Microsoft

Transparency Note: Custom Neural Voice

Updated 6/21/2022

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 Custom Neural Voice

Custom Neural Voice is a Text-to-Speech (TTS) feature, part of Speech Service in Azure Cognitive Services, that allows customers to create a one-of-a-kind customized synthetic voice for their applications by providing their own audio data of their selected voice talents. For more information on Custom Neural Voice, see Overview of Custom Neural Voice.

Limited Access to Custom Neural Voice

Custom Neural Voice is a Limited Access service, and registration is required for access to some features. To learn more about Microsoft's Limited Access policy visit aka.ms/limitedaccesscogservices. Certain features are only available to Microsoft managed customers and partners, and only for certain use cases selected at the time of registration.

Approved use cases

The following use cases are approved for customers:

- **Media: Educational or interactive learning:** For use to create a fictional brand or character voice for reading or speaking educational materials, online learning, interactive lesson plans, simulation learning, standardized testing, or guided museum tours.
- **Media: Entertainment:** For use to create a fictional brand or character voice for reading or speaking entertainment content for video games, movies, TV, recorded music, podcasts, audio books, or augmented or virtual reality.
- **Media: Journalistic or news:** For use to create voices for reading news or journalistic content that must be accompanied by a published, text version of the same content.
- **Media: Marketing:** For use to create a fictional brand or character voice for reading or speaking marketing and product or service media, product introductions, business promotion, or advertisements.
- **Media: Self-authored content:** For use to create a voice for reading content authored by the voice talent except where the voice is used to enhance the authority or credibility of the content in connection with financial, health, legal, political, or spiritual matters.
- **Accessibility Features:** For use in audio description systems, narration, or to facilitate communication by speaking impaired individuals.
- **Interactive Voice Response (IVR) Systems:** For use to create voices for call center operations, telephony systems, or responses for phone interactions.
- **Public Service Announcement**: For use to create a fictional brand or character voice for announcements for public venues.

- **Translation and Localization**: For use in real-time translation applications for translating conversations in different languages or translating audio media.
- **Virtual Assistant or Chatbot:** For use to create a fictional brand or character voice for smart assistants in or for virtual web assistants, appliances, cars, home appliances, toys, control of IoT devices, navigation systems, reading out personal messages, virtual companions, or customer service scenarios.

Considerations when using Custom Neural Voice

The ability to produce synthetic media generatively, rapidly, and at scale offers unique opportunities for augmenting personal and creative expression, but also poses unprecedented challenges to public safety by making it easier to misappropriate, misinform, mislead, propagandize, or libel; while simultaneously undermining the believability of legitimate recordings and other digital artifacts. For this reason, Microsoft has established the following Code of Conduct that prohibits certain uses of Custom Neural Voice.

In addition to reviewing the Code of Conduct when choosing a use case to use Custom Neural Voice, take the following considerations into account:

- **Avoid photo realistic avatars with synthetic voices to represent real people** - One of the key principles of the responsible use of Custom Neural Voice is to ensure that our consumers understand and expect the content they are interacting with is synthetic as opposed to being real. Pairing a photo realistic avatar with the custom neural voice of a real person could potentially create an illusion to consumers that they are interacting with a real and known person. This would erode trust in the application and potentially cause harm to consumers.
- **Carefully consider using a synthetic voice with contents without editorial control** - Synthetic voice can sound like a human, which could amplify the effect of fake or misleading content.

Characteristics and limitations of Custom Neural Voice

Custom Neural Voice enables you to create brand voices for your apps. To create a Neural TTS voice that sounds like a specific person or style for your apps, the recordings of the voice talent embodying that personality and style are used. The audio data is then uploaded to the Custom Voice platform where a machine learning model is built using transfer learning technology and our multi-lingual, multi-speaker base model. While the custom voice models can speak in a voice that sounds like the voice talent, the similarity and naturalness of the voice depends on a number of factors, including the size of the training data, the quality of the recording, the accuracy of the transcript file, how well the recorded voice in the training data matches the personality of the designed voice for your intended use case, as well as other factors.

