Nuance Audio Input Specification

NUANCE MOBILE DIVISION

© 2015-2016 Nuance Communications, Inc. All rights reserved.
ABOUT THIS DOCUMENT

This document describes requirements and best practices for the audio input path for Nuance Developers platform.

Nuance Developers platform supports different concepts of audio signal acquisition (the user’s speech) for speech recognition:

- Close-talk (10-25 cm) audio input with a microphone on the remote control device,
- Close-talk audio input from the primary Mobile device, e.g.: Smartphone / Tablet,
- Distant-talk (25-100 cm) audio input from a microphone array, e.g.: TV array or PC/laptop array.

This document provides general signal acquisition recommendations and audio specifications, not specific input or acquisition options. Please contact Nuance for additional guidance for specific audio acquisition advice.

GENERAL AUDIO REQUIREMENTS:

Nuance Recommendations on Audio Path

<table>
<thead>
<tr>
<th>Audio Chain Characteristic</th>
<th>Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency Response</td>
<td>The front end frequency response must be flat within +/-4 dB over the range of 100 Hz to 8 kHz</td>
</tr>
<tr>
<td>Gain</td>
<td>Gain will be set once during initialization to a level just below clipping (2 bits headroom) when speaking with loud voice at normal distance from the microphone</td>
</tr>
<tr>
<td>Signal Resolution</td>
<td>16-bit A/D Converter is recommended to provide an effective dynamic range of 12 bits for the speech signal while accounting for the 2-bit headroom.</td>
</tr>
<tr>
<td>Transients</td>
<td>Duration of any transients shall not exceed 50 ms following command to start recording. Long-tailed transients shall settle to within noise floor limits within this period.</td>
</tr>
<tr>
<td>Low Frequency Roll-off</td>
<td>Maximum 6 dB attenuation at 200 Hz; minimum 24 dB attenuation 0-100 Hz</td>
</tr>
<tr>
<td>Anti-Aliasing</td>
<td>The attenuation of the anti-alias filter at the Nyquist frequency should be 20dB or better with a high order roll-off.</td>
</tr>
<tr>
<td>Sampling Rate</td>
<td>Recommended sampling rate is 16 kHz, 16-bit resolution yielding signal bandwidth of 250-7300 Hz</td>
</tr>
<tr>
<td>Compression</td>
<td>Several high-quality audio compression codecs are supported for Nuance Cloud</td>
</tr>
</tbody>
</table>
Response Latency

Response time is critical in terms of user experience and is affected by a number of variables including audio processing. For this reason, Nuance Cloud Services (NCS) is designed to process audio in real-time while the user is speaking. Therefore, it is required that audio is streamed to NCS in real-time. For example, 3 seconds of audio should be streamed over 3 clock seconds (1X real-time), additional latency results when audio is buffered and streamed significantly faster than real-time to catch up on streaming backlog.

Activation

Typically, start-of-speech is signaled by a manual event (button press) or wake-up word; end-of-speech is either automatically detected by CODEC/DSP or manually by a user event. (See below for additional information)

Expected SNR Range:

The expected SNR range is such that the median SNR is approximately 18-20dB, no more than 10% of samples have SNR<11 and no more than 25% of recordings have SNR<15dB.

Digital Silence

There should be no digital silence at any point in the audio whose duration is longer than a few 10-20 samples. This includes DC value 0, DC constant value and samples that have no energy above 200Hz.

DC Offset

Any DC offset transient should be filtered out.

Clipping

Clipping (signal values exceeding the dynamic range of the system) can seriously affect ASR performance depending on the amount of clipping present in each sample, and how many samples (user utterances) are clipped. In general, gain should be set to avoid clipped samples. The number of clipped frames in a sample should not exceed 10%, and the total number of samples with clipping should be less the 2% of the total.

MICROPHONE

Microphone Specification

- Directional characteristics
  Both, omni-directional or directional (e.g. cardioid) microphone types are supported.

- Equivalent noise level
  The equivalent noise level of the microphone self noise shall not exceed 30 dB_{SPL} (A-weighted).

- Frequency response
  The frequency response of the microphone shall support a frequency range up to at least 8 kHz. A roll-off characteristic between 7 kHz and 8 kHz is allowed. However, at 200 Hz the response should be at least -18 dB (with respect to 1 kHz).
• **Microphone Sensitivity:**

It is recommended to use a microphone with a sensitivity of $300 \text{ mV}_{\text{RMS}}/\text{Pa} \pm 3\text{ dB (f = 1 kHz)}$, where $\text{mV}_{\text{RMS}}$ is the root mean square voltage and 1 Pa corresponds to a sound pressure level of 94 dB$\text{SPL}$. Thus, a peak to peak voltage of $85 \text{ mV}_{\text{PP}}$ is obtained at 74 dB$\text{SPL}$. Other sensitivities are possible if the maximum input sound level still conforms with the requirements and the sensitivity of the ADC is adjusted accordingly.

