Speech adaptation

Beta

This product or feature is in a pre-release state and might change or have limited support. For more information, see the product launch stages (https://cloud.google.com/products/#product-launch-stages).

This page describes how to improve the accuracy of speech transcription results from Speech-to-Text.

When you send a request to Cloud Speech-to-Text, you can include a list of phrases to act as "hints" during speech transcription. Providing these hints, a technique called *speech adaptation*, helps the Speech-to-Text API to recognize the specified phrases in your input audio files.

Using speech adaptation

You can help Cloud Speech-to-Text to recognize one specific alternative—like "weather"—more frequently by using speech adaptation. To use speech adaptation, you provide a *context* to Cloud Speech-to-Text. A context holds a list of strings to act as hints to Cloud Speech-to-Text. When you provide this context, you increase the probability that Cloud Speech-to-Text recognizes the words in that context when transcribing your source audio data.

You might use speech adaptation in a few ways:

- Improve the accuracy for unique words and phrases that tend to occur very frequently in your audio data. For example, if there is a set of voice commands typically spoken by your users, you can provide those commands as speech adaptations.
- Expand the vocabulary of words recognized by Cloud Speech-to-Text. Cloud Speech-to-Text includes a very large vocabulary. However, if your audio data contains proper names or domain-specific words, you can add them to the phrases provided in the context sent in your transcription request.
- Improve the accuracy of speech transcription when the supplied audio contains noise or is not very clear.

Example of speech adaptation

Consider a scenario where the audio you want to transcribe includes a speaker saying "weather" frequently. When Speech-to-Text encounters audio data with the word "weather," you want it to transcribe the word as "weather" more often than "whether." In this case, you might use speech adaptation to bias Cloud Speech-to-Text towards recognizing "weather."

To use speech adaptation, you create a context containing the words or phrases that you want to recognize. You assign a `SpeechContext` (https://cloud.google.com/speech-to-text/docs/reference/rest/v1p1beta1/RecognitionConfig#SpeechContext) object to the `speechContexts` field of the `RecognitionConfig` (https://cloud.google.com/speech-to-text/docs/reference/rest/v1p1beta1/RecognitionConfig) object in your request to the Speech-to-Text API.

The following snippet shows a part of a JSON payload sent to the Speech-to-Text API. The JSON snippet provides the word "weather" for speech adaptation.

```
"config": {
    "encoding":"LINEAR16",
    "sampleRateHertz": 8000,
    "languageCode":"en-US",
    "speechContexts": [{
      "phrases": ["weather"]
    }]
}
```

Groups of words and phrases

In addition to single words, you can provide multi-word phrases for speech adaptation. When you provide a phrase, Cloud Speech-to-Text is more likely to recognize those words in sequence. Providing a phrase also increases the probability of recognizing portions of the phrase, including individual words.

The following snippet shows a part of a JSON payload sent to the Speech-to-Text API. The JSON snippet includes an array of multi-word phrases assigned to the `phrases` field of a `SpeechContext` object.

```
"config": {
    "encoding":"LINEAR16",
    "sampleRateHertz": 8000,
    "languageCode":"en-US",
    "speechContexts": [{
      "phrases": ["weather is hot", "weather is cold"]
    }]
}
```

Note: See the content limits (https://cloud.google.com/speech-to-text/quotas#content) for limits on the number and size of these phrases.

Using boost adaptation

You can also fine tune speech adaptation results. One technique, called *boost adaptation*, allows you to set some terms to have even more weight with Cloud Speech-to-Text than others.

Take an example where you have a lot of recordings of people asking about the "fare to get into the county fair." In this case, you want Cloud Speech-to-Text to recognize "fair" and "fare" more often than, say, "hare." Extending the example, the word "fair" occurs more frequently than "fare," when you listen to a representative sample of your data.

In this case, you might want to boost both "fair" and "fare." However, because "fair" occurs more than "fare," you want Speech-to-Text API to pick "fair" more frequently.

By default, speech adaptation provides a relatively small effect, especially for one-word phrases. With boost adaptation, you can increase the relative weight of a word or phrase provided for speech adaptation. Setting boost adaptation for a phrase further increases the likelihood that the Speech-to-Text API recognizes the phrase in your source audio.

Note: Refer to the list of supported features by language (https://cloud.google.com/speech-to-text/docs/supported-features-languages) to see whether boost adaptation is available for your audio data.