Fairness

At Microsoft, we strive to empower every person on the planet to do more. An essential part of this goal is working to create technologies and products that are fair and inclusive. Fairness is a multi-dimensional, socio-technical topic and impacts many different aspects of our product development. You can learn more about Microsoft's approach to fairness here.

One dimension we need to consider is how well the system performs for different groups of people. Research has shown that without conscious effort focused on improving performance for all groups, it is often possible for the performance of an AI system to vary across groups based on factors such as race, ethnicity, gender, and age.

Custom Neural Voice models are trained using transfer learning technology based on our multi-lingual, multi-speaker base model, with the recording samples of human voices that you provide. The richer the base model, the more powerful the transfer learning, the less training data is required. While our base model is built using a

wide range of speech data that incudes different speaking accents, age groups, genders, across more than 50 languages (see list of supported languages here), it's possible that certain demographic groups are not well represented in some languages. For example, we cover less speech data from children than from adults. To accommodate this limitation, we require at least 300 lines of recordings (or, around 30 minutes of speech) to be prepared as training data for Custom Neural Voice, and we recommend 2,000 lines of recordings (2-3 hours of speech) to create a voice for production use. Each line of recording (a.k.a. "utterance") consists of a normal sentence or a short phrase that is read by your chosen voice talent. With 2,000 utterances, our system can learn the target voice characteristics well even if the base model doesn't include a similar speaker. The evaluation is done using SMOS measurement as described below. We recommend lower score for SMOS measurement across languages as judges need to compare the recording from primary language with voice in secondary language.

Each application is different, and our base model may not match your context or cover all scenarios required for your use case. We encourage developers to thoroughly evaluate synthesized voice quality for the service with real-world data that reflects your use case, including testing with users from different demographic groups and with different speech characteristics. Please see the Quality of the voice model trained section for best practices for building high quality voice models.

In addition to ensuring similar performance, it is important to consider how to minimize risks of stereotyping and erasure that may result from synthetic voices. For example, if you are creating a custom neural voice for a smart voice assistant, carefully consider what voice is appropriate to create, and seek diverse perspectives from a variety of backgrounds. When building and evaluating your system, seek diverse input.

Quality of the voice model trained

You can measure the quality of the voice model produced by Custom Neural Voice through listening to the samples generated by the service. A quantitative approach is to use MOS (Mean Opinion Score) to measure naturalness. In MOS, independent human judges score the naturalness of the voice samples they are presented with using a scale of 1 to 5, with 1 being the worst quality and 5 being the best quality. An average score is then calculated for the report. When selecting judges, it is recommended to include people with a variety of backgrounds, including judges from different demographic groups.

In custom voice quality evaluations, MOS gap is usually used to compare the quality of the TTS voice model against a human recording. The quality of a voice model created by Custom Neural Voice compared to that of a human recording is expected to be close, with a gap of no more than 0.5 in the MOS score.

In addition, you can use Similarity MOS (SMOS) to measure how similar the custom voice sounds compared to the original human recordings. With SMOS studies, judges are asked to listen to a set of paired audios, one generated using the custom voice, the other from the original human recordings in the training data, and rate if the two audios in each pair are spoken by the same person, using a five-point scale (1 being the lowest, 5 the highest). The average score is reported as the SMOS score. We recommend that a good custom neural voice should achieve an SMOS higher than 4.0.

Besides measuring naturalness with MOS and SMOS, you can also assess the intelligibility of the voice model by checking the pronunciation accuracy of the generated speech. This is done by having the judges listen to a set of testing samples, determine whether they can understand the meaning and indicate any words that were unintelligible to them. Intelligibility rate is calculated using the percentage of the correctly intelligible words among the total number of words tested (i.e., the number of intelligible words/the total number of words tested * 100%). Normally a usable TTS engine needs to reach a score of > 98% for intelligibility.

