Microsoft

# Transparency Note: Pronunciation Assessment

Updated 5/21/2021

# Table of Contents

# What is a Transparency Note?

An AI system includes not only the technology, but also the people who will use it, the people who will be affected by it, and the environment in which it is deployed. Creating a system that is fit for its intended purpose requires an understanding of how the technology works, what its capabilities and limitations are, and how to achieve the best performance. Microsoft's Transparency Notes are intended to help you understand how our AI technology works, the choices system owners can make that influence system performance and behavior, and the importance of thinking about the whole system, including the technology, the people, and the environment. You can use Transparency Notes when developing or deploying your own system, or share them with the people who will use or be affected by your system.

Microsoft's Transparency Notes are part of a broader effort at Microsoft to put our AI Principles into practice. To find out more, see the Microsoft AI principles.

# Introduction to Pronunciation Assessment

The Pronunciation Assessment API takes both audio and reference text as inputs to evaluate speech pronunciation and gives speakers feedback on the accuracy and fluency of spoken audio, by comparing a machine generated transcript of the input audio with the reference text. The feature currently supports US English and is available in all Speech-To-Text public regions. For more information, see the region list.

With Pronunciation Assessment, language learners can practice, get instant feedback, and improve their pronunciation so that they can speak and present with confidence. Educators can use Pronunciation Assessment to evaluate the pronunciation of multiple speakers in real time.

## The basics of Pronunciation Assessment

The Pronunciation Assessment API provides speech evaluation results compared to reference text by using a machine learning-based approach that correlates highly with speech assessments conducted by native experts. The pronunciation Assessment model was trained with 100,000+ hours of speech data from US English speakers. It can provide accurate results when people miss, repeat, or add phrases compared to the reference text. It also enables rich configuration parameters to support flexibility in using the API, such as setting **Granularity** to change information granularity in evaluation. (For more information, please see more in sample code).

Pronunciation Assessment evaluates three aspects of pronunciation: accuracy, fluency, and completeness. It also provides evaluations at multiple levels of granularity and returns accuracy scores for specific phonemes, words, sentences, or even whole articles. For more information, see how to use Speech SDK for Pronunciation Assessment.

The following table describes the key results. For more information, see the full response parameters. By using natural language processing (NLP) techniques and the **EnableMiscue** settings, Pronunciation Assessment can detect errors such as extra, missing, or repeated words when compared to the reference text. This information helps obtain more accurate scoring to be used as diagnosis information. This capability is useful for longer paragraphs of text.

| Parameter | Description |
|---|---|
| AccuracyScore | Gives a 0-5 score on pronunciation accuracy of the speech with phoneme, word, or full-text-level granularity. Accuracy indicates how closely the speech matches a native speaker's pronunciation in different granularities, from phoneme to full text. |
| FluencyScore | Gives a 0-5 score on fluency of the speech. Fluency indicates how closely the speech matches a native speaker's use of silent breaks between words. |
| CompletenessScore | Gives a 0-5 score on completeness of the speech. Determined by calculating the ratio of pronounced words to the reference text input. |
| ErrorType | This value indicates whether a word is omitted, inserted, or mispronounced compared to the reference text. Possible values are None (meaning no error on this word), Omission, Insertion, Repetition, and Mispronunciation. |

Another set of parameters returned by Pronunciation Assessment are Offset and Duration (referred to together as the "timestamp") The timestamp of speech is returned in structured JSON format. Pronunciation Assessment can calculate pronunciation errors on each phoneme. Pronunciation Assessment can also flag the errors to specific timestamps in the input audio. Customers developing applications can use the signal to offer a learning path to help students focus on the error in multiple ways. For example, the application can highlight the original speech, reply to the audio to compare it with standard pronunciation, or recommend similar words to practice with.

| Parameter | Description |
|---|---|
| Offset | The time (in 100-nanosecond units) at which the recognized speech begins in the audio stream. |
| Duration | The duration (in 100-nanosecond units) of the recognized speech in the audio stream. |

