sports event.
`style="sports_commentary_excited"`	Expresses an intensive and energetic tone for broadcasting exciting moments in a sports event.
`style="whispering"`	Expresses a soft tone that's trying to make a quiet and gentle sound.
`style="terrified"`	Expresses a scared tone, with a faster pace and a shakier voice. It sounds like the speaker is in an unsteady and frantic status.
`style="unfriendly"`	Expresses a cold and indifferent tone.

The following table has descriptions of each supported `role` attribute:

Expand table

Role	Description
`role="Girl"`	The voice imitates a girl.
`role="Boy"`	The voice imitates a boy.
`role="YoungAdultFemale"`	The voice imitates a young adult female.
`role="YoungAdultMale"`	The voice imitates a young adult male.
`role="OlderAdultFemale"`	The voice imitates an older adult female.
`role="OlderAdultMale"`	The voice imitates an older adult male.
`role="SeniorFemale"`	The voice imitates a senior female.
`role="SeniorMale"`	The voice imitates a senior male.

mstts express-as examples

For information about the supported values for attributes of the `mstts:express-as` element, see Use speaking styles and roles.

Style and degree example

You use the `mstts:express-as` element to express emotions like cheerfulness, empathy, and calm. You can also optimize the voice for different scenarios like customer service, newscast, and voice assistant.

The following SSML example uses the `<mstts:express-as>` element with a `sad` style degree of `2`.

XML

```xml
<speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis"
xmlns:mstts="https://www.w3.org/2001/mstts" xml:lang="zh-CN">
    <voice name="zh-CN-XiaomoNeural">
        <mstts:express-as style="sad" styledegree="2">
            快走吧，路上一定要注意安全，早去早回。
        </mstts:express-as>
    </voice>
</speak>
```

Role example

Apart from adjusting the speaking styles and style degree, you can also adjust the `role` parameter so that the voice imitates a different age and gender. For example, a male voice can raise the pitch and change the intonation to imitate a female voice, but the voice name isn't changed.

This SSML snippet illustrates how the `role` attribute is used to change the role-play for `zh-CN-XiaomoNeural`.

XML

```xml
<speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis"
xmlns:mstts="https://www.w3.org/2001/mstts" xml:lang="zh-CN">
    <voice name="zh-CN-XiaomoNeural">
        女儿看见父亲走了进来，问道：
        <mstts:express-as role="YoungAdultFemale" style="calm">
            “您来的挺快的，怎么过来的？”
        </mstts:express-as>
```

```
            父亲放下手提包，说：
            <mstts:express-as role="OlderAdultMale" style="calm">
                “刚打车过来的，路上还挺顺畅。”
            </mstts:express-as>
        </voice>
    </speak>
```

Custom neural voice style example

You can train your custom neural voice to speak with some preset styles such as `cheerful`, `sad`, and `whispering`. You can also train a custom neural voice to speak in a custom style as determined by your training data. To use your custom neural voice style in SSML, specify the style name that you previously entered in Speech Studio.

This example uses a custom voice named **my-custom-voice**. The custom voice speaks with the `cheerful` preset style and style degree of `2`, and then with a custom style named **my-custom-style** and style degree of `0.01`.

XML

```xml
<speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis"
xmlns:mstts="https://www.w3.org/2001/mstts" xml:lang="en-US">
    <voice name="my-custom-voice">
        <mstts:express-as style="cheerful" styledegree="2">
            That'd be just amazing!
        </mstts:express-as>
        <mstts:express-as style="my-custom-style" styledegree="0.01">
            What's next?
        </mstts:express-as>
    </voice>
</speak>
```

Speaker profile ID

You use the `mstts:ttsembedding` element to specify the `speakerProfileId` property for a personal voice. Personal voice is a custom neural voice that's trained on your own voice or your customer's voice. For more information, see create a personal voice.

The following SSML example uses the `<mstts:ttsembedding>` element with a voice name and speaker profile ID.