Measurement	Definition	How it is calculated	Recommended text size	Recommended score
MOS	Mean Opinion Score of the quality of the audios	Average of the rating scores of each judge on each audio	> 30 generated audios > 20 judges on each audio	> 4.0 (normally requires the MOS of the human recording is higher than 4.5)
MOS gap	The MOS score difference between human recordings and the generated audios	The MOS score on the human recordings minus the MOS score on the generated audios	> 10 human recordings > 30 generated audios > 20 judges on each audio	<0.5
SMOS	The similarity of the generated audios to the human recordings	Average of the rating scores of the similarity level on each pair of audios	> 40 pairs > 20 judges on each pair	>4.0 >3.5 (secondary language)
Intelligibility	The pronunciation accuracy of the generated speech at the word level	Percentage of the correctly intelligible words among the total number of words tested	> 60 generated audios > 10 judges on each audio	>98%

Best practices to improve the quality of the voice model

Creating a great custom voice requires careful quality control in each step, from voice design, data preparation, to the deployment of the voice model to your system.

Persona design

Before building a custom neural voice, particularly if it is for a fictitious voice, it is a good practice to design a persona of the voice that would represent your brand, fit well with the user scenarios as well as resonate with your intended audience. This can be specified in a persona brief: a document that describes the features of the voice and the fictitious or real person behind the voice. This persona brief document will help guide the process of creating a custom voice model including the recording scripts, the voice talent selection, and the training and tuning of the voice.

Script selection

Your recording script defines the training dataset for your voice model and is the starting point of any custom voice recording session. Your recording script must be carefully selected to represent the user scenarios for your voice. For example, if you are going to use the voice model for your customer service bot, you may want to use the phrases from your bot conversations to create the recording script. To create a voice for reading stories, you can use a relevant story script for your recordings.

Follow the guidance here to prepare your script in more detail.

Note

When preparing your recording script, make sure you include a statement sentence to acquire the voice talent acknowledgement for using their voice data to create a TTS voice model and generate synthetic speech. You can find the statement in multiple languages **here**. The language of the verbal statement must be the same as your recording.

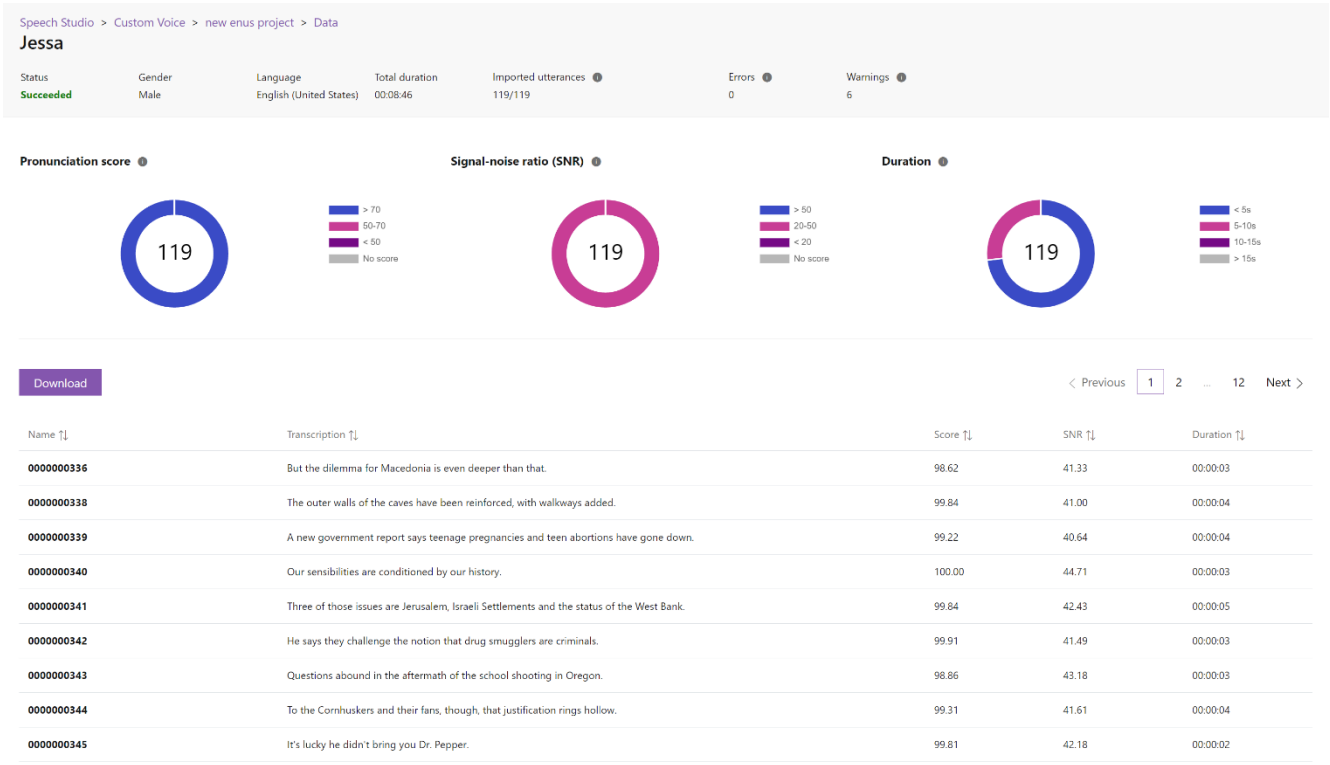
As a technical safeguard intended to prevent misuse of Custom Neural Voice, Microsoft will use this recording to verify that the voice talent's voice in the script matches the voice provided in the training data through **Speaker Verification**. Read more about this process in the **Data and Privacy section here**.