## Example use cases

Pronunciation Assessment can be used for remote learning, exam practice, or other scenarios that demand pronunciation feedback. The following examples are use cases that are deployed or that we've designed for customers using pronunciation Assessment:

- **Educational service provider**: Providers can build applications with the use of Pronunciation Assessment to help students practice language learning remotely with real-time feedback. This use case is typical when an application needs to support real-time feedback. We support streaming upload on audio files for immediate feedback.
- **Education in a game**: App developers, for example, can build a language learning app by combining comprehensive lessons in games with state-of-the-art speech technology to help children learn English. The program can cover a wide range of English skills, such as speaking, reading, and listening, and also train children on grammar and vocabulary, with Pronunciation Assessment used to support children as they learn to speak English. These multiple learning formats ensure that children learn English with ease based on a fun learning style.
- **Education in a communication app**: Microsoft Teams Reading Progress assists the teacher in evaluating a student's speaking assignment with autodetection assistance for omission, insertion, and

mispronunciation. It also enables students to practice pronunciation more conveniently before they submit their homework.

## Considerations when choosing other use cases

Online learning has grown rapidly as schools and organizations adapt to new ways of connecting and methods of education. Speech technology can play a significant role in making distance learning more engaging and accessible to students of all backgrounds. With Azure Cognitive Services, developers can quickly add speech capabilities to applications, bringing online learning to life.

One key element in language learning is improving pronunciation skills. For new language learners, practicing pronunciation and getting timely feedback are essential to becoming a more fluent speaker. For the solution provider that seeks to support learners or students in language learning, the ability to practice anytime, anywhere by using Pronunciation Assessment would be a good fit for this scenario. It can also be integrated as a virtual assistant for teachers and help to improve their efficiency.

The following recommendations pertain to use cases where Pronunciation Assessment should be used carefully:

- **Include a human-in-the-loop for any formal examination scenarios**: Pronunciation Assessment system is powered by AI systems, and external factors like voice quality and background noise may impact the accuracy. A human-in-the-loop in formal examinations ensures the assessment results are as expected.
- **Consider using different thresholds for different scenarios**: Currently, the Pronunciation Assessment score only represents the similarity distance to the native speakers used to train the model. Such similarity distance can be mapped toward different scenarios with rule-based conditions or weighted counting to help provide pronunciation feedback. For example, the grading method for children's learning might not be as strict as that for adult learning. Customers can consider setting a lower threshold on mispronunciation detection, such as 3 instead of the default 4 for adult learning. Considering the specific scenario can help customers design better experiences for the targeted user group.
- **Consider the ability to account for miscues**: When the scenario involves reading long paragraphs, users are likely to find it hard to follow the reference text without making mistakes. These mistakes, including omission, insertion, and repetition, are counted as miscues. With **EnableMiscue** enabled, the pronounced words will be compared to the reference text, and will be marked with Omission, Insertion, Repetition based on the comparison.

# Characteristics and limitations of Pronunciation Assessment

As a part of the Azure Cognitive Services Speech service, pronunciation assessment empowers end-to-end education solutions for computer-assisted language learning. Pronunciation Assessment involves multiple criteria to assess learners' performance at multiple levels of detail, with perceptions similar to human judges.

## How accurate is Pronunciation Assessment?

Pronunciation Assessment feature provides objective scores, like pronunciation accuracy and fluency degree, for language learners in computer-assisted language learning. The performance of pronunciation assessment depends on Azure Cognitive Services Speech-To-Text transcription accuracy with the use of a submitted transcription as reference, and inter-rater agreement between the system and human judges. For a definition of Speech-To-Text accuracy, see Characteristics and limitations for using speech-to-text.

The following sections are designed to help you understand key concepts about accuracy as they apply to using Pronunciation Assessment.

### Language of accuracy

The accuracy of Speech-To-Text affects pronunciation assessment. Word error rate (WER) is used to measure Speech-To-Text accuracy as the industry standard. WER counts the number of incorrect words identified during recognition and then divides by the total number of words provided in the correct transcript, which is often created by human labelling.