Setting boost values

When using boost adaptation, you assign a weighted value to a speech adaptation context. The Cloud Speech-to-Text refers to this weighted value when selecting a possible transcription for words in your audio data. The higher you set your value, you increase the likelihood that Cloud Speech-to-Text chooses the phrase from the possible alternatives.

Higher boost values can result in fewer false negatives—cases where the utterance occurred in the audio but wasn't recognized by Cloud Speech-to-Text. However, boost adaptation can also increase the likelihood of false positives—that is, cases where the audio file doesn't contain the utterance but appears in the transcription.

Boost values must be a float value greater than 0. The practical maximum limit for boost adaptation is 20.0, although you can set higher values. For best results, you should experiment by using some initial value and adjust up or down as needed.

Varying boost values for different phrases

You can provide different values for different phrases. You should assign the highest values of boost to the phrases that occur most often in your audio data. With other common phrases, you might want to bias Cloud Speech-to-Text towards them, but not to the same degree as the most frequent or common phrases in your audio data.

Example of boost adaptation

Continuing the previous example, to set the boost adaptation for both "fair" and "fare" in your speech transcription request, you would set two `SpeechContext` objects to the `speechContexts` array of the `RecognitionConfig` (https://cloud.google.com/speech-to-text/docs/reference/rest/v1p1beta1/RecognitionConfig) object. For each `SpeechContext` object, you set a `boost` value as a non-negative float value.

The following snippet shows an example of a JSON payload sent to the Speech-to-Text API. The JSON snippet includes a `RecognitionConfig` object that uses boost adaptation to differently weight the words "fair" and "fare."

```
"config": {
    "encoding":"LINEAR16",
    "sampleRateHertz": 8000,
    "languageCode":"en-US",
    "speechContexts": [{
      "phrases": ["county fair"],
      "boost": 15
     }, {
      "phrases": ["fare"],
      "boost": 2
    }]
  }
```

Classes

Consider an example where your audio data includes recordings of people saying their street address. You might have an audio recording of someone saying "My house is 123 Main Street, the fourth house on the left." In such a case, you want Cloud Speech-to-Text to recognize the first sequence of numerals ("123") as an address rather than as an ordinal ("one-hundred twenty-third").

However, not all people live at "123 Main Street." It becomes impractical to assign every possible street address as contexts for speech adaptation.

For a case like this, you use a *class*. Classes represent groups of words that represent common concepts that occur in natural language. A class allows you to use speech adaptation to improve transcription accuracy for large groups of words that map to a common concept. In the previous case, you can set speech adaptation such that Cloud Speech-to-Text more accurately transcribes phrases like "123 Main Street" and "987 Grand Boulevard" because they are both address numbers.

Note: Refer to the list of supported class tokens (https://cloud.google.com/speech-to-text/docs/class-tokens) to see which tokens are available for your language.

Class tokens

To use a class in speech adaptation, you assign a *class token* in in your `SpeechContext` object. For example, to improve the transcription of addresses from your source audio, you provide the value `$ADDRESSNUM` to your `SpeechContext` object.

For more examples see the list of supported class tokens (https://cloud.google.com/speech-to-text/docs/class-tokens).

You can use classes either as stand-alone phrases or embedded in larger phrases, such as `my address is $ADDRESSNUM`. If you use an invalid or malformed class token, Cloud Speech-to-Text ignores the token without triggering an error but still uses the rest of the phrase for context.

The following snippet shows an example of a JSON payload sent to the Speech-to-Text API. The JSON snippet includes a `SpeechContext` object that uses a class token.

```
"config": {
  "encoding":"LINEAR16",
  "sampleRateHertz": 8000,
  "languageCode":"en-US",
  "speechContexts": [{
    "phrases": ["$ADDRESSNUM"]
   }]
}
```

Note: Class availability varies by transcription model (https://cloud.google.com/speech-to-text/docs/transcription-model) and language. For en-us, the enhanced `phone_call` model currently supports the following classes: `ADDRESSNUM`, `DAY`, `FULLPHONENUM`, `MONEY`, `MONTH`, `OOV_CLASS_DIGIT_SEQUENCE`, `OPERAND`, `PERCENT`, `POSTALCODE`, `TIME`, `YEAR`. The `video` model only supports `OOV_CLASS_DIGIT_SEQUENCE`.

What's next

- Learn how to use speech adaptation in a request to Cloud Speech-to-Text (https://cloud.google.com/speech-to-text/docs/context-strength).
- Review the list of supported class tokens (https://cloud.google.com/speech-to-text/docs/class-tokens)