XML

```xml
<speak version='1.0' xmlns='http://www.w3.org/2001/10/synthesis'
xmlns:mstts='http://www.w3.org/2001/mstts' xml:lang='en-US'>
    <voice xml:lang='en-US' xml:gender='Male' name='PhoenixV2Neural'>
    <mstts:ttsembedding speakerProfileId='your speaker profile ID here'>
    I'm happy to hear that you find me amazing and that I have made your trip
planning easier and more fun. 我很高兴听到你觉得我很了不起，我让你的旅行计划更轻松、
更有趣。Je suis heureux d'apprendre que vous me trouvez incroyable et que j'ai
rendu la planification de votre voyage plus facile et plus amusante.
    </mstts:ttsembedding>
    </voice>
</speak>
```

Adjust speaking languages

By default, multilingual voices can autodetect the language of the input text and speak in the language of the default locale of the input text without using SSML. Optionally, you can use the `<lang xml:lang>` element to adjust the speaking language for these voices to set the preferred accent such as `en-GB` for British English. You can adjust the speaking language at both the sentence level and word level. For information about the supported languages for multilingual voice, see Multilingual voices with the lang element for a table showing the `<lang>` syntax and attribute definitions.

The following table describes the usage of the `<lang xml:lang>` element's attributes:

Expand table

Attribute	Description	Required or optional
`xml:lang`	The language that you want the neural voice to speak.	Required to adjust the speaking language for the neural voice. If you're using `lang xml:lang`, the locale must be provided.

Note

The `<lang xml:lang>` element is incompatible with the `prosody` and `break` elements. You can't adjust pause and prosody like pitch, contour, rate, or volume in this element.

Non-multilingual voices don't support the `<lang xml:lang>` element by design.

Multilingual voices with the lang element

Use the multilingual voices section to determine which speaking languages the Speech service supports for each neural voice, as demonstrated in the following example table. If the voice doesn't speak the language of the input text, the Speech service doesn't output synthesized audio.

Expand table

Voice	Auto-detected language number	Auto-detected language (locale)	All locales number	All languages (locale) supported from SSML
`en-US-AndrewMultilingualNeural`[1] (Male) `en-US-AvaMultilingualNeural`[1] (Female) `en-US-BrianMultilingualNeural`[1] (Male) `en-US-EmmaMultilingualNeural`[1] (Female)	77	Afrikaans (`af-ZA`), Albanian (`sq-AL`), Amharic (`am-ET`), Arabic (`ar-EG`), Armenian (`hy-AM`), Azerbaijani (`az-AZ`), Bahasa Indonesian (`id-ID`), Bangla (`bn-BD`), Basque (`eu-ES`), Bengali (`bn-IN`), Bosnian (`bs-BA`), Bulgarian (`bg-BG`), Burmese (`my-MM`), Catalan (`ca-ES`), Chinese Cantonese (`zh-HK`), Chinese Mandarin (`zh-CN`), Chinese Taiwanese (`zh-TW`), Croatian (`hr-HR`), Czech (`cs-CZ`), Danish (`da-DK`), Dutch (`nl-NL`), English (`en-US`), Estonian (`et-EE`), Filipino (`fil-PH`), Finnish (`fi-FI`), French (`fr-FR`), Galician (`gl-ES`),	91	Afrikaans (South Africa) (`af-ZA`), Albanian (Albania) (`sq-AL`), Amharic (Ethiopia) (`am-ET`), Arabic (Egypt) (`ar-EG`), Arabic (Saudi Arabia) (`ar-SA`), Armenian (Armenia) (`hy-AM`), Azerbaijani (Azerbaijan) (`az-AZ`), Basque (Basque) (`eu-ES`), Bengali (India) (`bn-IN`), Bosnian (Bosnia and Herzegovina) (`bs-BA`), Bulgarian (Bulgaria) (`bg-BG`), Burmese (Myanmar) (`my-MM`), Catalan (Spain) (`ca-ES`), Chinese (Cantonese, Traditional) (`zh-HK`), Chinese (Mandarin, Simplified) (`zh-CN`), Chinese (Taiwanese Mandarin) (`zh-TW`), Croatian (Croatia) (`hr-HR`), Czech (Czech) (`cs-CZ`), Danish (Denmark) (`da-DK`), Dutch (Belgium) (`nl-BE`), Dutch (Netherlands) (`nl-NL`), English (Australia) (`en-AU`), English (Canada) (`en-CA`),