### Comparing Pronunciation Assessment to Human Judges

The Pearson correlation coefficient is used to measure the correlation between pronunciation assessment API generated scores and scores generated by human judges. The Pearson correlation coefficient is a measure of linear correlation for two given sequences. It's widely used to measure the difference between automatically generated machine results and human-annotated labels. This coefficient assigns a value between –1 to 1, where 0 is no correlation, negative value means the prediction is opposite to the target, and positive value means how prediction is aligned with the target.

The proposed guidelines for a Pearson correlation coefficient interpretation are shown in the following table. The strength signifies the relationship correlation between two variables and reflects how consistently the machine result aligns with human labels. Values that are close to 1 indicate a stronger correlation.

| Strength of Association | Coefficient Value | Detail |
|---|---|---|
| **Low** | 0.1 to 0.3 | The autogenerated scores from an automatic system aren't significantly aligned with the perception of humans. |
| **Medium** | 0.3 to 0.5 | The autogenerated scores from an automatic system are aligned with the perception of humans, but differences still exist, and people might not agree with the result. |
| **High** | 0.5 to 1.0 | The autogenerated scores from an automatic system are aligned with the perception of humans, and people are willing to agree with the system results. |

In our evaluations, Microsoft Pronunciation Assessment has performed >0.5 Pearson correlation with human judges' results, which indicates the autogenerated results are highly consistent with the judgment of human experts.

## System limitations and best practices to improve system accuracy

- Pronunciation Assessment works better with higher-quality audio input. We recommend an input quality of 16 kHz or higher.
- Pronunciation Assessment quality is also affected by the distance of the speaker from the microphone. Recordings should be made with the speaker close to the microphone, and not over a remote connection.
- Pronunciation Assessment doesn't support a mixed lingual assessment scenario; it only supports US English.

- Pronunciation Assessment doesn't support a multi-speaker assessment scenario. The audio should include only one speaker for each assessment.
- Pronunciation Assessment compares the submitted audio to native speakers in general conditions. The speaker should maintain a normal speaking speed and volume, and avoid shouting or otherwise raising their voice.
- Pronunciation Assessment is a new feature of the Azure Speech to Text service. Accuracy improvement on Speech-To-Text will also benefit this feature. Customers can further improve the customized Speech-To-Text accuracy using their own data. For more information, see Improve Speech to Text accuracy with Custom Speech and Improve Speech to Text accuracy with tenant models.
- Pronunciation assessment performs better in an environment with little background noise. Current Speech-To-Text models accommodate noise in general conditions. Noisy environments or multiple people speaking at the same time might lead to lower confidence of the evaluation. To handle difficult cases better, you can suggest that the speaker should repeat a pronunciation if they score below a certain threshold.

## Evaluating Pronunciation Assessment in your applications

Pronunciation Assessment's performance will vary depending on the real-world uses that customers implement. In order to ensure optimal performance in their scenarios, customers should conduct their own evaluations of the solutions they implement using Pronunciation Assessment.

- Before using Pronunciation Assessment in your applications, consider whether this product performs well in your scenario. Collect real-life data from the target scenario, test how Pronunciation Assessment performs, and make sure Speech-To-Text and Pronunciation Assessment can deliver the accuracy you need, see Evaluate and improve Azure Cognitive Services Custom Speech accuracy.
- Select suitable thresholds for the target scenario. Pronunciation Assessment provides accuracy scores at different levels and you may need to consider the threshold employed in real-use. For example, the grading method for children's learning might not be as strict as that for adult learning. Customers can consider setting a lower threshold on mispronunciation detection, such as 3 instead of the default 4 for adult learning. Considering the specific scenario can help customers design better experiences for the targeted user group.

## Learn more about responsible AI

Microsoft AI principles

Microsoft responsible AI resources

Microsoft Azure Learning courses on responsible AI

## Learn more about Pronunciation Assessment

How to use pronunciation assessment

## Contact us

Give us feedback on this document

## About this document



Published: 05/21/2021

Last updated: 05/21/2021