Basic Speech to Text

Converts speech to text using the generic recognizer. Context can be specified. Sound files can be sent all at once (non-streaming) or in pieces (streaming).

URL:

```
POST https://api.att.com/speech/v3/speechToText
```

Common request headers:

Header Name	Required	Description
Authorization	Required	Has the value **Bearer {token},** where **{token}** is your access token.
Accept	Optional	The format of the data that should be returned. Valid values are **application/json** and **application/xml**. Default is **application/json**.
Content-Type	Required	The type of audio file that you are sending. See the Speech API documentation for a list of valid values.
Content-Length	Optional	The size of the audio file. Only applies to non-streaming requests.
Transfer-Encoding	Optional	Set to **chunked** if a streaming request.
X-SpeechContext	Optional	The context that is applied to the recognition. Valid values are: **BusinessSearch**, **Gaming**, **Generic** (default), **QuestionAndAnswer**, **SMS**, **SocialMedia**, **TV**, **VoiceMail**, and **WebSearch**.
X-Arg	Optional	Specifies multiple parameters, such as language, acoustical model, social media settings, etc. See the documentation for more detail.

The POST body contains the audio data.

Important response elements:

Element Name	Description
Status	That status of the request: **Ok** or **Speech Not Recognized**
ResponseId	The response identifier for this transaction.
NBest	An array of results.
Hypothesis	The transcription of the audio.
Info	Structure that holds additional information about recognition results. See table below.
LanguageId	The language used to recognize the speech. Takes the format language code, hyphen, country code, all lowercase. Example: **en-us**.
Confidence	Indicates how confident the algorithm is that the **Hypothesis** is correct. It's a value between 0 and 1, where 1 is the highest confidence.
Grade	As assessment of the result quality with the values **accept**, **confirm**, and **reject**. However, in most cases, using **Confidence** is more accurate.
ResultText	A modified version of the Hypothesis, often with formatting applied to words such as numbers and times.
Words	An array of strings that represent words in the **ResultText**.
WordScores	An array of floats that indicate confidence scores for each of the words in **Words**. Each float has a value between 0 and 1, where 1 is the highest confidence.

Basic Speech to Text: Example

Sample Request

```
POST https://api.att.com/speech/v3/speechToText
Accept: application/json
Content-Type: audio/wav
Content-Length: 12384
Authorization: Bearer {token}

{Sound file}
```

URL

The format to be returned: **application/json** or **application/xml**.

The format of the audio data in the POST body.

Required for authorization. Always takes the form **Bearer {token}**, where **{token}** is your access token.

POST body: the audio file

Sample Response

Indicates recognition was successful

Confidence of the result

Result of conversion to text

```
{
  "Recognition": {
    "Status": "OK",
    "ResponseId": "f723e9529",
    "Info": {
      "metrics": {
        "audioTime": 5.3,
        "audioBytes": 467968
      }
    },
    "NBest": [
      {
        "WordScores": [
          0.61,
          1,
          1,
          1,
          1,
          0.93
        ],
        "Confidence": 0.92333335,
        "Grade": "accept",
        "ResultText": "I will be home by six",
        "Words": [
          "I",
          "will",
          "be",
          "home",
          "by",
          "six"
        ],
        "LanguageId": "en-US",
        "Hypothesis": "I will be home by six"
      }
    ]
  }
}
```

Confidence of each word

Individual words

Result before any formatting

Speech to Text Custom

Converts speech to text using grammars or hints. (Grammars constrain the results to what is specified in the grammar file. Hints make use of both the grammar file and the generic recognizer.) Grammar and hint files follow the W3C Speech Recognition Grammar Specification Version 1.0.

Sound files can be sent all at once (non-streaming) or in pieces (streaming).

URL:

```
POST https://api.att.com/speech/v3/speechToText
```

Note: The URL is case sensitive.

