Speech Interaction Service

Product Introduction

Issue 03
Date 2019-08-30
Contents

1 What Is SIS......................................................................................................................................1
2 Application Scenarios...................................................................................................................5
3 Related Services........................................................................................................................... 14
4 Restrictions and Limitations..................................................................................................... 15
   4.1 ASRC............................................................................................................................................................................15
   4.2 RASR............................................................................................................................................................................15
   4.3 ASR.............................................................................................................................................................................. 15
   4.4 TTS............................................................................................................................................................................... 16
   4.5 TTSC............................................................................................................................................................................ 16
5 Monitoring.................................................................................................................................... 17
   5.1 ASR Monitoring Metrics.............................................................................................................................................. 17
   5.2 TTS Monitoring Metrics...............................................................................................................................................18
   5.3 Viewing Metrics............................................................................................................................................................18
A Change History........................................................................................................................... 20
Speech Interaction Service (SIS) provides a man-machine interaction mode by open Application Programming Interfaces (APIs). You can obtain the speech interaction result by real-time access and API calling. SIS consists of the following sub-services:

- **ASR Customization (ASRC):** leverages deep learning technology to optimize speech recognition for specific fields and allows you to define language models. ASRC provides the Sentence Transcription and Long Audio Transcription functions.
- **Real-Time ASR (RASR):** converts continuous audio streams into text in real time, enabling faster speech recognition.
- **Automatic Speech Recognition (ASR):** converts audio recordings into text.
- **Text To Speech (TTS):** converts text into lifelike voices.
- **Text To Speech Customization (TTSC):** leverages advanced speech technologies from Huawei, in combination with deep learning algorithms, to convert text into smooth, natural speech.

### ASRC

ASRC provides the Sentence Transcription and Long Audio Transcription functions. Sentence Transcription recognizes audio recordings with a shorter duration at a faster speed, and Long Audio Transcription performs well in recognizing audio recordings with a longer duration.

- Sentence Transcription converts audio recordings whose duration is less than 1 minute into text. After binary data is uploaded, the system automatically processes and converts it into text.
- Long Audio Transcription recognizes long audio recordings and converts them into text. It features good scalability, provides different models for different domains, and supports hot word customization.

### ASRC Advantages

- **High Recognition Rate**
  Utilizes the deep learning technology to optimize the corpus for specific scenarios and fields, enabling an industry-leading recognition rate.
- **Cutting-Edge Technologies**
  Combines mature speech recognition algorithms currently in active use in the industry with the latest research to empower enterprises with unique competitive advantages.
**RASR**

You can access and call APIs to obtain the speech recognition result in real time.

**RASR Advantages**

- **High Recognition Accuracy**
  Adopts the latest generation of speech recognition and Deep Neural Network (DNN) technologies to greatly improve the anti-noise performance and recognition accuracy.

- **High Speed**
  Integrates the language models, dictionaries, and acoustic models into a large neural network featuring impressive optimizations in the engineering to greatly increase the decoding speed, achieving faster recognition.

- **Multiple Recognition Modes**
  Supports multiple real-time speech recognition modes, including streaming, continuous, and single-sentence, to suit different application scenarios.

- **Customization Service**
  Allows you to customize the language-layer model in a specific vertical domain to better recognize proprietary words and industry terms, adding a significant boost to accuracy.

**RASR Functions**

- **Text Timestamps**
  Generates specific timestamps for the audio conversion result, so that you can quickly find the spot in the original audio clip to confirm the text and adopt if needed.

- **Intelligent Text Segmentation**
  By extracting semantic features of the context and combining voice features, intelligently segments sentences and adds punctuation marks to improve the readability of the output text.

- **Hybrid Recognition**
  Supports recognition of English letters/words and digits included in Chinese sentences.

- **Instant Result Output**
  Continuously recognizes voice streams, outputs results in real time, and automatically corrects the content based on the context language model.

- **Automatic VAD**
  Performs voice activity detection (VAD) on the input voice streams to improve recognition efficiency and accuracy.

- **Flexible Access Modes**
  Supports access over WebSocket and MRCP interfaces.
ASR

ASR converts audio recordings whose duration is within 1 minute and whose size is less than 4 MB into text. After a complete audio file is uploaded, the system automatically converts it into text.