Voice	Auto-detected language number	Auto-detected language (locale)	All locales number	All languages (locale) supported from SSML
		Georgian (`ka-GE`), German (`de-DE`), Greek (`el-GR`), Hebrew (`he-IL`), Hindi (`hi-IN`), Hungarian (`hu-HU`), Icelandic (`is-IS`), Irish (`ga-IE`), Italian (`it-IT`), Japanese (`ja-JP`), Javanese (`jv-ID`), Kannada (`kn-IN`), Kazakh (`kk-KZ`), Khmer (`km-KH`), Korean (`ko-KR`), Lao (`lo-LA`), Latvian (`lv-LV`), Lithuanian (`lt-LT`), Macedonian (`mk-MK`), Malay (`ms-MY`), Malayalam (`ml-IN`), Maltese (`mt-MT`), Mongolian (`mn-MN`), Nepali (`ne-NP`), Norwegian Bokmål (`nb-NO`), Pashto (`ps-AF`), Persian (`fa-IR`), Polish (`pl-PL`), Portuguese (`pt-BR`), Romanian (`ro-RO`), Russian (`ru-RU`), Serbian (`sr-RS`), Sinhala (`si-LK`), Slovak (`sk-SK`), Slovene (`sl-SI`), Somali (`so-SO`), Spanish (`es-ES`), Sundanese (`su-`		English (Hong Kong SAR) (`en-HK`), English (India) (`en-IN`), English (Ireland) (`en-IE`), English (United Kingdom) (`en-GB`), English (United States) (`en-US`), Estonian (Estonia) (`et-EE`), Filipino (Philippines) (`fil-PH`), Finnish (Finland) (`fi-FI`), French (Belgium) (`fr-BE`), French (Canada) (`fr-CA`), French (France) (`fr-FR`), French (Switzerland) (`fr-CH`), Galician (Galician) (`gl-ES`), Georgian (Georgia) (`ka-GE`), German (Austria) (`de-AT`), German (Germany) (`de-DE`), German (Switzerland) (`de-CH`), Greek (Greece) (`el-GR`), Hebrew (Israel) (`he-IL`), Hindi (India) (`hi-IN`), Hungarian (Hungary) (`hu-HU`), Icelandic (Iceland) (`is-IS`), Indonesian (Indonesia) (`id-ID`), Irish (Ireland) (`ga-IE`), Italian (Italy) (`it-IT`), Japanese (Japan) (`ja-JP`), Javanese (Indonesia) (`jv-ID`), Kannada (India) (`kn-IN`), Kazakh (Kazakhstan) (`kk-KZ`), Khmer (Cambodia) (`km-KH`), Korean (Korea) (`ko-KR`), Lao (Laos) (`lo-LA`), Latvian (Latvia) (`lv-LV`), Lithuanian (Lithuania) (`lt-LT`), Macedonian (North Macedonia) (`mk-MK`), Malay (Malaysia) (`ms-`

Voice	Auto-detected language number	Auto-detected language (locale)	All locales number	All languages (locale) supported from SSML
		`ID`), Swahili (`sw-KE`), Swedish (`sv-SE`), Tamil (`ta-IN`), Telugu (`te-IN`), Thai (`th-TH`), Turkish (`tr-TR`), Ukrainian (`uk-UA`), Urdu (`ur-PK`), Uzbek (`uz-UZ`), Vietnamese (`vi-VN`), Welsh (`cy-GB`), Zulu (`zu-ZA`)		`MY`), Malayalam (India) (`ml-IN`), Maltese (Malta) (`mt-MT`), Mongolian (Mongolia) (`mn-MN`), Nepali (Nepal) (`ne-NP`), Norwegian (Bokmål, Norway) (`nb-NO`), Pashto (Afghanistan) (`ps-AF`), Persian (Iran) (`fa-IR`), Polish (Poland) (`pl-PL`), Portuguese (Brazil) (`pt-BR`), Portuguese (Portugal) (`pt-PT`), Romanian (Romania) (`ro-RO`), Russian (Russia) (`ru-RU`), Serbian (Cyrillic, Serbia) (`sr-RS`), Sinhala (Sri Lanka) (`si-LK`), Slovak (Slovakia) (`sk-SK`), Slovenian (Slovenia) (`sl-SI`), Somali (Somalia) (`so-SO`), Spanish (Mexico) (`es-MX`), Spanish (Spain) (`es-ES`), Sundanese (Indonesia) (`su-ID`), Swahili (Kenya) (`sw-KE`), Swedish (Sweden) (`sv-SE`), Tamil (India) (`ta-IN`), Telugu (India) (`te-IN`), Thai (Thailand) (`th-TH`), Turkish (Türkiye) (`tr-TR`), Ukrainian (Ukraine) (`uk-UA`), Urdu (Pakistan) (`ur-PK`), Uzbek (Uzbekistan) (`uz-UZ`), Vietnamese (Vietnam) (`vi-VN`), Welsh (United Kingdom) (`cy-GB`), Zulu (South Africa) (`zu-ZA`)