Preparing training data

We recommend that the audio recordings should be captured in a professional quality recording studio so that high signal-to-noise ratio is achieved without distortion, and other defects such as RF interference and plosives are minimized. Follow the recording guidance here.

The quality of the voice model trained heavily depends on the recorded voice used for training. Consistent volume, speaking rate, speaking pitch, and consistency in expressive mannerisms of speech are essential to create a great custom neural voice. Custom Neural Voice creates a synthetic voice model that mimics the voice in the training data. The quality of the recording of the voice talent is the upper bound of the quality of the Custom Neural Voice model. The voice model will capture the styles, accents, and other characteristics of the voice, including defects such as noises and mispronunciations.

Unexpected errors such as mismatching of the transcript to the recordings can introduce mistakes in the pronunciation labeling to the training system. The Speech Studio provides capabilities for users to evaluate the pronunciation accuracy, identify noises, check the audio length for each utterance in the training dataset, and filter out unqualified recordings. For example, in the dataset detail view, you can check the pronunciation score, the signal-to-noise ratio and the audio length for your training data.



It is nevertheless not possible to provide 100% accurate data review results automatically. As the developer creating the synthetic voice, you are responsible for reviewing and ensuring the audio quality of the training data is sufficient for voice model building.

Tuning and adjustment

The style and the characteristics of the trained voice model depend on the style and the quality of the recordings from the voice talent used for training. However, several adjustments can be made using SSML (Speech Synthesis Markup Language) when you make the API calls to your voice model to create synthetic speech. SSML is the markup language used to communicate with the TTS service to convert text into audio. The adjustments include change of pitch, rate, intonation, and pronunciation correction. If the voice model is built with multiple styles, SSML can also be used to switch the styles.

All of the SSML markups mentioned above can be passed directly to the API. We also provide an online tool, Audio Content Creation, that allows customers to tune using a friendly UI.

Learn more about responsible AI

Audio Content Creation

Microsoft responsible AI resources

Microsoft Azure Learning courses on responsible AI

Learn more about Custom Neural Voice

What is Custom Neural Voice?

Contact us

Give us feedback on this document

About this document

Published: 06/11/2021

Last updated: 6/21/2022