**ASR Advantages**

- **High Accuracy**
  Employs deep learning technologies to achieve a speech recognition accuracy of over 95%.
- **Strong Language Support**
  Supports recognition of speeches in Chinese Mandarin to meet the application requirements in various scenarios.
- **Solid Reliability**
  Proven stability after years of experience in complex enterprise customer scenarios.
- **High Efficiency**
  Provides standard RESTful APIs and service SDKs to facilitate use and integration while reducing labor and business costs.

TTS

TTS converts text into lifelike voices. It provides speech services with customizable timbres, volumes, and speeds for enterprises and individuals.

**TTS Advantages**

- **High Accuracy**
  Employs deep learning technologies to rapidly synthesize natural, fluent human speech.
- **Customization**
  Allows you to customize the timbre, tone, and speed of spoken text based on your needs.
- **Solid Reliability**
  Proven stability after years of experience in complex enterprise customer scenarios.
- **High Efficiency**
  Provides standard RESTful APIs and service SDKs to facilitate use and integration while reducing labor and business costs.

TTSC

TTSC can handle both Chinese and English, and provides customizable playback. You can adjust the pitch, speed, or volume as needed.

**TTSC Advantages**

- **Customization**
  Provides custom speech services with better synthesis effect for enterprise customers.
- **Multiple Languages and Timbres**
  TTSC can handle both Chinese and English and provides customizable playback (male, female, child's voices for you to select). You can adjust the speed or volume as needed.
- **Smooth and Natural**
  The speech converted from text is natural, clear, and lifelike, meeting requirements of various application scenarios.
Table 2-1 lists the application scenarios of ASRC.

<table>
<thead>
<tr>
<th>Application Scenario</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voice Customer Service Inspection</td>
<td>ASRC recognizes the speech of the customer service personnel and customer, converts the speech into text, and checks whether it contains any violation, sensitive word, or phone number through text retrieval.</td>
</tr>
</tbody>
</table>

**Figure 2-1** Voice customer service inspection
<table>
<thead>
<tr>
<th>Application Scenario</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meeting Minutes Taking</td>
<td>ASRC quickly recognizes the audio of meeting minutes and converts it into text, facilitating automatic and efficient meeting minutes recording.</td>
</tr>
<tr>
<td></td>
<td><strong>Figure 2-2 Meeting minutes taking</strong></td>
</tr>
<tr>
<td>Voice Message</td>
<td>ASRC converts voice messages you send or receive into text to deliver higher reading efficiency and interaction experience.</td>
</tr>
<tr>
<td></td>
<td><strong>Figure 2-3 Voice message</strong></td>
</tr>
</tbody>
</table>
ASR

Table 2-2 shows the application scenarios of ASR.

Table 2-2 ASR application scenarios

<table>
<thead>
<tr>
<th>Application Scenario</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voice Search</td>
<td>ASR allows users to search with their voice, improving search efficiency. It supports voice search in various scenarios, such as map navigation and web page search.</td>
</tr>
</tbody>
</table>

Figure 2-5 Voice search
ASR integrates the voice wakeup service so that voice commands sent to terminals initiate operations in real time, improving the interaction experience between human and machine.

**Figure 2-6** Human-machine interaction

<table>
<thead>
<tr>
<th>Application Scenario</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human-Machine Interaction</td>
<td>ASR integrates the voice wakeup service so that voice commands sent to terminals initiate operations in real time, improving the interaction experience between human and machine.</td>
</tr>
</tbody>
</table>

**TTS**

**Table 2-3** shows the application scenarios of TTS.

**Table 2-3** TTS application scenarios

<table>
<thead>
<tr>
<th>Application Scenario</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voice Navigation</td>
<td>TTS uses vehicle-mounted navigational data to provide you with accurate, customizable voice navigation services.</td>
</tr>
</tbody>
</table>