[1] Those are neural multilingual voices in Azure AI Speech. All multilingual voices can speak in the language in default locale of the input text without using SSML. However, you can still use the `<lang xml:lang>` element to adjust the speaking accent of each language to set preferred accent such as British accent (`en-GB`) for English. The primary locale for each voice is indicated by the prefix in its name, such as the voice `en-US-AndrewMultilingualNeural`, its primary locale is `en-US`.

> ⓘ **Note**
>
> Multilingual voices don't fully support certain SSML elements, such as `break`, `emphasis`, `silence`, and `sub`.

Lang examples

For information about the supported values for attributes of the `lang` element, see Adjust speaking language.

You must specify `en-US` as the default language within the `speak` element, whether or not the language is adjusted elsewhere. In this example, the primary language for `en-US-AvaMultilingualNeural` is `en-US`.

This SSML snippet shows how to use `<lang xml:lang>` to speak `de-DE` with the `en-US-AvaMultilingualNeural` neural voice.

XML

```xml
<speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis"
xmlns:mstts="https://www.w3.org/2001/mstts" xml:lang="en-US">
    <voice name="en-US-AvaMultilingualNeural">
        <lang xml:lang="de-DE">
            Wir freuen uns auf die Zusammenarbeit mit Ihnen!
        </lang>
    </voice>
</speak>
```

Within the `speak` element, you can specify multiple languages including `en-US` for text to speech output. For each adjusted language, the text must match the language and be wrapped in a `voice` element. This SSML snippet shows how to use `<lang xml:lang>` to change the speaking languages to `es-MX`, `en-US`, and `fr-FR`.

XML

```xml
<speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis"
xmlns:mstts="https://www.w3.org/2001/mstts" xml:lang="en-US">
    <voice name="en-US-AvaMultilingualNeural">
        <lang xml:lang="es-MX">
            ¡Esperamos trabajar con usted!
        </lang>
        <lang xml:lang="en-US">
           We look forward to working with you!
        </lang>
        <lang xml:lang="fr-FR">
            Nous avons hâte de travailler avec vous!
        </lang>
    </voice>
</speak>
```

Adjust prosody

You can use the `prosody` element to specify changes to pitch, contour, range, rate, and volume for the text to speech output. The `prosody` element can contain text and the following elements: `audio`, `break`, `p`, `phoneme`, `prosody`, `say-as`, `sub`, and `s`.

Because prosodic attribute values can vary over a wide range, the speech recognizer interprets the assigned values as a suggestion of what the actual prosodic values of the selected voice should be. Text to speech limits or substitutes values that aren't supported. Examples of unsupported values are a pitch of 1 MHz or a volume of 120.

