Best practices

This document contains recommendations on how to provide speech data to the Speech-to-Text API. These guidelines are designed for greater efficiency and accuracy as well as reasonable response times from the service. Use of the Speech-to-Text API works best when data sent to the service is within the parameters described in this document.

If you follow these guidelines and don't get the results you expect from the API, see Troubleshooting & Support (https://cloud.google.com/speech-to-text/docs/support).

For optimal results...	If possible, avoid...
Capture audio with a sampling rate of 16,000 Hz or higher.	Lower sampling rates may reduce accuracy. However, avoid re-sampling. For example, in telephony the native rate is commonly 8000 Hz, which is the rate that should be sent to the service.
Use a lossless codec to record and transmit audio. **FLAC** or **LINEAR16** is recommended.	Using mp3, mp4, m4a, mu-law, a-law or other lossy codecs during recording or transmission may reduce accuracy. If your audio is already in an encoding not supported by the API, transcode it to lossless **FLAC** or **LINEAR16**. If your application must use a lossy codec to conserve bandwidth, we recommend the **AMR_WB**, **OGG_OPUS** or **SPEEX_WITH_HEADER_BYTE** codecs, in that preferred order.
The recognizer is designed to ignore background voices and noise without additional noise-canceling. However, for optimal results, position the microphone as close to the user as possible, particularly when background noise is present.	Excessive background noise and echoes may reduce accuracy, especially if a lossy codec is also used.
If you are capturing audio from more than one person, and each person is recorded on a separate channel, send each channel separately to get the best recognition results. However, if all speakers are mixed in a single channel recording, send the recording as is.	Multiple people talking at the same time, or at different volumes may be interpreted as background noise and ignored.

For optimal results...	If possible, avoid...
Use word and phrase hints to add names and terms to the vocabulary and to boost the accuracy for specific words and phrases.	The recognizer has a very large vocabulary, however terms and proper names that are out-of-vocabulary will not be recognized.
For short queries or commands, use **`StreamingRecognize`** with **`single_utterance`** set to true. This optimizes the recognition for short utterances and also minimizes latency.	Using **`Recognize`** or **`LongRunningRecognize`** for short query or command usages.

Sampling rate

If possible, set the sampling rate of the audio source to 16000 Hz. Otherwise, set the sample_rate_hertz (https://cloud.google.com/speech-to-text/docs/reference/rpc/google.cloud.speech.v1#recognitionconfig) to match the native sample rate of the audio source (instead of re-sampling).

Frame size

Streaming recognition recognizes live audio as it is captured from a microphone or other audio source. The audio stream is split into frames and sent in consecutive **`StreamingRecognizeRequest`** messages. Any frame size is acceptable. Larger frames are more efficient, but add latency. A 100-millisecond frame size is recommended as a good tradeoff between latency and efficiency.

Audio pre-processing

It's best to provide audio that is as clean as possible by using a good quality and well-positioned microphone. However, applying noise-reduction signal processing to the audio before sending it to the service typically reduces recognition accuracy. The service is designed to handle noisy audio.

For best results:

- Position the microphone as close as possible to the person that is speaking, particularly when background noise is present.
- Avoid audio clipping.
- Do not use automatic gain control (AGC).
- All noise reduction processing should be disabled.
- Listen to some sample audio. It should sound clear, without distortion or unexpected noise.

Request configuration

Make sure that you accurately describe the audio data sent with your request to the Speech-to-Text API. Ensuring that the RecognitionConfig (https://cloud.google.com/speech-to-text/docs/reference/rpc/google.cloud.speech.v1#recognitionconfig) for your request describes the correct `sampleRateHertz`, `encoding`, and `languageCode` will result in the most accurate transcription and billing for your request.

Last updated January 17, 2019.