**Figure 2-7** Voice navigation
<table>
<thead>
<tr>
<th>Application Scenario</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audible Reading</td>
<td>TTS converts books, magazines, and news articles into human voices, providing a convenient audio file so you can obtain the latest news in the car, on the way to work, or at the gym.</td>
</tr>
<tr>
<td>Figure 2-8 Audible reading</td>
<td></td>
</tr>
<tr>
<td>Telephone Follow-Up</td>
<td>TTS converts telephone follow-up content into human voices to communicate with customers, improving the overall convenience of calling customers.</td>
</tr>
<tr>
<td>Figure 2-9 Telephone follow-up</td>
<td></td>
</tr>
<tr>
<td>Application Scenario</td>
<td>Description</td>
</tr>
<tr>
<td>----------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Smart Education</td>
<td>TTS converts textbooks into life-like voices to simulate classroom teaching, helping teachers find new, innovative ways to better the education of their students.</td>
</tr>
</tbody>
</table>

**Figure 2-10 Smart education**

---

**RASR**

*Table 2-4* shows the application scenarios of RASR.
<table>
<thead>
<tr>
<th>Application Scenario</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Live Subtitling</td>
<td>RASR converts the audio from live video streams into subtitles in real time, optimizing the viewing experience while bringing much added convenience to content monitoring.</td>
</tr>
<tr>
<td></td>
<td><strong>Figure 2-11</strong> Live subtitling</td>
</tr>
<tr>
<td>Real-Time Conference Recording</td>
<td>RASR converts the audio in a video or conference call into text in real time, and allows you to quickly verify, modify, and retrieve the text.</td>
</tr>
<tr>
<td></td>
<td><strong>Figure 2-12</strong> Real-time conference recording</td>
</tr>
</tbody>
</table>
### Application Scenario

<table>
<thead>
<tr>
<th>Application Scenario</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instant Text Generation</td>
<td>RASR records your speech and converts it into text on mobile apps, such as voice input.</td>
</tr>
</tbody>
</table>

**Figure 2-13 Instant text generation**

### TTSC

*Table 2-5* shows the application scenarios of TTSC.

#### Table 2-5 TTSC application scenarios

<table>
<thead>
<tr>
<th>Application Scenario</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human-Machine Interaction</td>
<td>TTSC implements high-quality and natural interaction between human beings and machines.</td>
</tr>
</tbody>
</table>

**Figure 2-14 Human-machine interaction**
<table>
<thead>
<tr>
<th>Application Scenario</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smart Customer Service</td>
<td>With TTSC, contact centers can engage customers with natural sounding voices.</td>
</tr>
</tbody>
</table>

**Figure 2-15 Smart customer service**

![Image of a computer screen with a feedback form and a person on a phone call.]
3 Related Services

Identity and Access Management (IAM)

IAM provides the user authentication and authorization functions for SIS. For more information about IAM, see the Identity and Access Management User Guide.

Object Storage Service (OBS)

ASR APIs can be temporarily authenticated or anonymously and publicly authorized to obtain data from OBS for processing. Long Audio Transcription can be temporarily authenticated to obtain data from OBS for processing. For more information about OBS, see the Object Storage Service API Reference and Object Storage Service Developer Guide.
4 Restrictions and Limitations

4.1 ASRC

Sentence Transcription
- Currently, the sub-service is available in CN North-Beijing1 and CN North-Beijing4.
- pcm16k16bit, pcm8k16bit, ulaw16k8bit, ulaw8k8bit, alaw16k8bit, wav, amr, and amrwb are supported.
- The audio duration cannot exceed 1 minute.
- Only recognition of speech in Chinese Mandarin is supported.

Long Audio Transcription
- Currently, the sub-service is available in CN North-Beijing1 and CN North-Beijing4.
- pcm16k16bit, pcm8k16bit, ulaw16k8bit, ulaw8k8bit, alaw16k8bit, alaw8k8bit, vox8k4bit, v3_8k4bit, WAV (supporting the pcm/ulaw/alaw/adcpcm coding format internally), MP3, M4A, ogg-speex, ogg-opus, and AMR are supported.
- The recording duration cannot exceed 5 hours, and the file size cannot exceed 300 MB. The recognition result is stored for 72 hours (starting from the transcription completion time).

4.2 RASR

- Currently, the sub-service is available in CN North-Beijing1 and CN North-Beijing4.
- The audio sampling rate is 8 kHz or 16 kHz, and the number of sampling bits is 8 bits or 16 bits.
- Only recognition of speech in Chinese Mandarin with a few English words is supported.