The following table describes the usage of the `prosody` element's attributes:

Expand table

Attribute	Description	Required or optional
`contour`	Contour represents changes in pitch. These changes are represented as an array of targets at specified time positions in the speech output. Sets of parameter pairs define each target. For example: `<prosody contour="(0%,+20Hz) (10%,-2st) (40%,+10Hz)">` The first value in each set of parameters specifies the location of the pitch	Optional

Attribute	Description	Required or optional
	change as a percentage of the duration of the text. The second value specifies the amount to raise or lower the pitch by using a relative value or an enumeration value for pitch (see `pitch`). Pitch contour doesn't work on single words and short phrases. It is recommended to adjust the pitch contour on whole sentences or long phrases.	
`pitch`	Indicates the baseline pitch for the text. Pitch changes can be applied at the sentence level. The pitch changes should be within 0.5 to 1.5 times the original audio. You can express the pitch as: • An absolute value: Expressed as a number followed by "Hz" (Hertz). For example, `<prosody pitch="600Hz">some text</prosody>.` • A relative value: ○ As a relative number: Expressed as a number preceded by "+" or "-" and followed by "Hz" or "st" that specifies an amount to change the pitch. For example: `<prosody pitch="+80Hz">some text</prosody>` or `<prosody pitch="-2st">some text</prosody>.` The "st" indicates the change unit is semitone, which is half of a tone (a half step) on the standard diatonic scale. ○ As a percentage: Expressed as a number preceded by "+" (optionally) or "-" and followed by "%", indicating the relative change. For example: `<prosody pitch="50%">some text</prosody>` or `<prosody pitch="-50%">some text</prosody>.` • A constant value: ○ `x-low` (equivalently 0.55,-45%) ○ `low` (equivalently 0.8, -20%) ○ `medium` (equivalently 1, default value) ○ `high` (equivalently 1.2, +20%) ○ `x-high` (equivalently 1.45, +45%)	Optional
`range`	A value that represents the range of pitch for the text. You can express `range` by using the same absolute values, relative values, or enumeration values used to describe `pitch`.	Optional
`rate`	Indicates the speaking rate of the text. Speaking rate can be applied at the word or sentence level. The rate changes should be within `0.5` to `2` times the original audio. You can express `rate` as: • A relative value: ○ As a relative number: Expressed as a number that acts as a multiplier of the default. For example, a value of `1` results in no change in the original rate. A value of `0.5` results in a halving of the original rate. A value of `2` results in twice the original rate.	Optional

Attribute	Description	Required or optional
	◦ As a percentage: Expressed as a number preceded by "+" (optionally) or "-" and followed by "%", indicating the relative change. For example: `<prosody rate="50%">some text</prosody>` or `<prosody rate="-50%">some text</prosody>`. • A constant value: ◦ `x-slow` (equivalently 0.5, -50%) ◦ `slow` (equivalently 0.64, -46%) ◦ `medium` (equivalently 1, default value) ◦ `fast` (equivalently 1.55, +55%) ◦ `x-fast` (equivalently 2, +100%)	
`volume`	Indicates the volume level of the speaking voice. Volume changes can be applied at the sentence level. You can express the volume as: • An absolute value: Expressed as a number in the range of `0.0` to `100.0`, from *quietest* to *loudest*, such as `75`. The default value is `100.0`. • A relative value: ◦ As a relative number: Expressed as a number preceded by "+" or "-" that specifies an amount to change the volume. Examples are `+10` or `-5.5`. ◦ As a percentage: Expressed as a number preceded by "+" (optionally) or "-" and followed by "%", indicating the relative change. For example: `<prosody volume="50%">some text</prosody>` or `<prosody volume="+3%">some text</prosody>`. • A constant value: ◦ `silent` (equivalently 0) ◦ `x-soft` (equivalently 0.2) ◦ `soft` (equivalently 0.4) ◦ `medium` (equivalently 0.6) ◦ `loud` (equivalently 0.8) ◦ `x-loud` (equivalently 1, default value)	Optional

Prosody examples

For information about the supported values for attributes of the `prosody` element, see Adjust prosody.

Change speaking rate example

This SSML snippet illustrates how the `rate` attribute is used to change the speaking rate to 30% greater than the default rate.

XML

```xml
<speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis" xml:lang="en-
US">
    <voice name="en-US-AvaMultilingualNeural">
        <prosody rate="+30.00%">
            Enjoy using text to speech.
        </prosody>
    </voice>
</speak>
```

Change volume example

This SSML snippet illustrates how the `volume` attribute is used to change the volume to 20% greater than the default volume.