Request headers:

Header Name	Required	Description
Authorization	Required	Has the value **Bearer {token},** where **{token}** is your access token.
Accept	Optional	The format of the data that should be returned. Valid values are **application/json** and **application/xml.** Default is **application/json**.
Content-Length	Optional	The size of the audio file. Only applies to non-streaming requests.
Content-Language	Optional	The language of the submitted audio. Inline Pronunciation Lexicon Specification (PLS) is required for rare words in all of the languages except for English in the United States of America (**en-us**). If the request uses one or more of the inline hints style grammars, then the only acceptable value for this parameter is **en-us**. If this request uses grammar files, then there are many languages that can be used. See the Speech API documentation for which languages are available.
Content-Type	Required	The content type for the overall MIME message. The only acceptable value for this parameter is **multipart/x-srgs-audio**.
Content-Type (MIME Pronunciation Part)	Optional	The content type for the pronunciation MIME part. If the pronunciation MIME part is included in the request message, then this parameter is required. The only acceptable value for this parameter is **application/pls+xml**.
Content-Type (MIME Grammar Part)	Optional	The content type for the grammar MIME part. The only acceptable value for this parameter is **application/srgs+xml**.
Content-Type (MIME Audio Part)	Optional	The content type for the audio MIME part, which is the type of audio file that you are sending. See the Speech API documentation for a list of valid values.
Content-Disposition (Each MIME Part)	Required	Specifies the content disposition of each MIME part. See table below.
Transfer-Encoding	Optional	Set to **chunked** if a streaming request.
X-SpeechContext	Optional	The context that is applied to the recognition, specifying whether to use grammar or hints. Valid values are: **GrammarList** and **GenericHints**. The default is **GenericHints**.
X-Arg	Optional	Specifies multiple parameters, such as using penalties to improve results, acoustical model, and how much data to return. See the documentation for more detail.

Speech to Text Custom (Continued)

The Content-Disposition for each MIME type has these values:

Part	Format
Pronunciation part	`form-data;name="x-dictionary";filename="`*`filename.pls`*`"` where *filename.pls* is the pronunciation file
Grammar part for inline grammar	`form-data;name="x-grammar";filename="`*`filename.grxml`*`"` where *`filename.grxml`* is the grammar file
Grammar part for single inline grammar	`form-data;name="x-grammar";filename="`*`filename.grxml`*`"` where *`filename.grxml`* is the grammar file
Grammar part for prefix hints plus Generic context	`form-data;name="x-grammar-prefix";filename="`*`filename.grxml`*`"` where *`filename.grxml`* is the grammar file
Grammar part for altgram hints plus Generic context	`form-data;name="x-grammar-altgram";filename="`*`filename.grxml`*`"` where *`filename.grxml`* is the grammar file
Audio part	`form-data;name="x-voice ";filename="`*`filename.wav`*`"` where *`filename.wav`* is the sound file

The POST body contains the audio data.

Common response elements:

Element Name	Description
Status	That status of the request: **Ok** if the recognition was successful; **Speech Not Recognized** if it was not.
ResponseId	The response identifier for this transaction.
NBest	An array of results.
Hypothesis	The transcription of the audio.
Info	Structure that holds additional information about recognition results. See table below.
LanguageId	The language used to recognize the speech. Takes the format language code, hyphen, country code, all lowercase. Example: **en-us**.
Confidence	Indicates how confident the algorithm is that the **Hypothesis** is correct. It's a value between 0 and 1, where 1 is the highest confidence.
Grade	As assessment of the result quality with the values **accept, confirm**, and **reject**. However, in most cases, using **Confidence** is more accurate.
ResultText	A modified version of the Hypothesis, often with formatting applied to words such as numbers and times.
Words	An array of strings that represent words in the **ResultText**.
WordScores	An array of floats that indicate confidence scores for each of the words in **Words**. Each float has a value between 0 and 1, where 1 is the highest confidence.

In addition, there are several elements that are returned if you are using the TV context. See the Speech API documentation for more information.

Speech to Text Custom: Example

Sample Request for using Grammar

```
POST https://api.att.com/speech/v3/speechToText
Accept: application/json
Content-Type: multipart/x-srgs-audio
Authorization: Bearer {token}


x-grammar={grammar grxml file};type=application/srgs+xml
x-voice=<wav file>;type=audio/wav
```

URL

The format to be returned: **application/json** or **application/xml**.

Content type is always **multipart/x-srgs-audi**o

Required for authorization. Always takes the form **Bearer {token}**, where **{token}** is your access token.