4.3 ASR

- Currently, the sub-service is available in CN North-Beijing1 only.
- Audio files in WAV, MP3, WMA, AMR, AC3, OGG, and AAC formats are supported. The audio sampling rate is 8 kHz or 16 kHz, and the number of sampling bits is 16 bits.
The Base64-encoded voice data cannot exceed 4 MB.
The audio duration cannot exceed 1 minute.
Only recognition of speech in Chinese Mandarin is supported.

4.4 TTS

- Currently, the sub-service is available in **CN North-Beijing1** only.
- Only Chinese and English text is supported and the text can contain a maximum of 500 Chinese characters or English letters.
- Currently, API calling concurrency cannot be ensured.
- The supported synthesis sampling rate is 16 kHz.

4.5 TTSC

- Currently, the sub-service is available in **CN North-Beijing1** and **CN North-Beijing4**.
- Only Chinese and English text is supported and the text can contain a maximum of 500 Chinese characters or English letters.
- Currently, API calling concurrency cannot be ensured.
- The supported synthesis sampling rates are 8kHz and 16 kHz.
5 Monitoring

5.1 ASR Monitoring Metrics

Function

This section describes metrics reported by ASR to Cloud Eye as well as their namespaces, list, and dimensions. You can use the management console or APIs provided by Cloud Eye to query the metric information generated for ASR.

Namespace

SYS.ASR

Metrics

<table>
<thead>
<tr>
<th>Metric ID</th>
<th>Metric</th>
<th>Description</th>
<th>Value Range</th>
<th>Monitored Object</th>
<th>Monitoring Period (Original Metric)</th>
</tr>
</thead>
<tbody>
<tr>
<td>successful_call_times_of_service</td>
<td>Successful Calls of Service</td>
<td>Number of successful API calls Unit: Times/min</td>
<td>≥ 0 times/min</td>
<td>API</td>
<td>1 minute</td>
</tr>
<tr>
<td>failed_call_times_of_service</td>
<td>Failed Calls of Service</td>
<td>Number of failed API calls Unit: Times/min</td>
<td>≥ 0 times/min</td>
<td>API</td>
<td>1 minute</td>
</tr>
</tbody>
</table>
5.2 TTS Monitoring Metrics

Function

This section describes metrics reported by TTS to Cloud Eye as well as their namespaces, list, and dimensions. You can use the management console or APIs provided by Cloud Eye to query the metric information generated for TTS.

Namespace

SYS.TTS

Metrics

<table>
<thead>
<tr>
<th>Metric ID</th>
<th>Metric</th>
<th>Description</th>
<th>Value Range</th>
<th>Monitored Object</th>
<th>Monitoring Period (Original Metric)</th>
</tr>
</thead>
<tbody>
<tr>
<td>successful_calls_of_service</td>
<td>Successful Calls of Service</td>
<td>Number of successful API calls</td>
<td>≥ 0 times/min</td>
<td>API</td>
<td>1 minute</td>
</tr>
<tr>
<td>failed_calls_of_service</td>
<td>Failed Calls of Service</td>
<td>Number of failed API calls</td>
<td>≥ 0 times/min</td>
<td>API</td>
<td>1 minute</td>
</tr>
</tbody>
</table>

Dimensions

<table>
<thead>
<tr>
<th>Key</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>call_of_interface</td>
<td>API</td>
</tr>
</tbody>
</table>

5.3 Viewing Metrics

1. Log in to the management console.
2. Choose Service List > Management & Deployment > Cloud Eye.
3. In the navigation pane on the left, choose Cloud Service Monitoring and select the service whose metrics are to be viewed.
4. Click View Metric in the Operation column.
5. In the monitoring area, you can select a duration to view the monitoring data.
   Monitoring data generated in the latest 1 hour, 3 hours, or 12 hours can be viewed.
## Change History

<table>
<thead>
<tr>
<th>Release Date</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2019-08-30</td>
<td>Added the following topic: TTSC</td>
</tr>
<tr>
<td>2019-07-30</td>
<td>Added the following topic: ASRC</td>
</tr>
<tr>
<td>2019-03-30</td>
<td>This issue is the first official release.</td>
</tr>
</tbody>
</table>