XML

```xml
<speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis" xml:lang="en-
US">
    <voice name="en-US-AvaMultilingualNeural">
        <prosody volume="+20.00%">
            Enjoy using text to speech.
        </prosody>
    </voice>
</speak>
```

Change pitch example

This SSML snippet illustrates how the `pitch` attribute is used so that the voice speaks in a high pitch.

XML

```xml
<speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis" xml:lang="en-
US">
    <voice name="en-US-AvaMultilingualNeural">
        Welcome to <prosody pitch="high">Enjoy using text to speech.</prosody>
    </voice>
</speak>
```

Change pitch contour example

This SSML snippet illustrates how the `contour` attribute is used to change the contour.

XML

```
<speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis" xml:lang="en-
US">
    <voice name="en-US-AvaMultilingualNeural">
        <prosody contour="(60%,-60%) (100%,+80%)" >
            Were you the only person in the room?
        </prosody>
    </voice>
</speak>
```

Adjust emphasis

You can use the optional `emphasis` element to add or remove word-level stress for the text. This element can only contain text and the following elements: `audio`, `break`, `emphasis`, `lang`, `phoneme`, `prosody`, `say-as`, `sub`, and `voice`.

> **Note**
>
> The word-level emphasis tuning is only available for these neural voices: `en-US-GuyNeural`, `en-US-DavisNeural`, and `en-US-JaneNeural`.
>
> For words that have low pitch and short duration, the pitch might not be raised enough to be noticed.

The following table describes the `emphasis` element's attributes:

Expand table

Attribute	Description	Required or optional
`level`	Indicates the strength of emphasis to be applied: • `reduced` • `none` • `moderate` • `strong` When the `level` attribute isn't specified, the default level is `moderate`. For details on each attribute, see emphasis element .	Optional

Emphasis examples

For information about the supported values for attributes of the `emphasis` element, see Adjust emphasis.

This SSML snippet demonstrates how you can use the `emphasis` element to add moderate level emphasis for the word "meetings."

XML

```xml
<speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis"
xmlns:mstts="https://www.w3.org/2001/mstts" xml:lang="en-US">
    <voice name="en-US-AndrewMultilingualNeural">
    I can help you join your <emphasis level="moderate">meetings</emphasis>
fast.
    </voice>
</speak>
```

Add recorded audio

The `audio` element is optional. You can use it to insert prerecorded audio into an SSML document. The body of the `audio` element can contain plain text or SSML markup spoken if the audio file is unavailable or unplayable. The `audio` element can also contain text and the following elements: `audio`, `break`, `p`, `s`, `phoneme`, `prosody`, `say-as`, and `sub`.

Any audio included in the SSML document must meet these requirements:

- The audio file must be valid **.mp3*, **.wav*, **.opus*, **.ogg*, **.flac*, or **.wma* files.

- The combined total time for all text and audio files in a single response can't exceed 600 seconds.
- The audio must not contain any customer-specific or other sensitive information.

> ⓘ **Note**
>
> The `audio` element is not supported by the Long Audio API. For long-form text to speech, use the batch synthesis API instead.

The following table describes the usage of the `audio` element's attributes:

⛶ Expand table

Attribute	Description	Required or optional
`src`	The URI location of the audio file. The audio must be hosted on an internet-accessible HTTPS endpoint. HTTPS is required. The domain hosting the file must present a valid, trusted TLS/SSL certificate. You should put the audio file into Blob Storage in the same Azure region as the text to speech endpoint to minimize the latency.	Required

Audio examples

For information about the supported values for attributes of the `audio` element, see Add recorded audio.

This SSML snippet illustrates how to use `src` attribute to insert audio from two .wav files.

XML

```xml
<speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis" xml:lang="en-
US">
    <voice name="en-US-AvaMultilingualNeural">
        <p>
            <audio src="https://contoso.com/opinionprompt.wav"/>
            Thanks for offering your opinion. Please begin speaking after the
beep.
            <audio src="https://contoso.com/beep.wav">
                Could not play the beep, please voice your opinion now.
            </audio>
        </p>
    </voice>
</speak>
```

Adjust the audio duration

Use the `mstts:audioduration` element to set the duration of the output audio. Use this element to help synchronize the timing of audio output completion. The audio duration can be decreased or increased between `0.5` to `2` times the rate of the original audio. The original audio is the audio without any other rate settings. The speaking rate is slowed down or sped up accordingly based on the set value.