The grammar file

The sound file

Sample Grammar File

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE grammar PUBLIC "-//W3C//DTD GRAMMAR 1.0//EN"
          "http://www.w3.org/TR/speech-grammar/grammar.dtd">
<grammar version="1.0" mode="voice" root="OTC" tag-format="semantics/1.0"
   xmlns="http://www.w3.org/2001/06/grammar">
  <rule id="OTC" scope="public">
    <one-of>
        <item>fexofenadine</item>
        <item>diphenhydramine</item>
        <item>brompheniramine</item>
        <item>loratadine</item>
        <item>clemastine</item>
        <item>chlorpheniramine</item>
        <item>certirizine</item>
    </one-of>
  </rule>
</grammar>
```

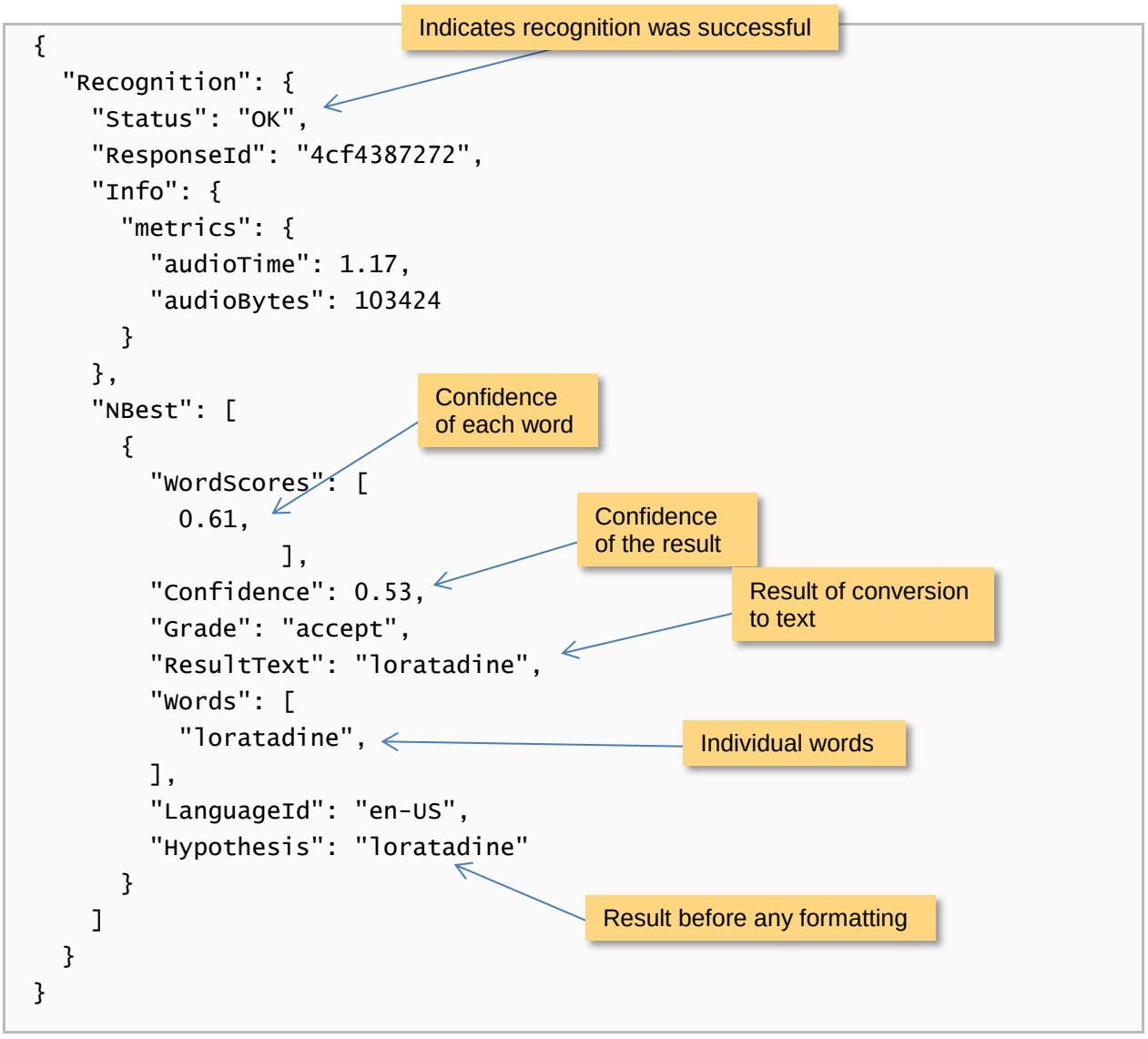
Name of root rule

Constrains to one of these values

Words to be recognized.

Speech to Text Custom: Example (Continued)

Sample Response




```
{
  "Recognition": {
    "Status": "OK",
    "ResponseId": "4cf4387272",
    "Info": {
      "metrics": {
        "audioTime": 1.17,
        "audioBytes": 103424
      }
    },
    "NBest": [
      {
        "WordScores": [
          0.61,
                ],
        "Confidence": 0.53,
        "Grade": "accept",
        "ResultText": "loratadine",
        "Words": [
          "loratadine",
        ],
        "LanguageId": "en-US",
        "Hypothesis": "loratadine"
      }
    ]
  }
}
```

Text to Speech

Converts text to speech, providing options for volume, speed, voice gender, and language. This request requires an OAuth token without user consent.

URL:

```
POST https://api.att.com/speech/v3/textToSpeech
```

Note: The URL is case sensitive.

Request headers:

Header Name	Required	Description
Authorization	Required	Has the value **Bearer {token},** where **{token}** is your access token.
Accept	Optional	The format of the sound file to be returned. Valid values are **application/amr** (Adaptive Multi-Rate codec), **audio/amr-wb** (Adaptive Multi-Rate Wideband codec), or **audio/x-wav** (Microsoft Waveform Audio File Format). Default is **audio/amr-wb**.
Content-Type	Required	The format of the POST body. Valid values are **text/plain** and **application/ssml+xml**.
Content-Length	Optional	The size of the POST body.
Content-Language	Optional	The language that of the supplied text. The acceptable values for this parameter are: **en-US** (English in the United States of America) and **es-US** (Spanish in the United States of America).
X-Arg	Optional	Specifies multiple parameters, such as volume, tempo, and voice. See the documentation for more information.

POST body:

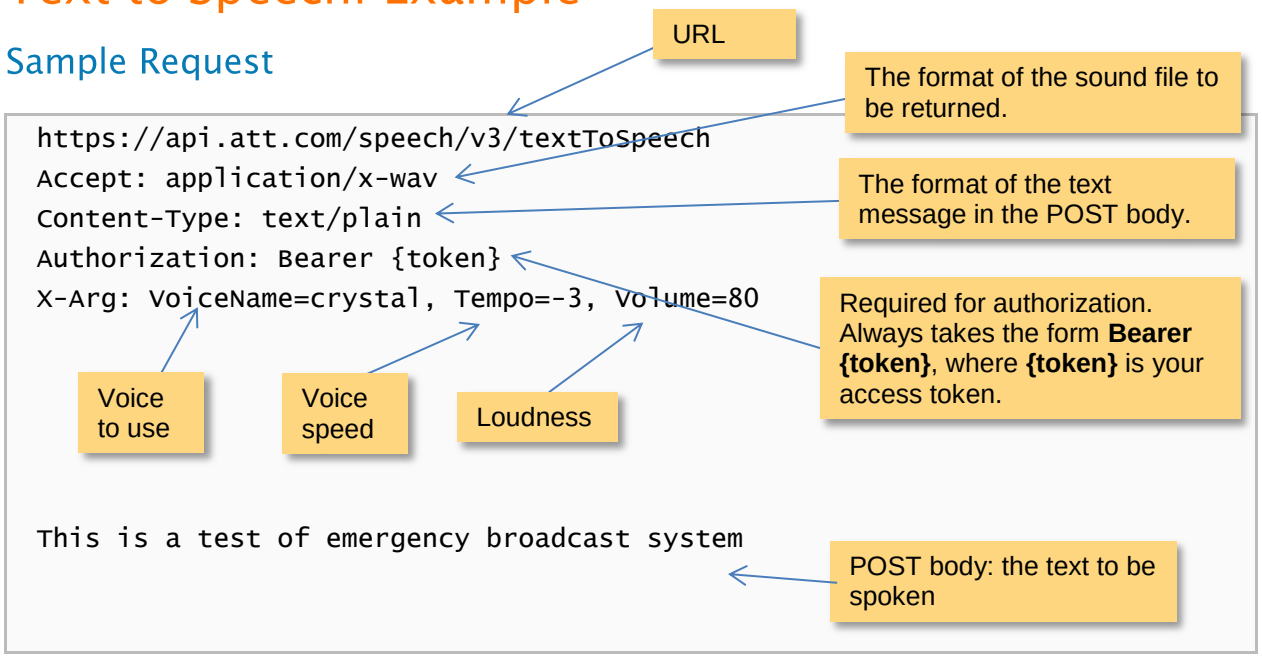
The POST body contains the text to be spoken

Response:

The response is the sound file with the spoken words.

Text to Speech: Example

Sample Request




```
https://api.att.com/speech/v3/textToSpeech
Accept: application/x-wav
Content-Type: text/plain
Authorization: Bearer {token}
X-Arg: VoiceName=crystal, Tempo=-3, Volume=80


This is a test of emergency broadcast system
```

Response: The Sound File