The audio duration setting applies to all input text within its enclosing `voice` element. To reset or change the audio duration setting again, you must use a new `voice` element with either the same voice or a different voice.

The following table describes the usage of the `mstts:audioduration` element's attributes:

Expand table

Attribute	Description	Required or optional
`value`	The requested duration of the output audio in either seconds, such as `2s`, or milliseconds, such as `2000ms`. The maximum value for output audio duration is 300 seconds. This value should be within `0.5` to `2` times the original audio without any other rate settings. For example, if the requested duration of your audio is `30s`, then the original audio must otherwise be between 15 and 60 seconds. If you set a value outside of these boundaries, the duration is set according to the respective minimum or maximum multiple. For output audio longer than 300 seconds, first generate the original audio without any other rate settings, then calculate the rate to adjust using the prosody rate to achieve the desired duration.	Required

mstts audio duration examples

For information about the supported values for attributes of the `mstts:audioduration` element, see Adjust the audio duration.

In this example, the original audio is around 15 seconds. The `mstts:audioduration` element is used to set the audio duration to 20 seconds or `20s`.

XML

```xml
<speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis"
xmlns:mstts="http://www.w3.org/2001/mstts" xml:lang="en-US">
<voice name="en-US-AvaMultilingualNeural">
<mstts:audioduration value="20s"/>
If we're home schooling, the best we can do is roll with what each day brings
and try to have fun along the way.
A good place to start is by trying out the slew of educational apps that are
helping children stay happy and smash their schooling at the same time.
</voice>
</speak>
```

Add background audio

You can use the `mstts:backgroundaudio` element to add background audio to your SSML documents or mix an audio file with text to speech. With `mstts:backgroundaudio`, you can loop an audio file in the background, fade in at the beginning of text to speech, and fade out at the end of text to speech.

If the background audio provided is shorter than the text to speech or the fade out, it loops. If it's longer than the text to speech, it stops when the fade out is finished.

Only one background audio file is allowed per SSML document. You can intersperse `audio` tags within the `voice` element to add more audio to your SSML document.

> ⓘ **Note**
>
> The `mstts:backgroundaudio` element should be put in front of all `voice` elements. If specified, it must be the first child of the `speak` element.
>
> The `mstts:backgroundaudio` element is not supported by the [Long Audio API](). For long-form text to speech, use the [batch synthesis API]() (Preview) instead.

The following table describes the usage of the `mstts:backgroundaudio` element's attributes:

Expand table

Attribute	Description	Required or optional
`src`	The URI location of the background audio file.	Required
`volume`	The volume of the background audio file. Accepted values: `0` to `100` inclusive. The default value is `1`.	Optional
`fadein`	The duration of the background audio fade-in as milliseconds. The default value is `0`, which is the equivalent to no fade in. Accepted values: `0` to `10000` inclusive.	Optional
`fadeout`	The duration of the background audio fade-out in milliseconds. The default value is `0`, which is the equivalent to no fade out. Accepted values: `0` to `10000` inclusive.	Optional

mstss backgroundaudio examples

For information about the supported values for attributes of the `mstts:backgroundaudi` element, see Add background audio.

XML

```xml
<speak version="1.0" xml:lang="en-US"
xmlns:mstts="http://www.w3.org/2001/mstts">
    <mstts:backgroundaudio src="https://contoso.com/sample.wav" volume="0.7"
fadein="3000" fadeout="4000"/>
    <voice name="en-US-AvaMultilingualNeural">
        The text provided in this document will be spoken over the background
audio.
    </voice>
</speak>
```

Next steps

- SSML overview
- SSML document structure and events
- Language and voice support for the Speech service

Feedback

Was this page helpful? 👍 Yes 👎 No

Provide product feedback | Get help at Microsoft Q&A