• **Maximum input sound level**

A headroom of 6 dB above the maximum expected sound pressure level at the microphone is recommended. Within this range the microphone should operate linearly. A typical value is e.g. 110 dB$\text{SPL}$. Larger headroom may lead to an increased level of noise. For smaller headroom there is the risk of non-linear distortions in the signal.

• **Digital full scale level**

The sensitivity of the ADC shall be adjusted such that the maximum expected signal amplitude is still represented without clipping. A typical value is e.g. 104 dB$\text{SPL}$.

It has to be ensured that the microphone signals do not clip. Clipping may occur due to a saturation of the analogue microphone signal or due to a saturation of the analogue-to-digital conversion (ADC). The analogue-to-digital converters should be adjusted such that the digital full-scale level is reached before the analogue microphone signal shows considerable non-linear distortions (such as saturation effects).

• **Maximum input sound level**

A headroom of 6 dB above the maximum expected sound pressure level at the microphone is recommended. Within this range the microphone should operate linearly. The distortion factor has to be less than 3% (THD) at 107 dB$\text{SPL}$ (equiv. to a SPL of 4.4 Pa). No significant distortion should occur up to the point where the AD converter clips. Larger headroom may lead to an increased level of noise. For smaller headroom there is the risk of non-linear distortions in the signal.

• **Digital full scale level**

The sensitivity of the ADC shall be adjusted such that the maximum expected signal amplitude is still represented without clipping. A typical value is e.g. 104 dB$\text{SPL}$.

As stated above, the preferred audio codec is raw PCM at 16 bit 16kHz (256kbps) without pre-processing steps such as gain control or noise reduction on the device, since these might interfere with the speech recognition engine.

Nuance prefers to review any built-in noise reduction and missing frame recovery algorithms with the manufacturer in an early stage of the project in order to make recommendations on which methods are best suited to affect ASR the least. To avoid costs and delays for investigating individual solutions, we strongly recommend using one of the compression / transmissions schemes below.
CODECS, COMPRESSION AND TRANSMISSION

Audio sampling

Even if compression algorithms are used to reduce the transmission bandwidth, the initial sampling of the audio in the device should be 16 kHz with 16 bit resolution. It is important to avoid recording and down-sampling to 8kHz and then up-sampling to 16kHz again. Otherwise there will be a mismatch in the signal bandwidth expected by the speech recognizer and the actual signal characteristics.

Audio Compression

An uncompressed signal generally provides better performance, but has high requirements in terms of transmission bandwidth.

Nuance Developers APIs have different supported and recommended CODECs as shown in the table below.

<table>
<thead>
<tr>
<th>Format</th>
<th>HTTP 1.0</th>
<th>SpeechKit</th>
<th>Web Sockets API</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCM 16k-16b—256 kbps</td>
<td>Supported</td>
<td>Supported</td>
<td>Supported</td>
</tr>
<tr>
<td>OPUS 16kHz CBR (complexity 10)</td>
<td>Not supported</td>
<td>Recommended</td>
<td>Not supported</td>
</tr>
<tr>
<td>PCM 8k-16b—128 kbps</td>
<td>Not supported</td>
<td>Supported</td>
<td>Not supported</td>
</tr>
<tr>
<td>SPEEX-WB—16kHz 27.8 kbps CBR</td>
<td>Recommended</td>
<td>Supported</td>
<td>Recommended</td>
</tr>
<tr>
<td>SPEEX NB—8kHz 24.6 kbps CBR</td>
<td>Supported</td>
<td>Supported</td>
<td>Supported</td>
</tr>
<tr>
<td>AMR-NB—12.2 kbps variable bit rate</td>
<td>Not supported</td>
<td>Not supported</td>
<td>Not supported</td>
</tr>
<tr>
<td>ADPCM—64 kbps</td>
<td>Not supported</td>
<td>Not supported</td>
<td>Not supported</td>
</tr>
<tr>
<td>G.711 (ulaw)—64 kbps</td>
<td>Not supported</td>
<td>Not supported</td>
<td>Not supported</td>
</tr>
</tbody>
</table>

Bluetooth

Ideally, the BT wideband audio profiles should be available to allow 16 kHz audio input, but Nuance speech engines also support narrow-band 8 kHz audio sampling.

Response Latency

Response latency is a critical performance consideration in terms of user experience, and is affected by a number of variables including audio processing. For this reason, Nuance Cloud Services (NCS) is designed to process audio in real time while the user is speaking. Therefore, it required that audio is streamed to NCS in real-time. For
example, 3 seconds of audio should be streamed over 3 clock seconds (1X real-time), additional latency results when audio is buffered and streamed significantly faster than real-time to catch up on streaming backlog.

Transmission

For Cloud Services, transmission bandwidth over the data channel is an important variable to estimate. The table below provides transmission requirements for two key supported Codecs:

<table>
<thead>
<tr>
<th>Context</th>
<th>Ave. Utterance (seconds)</th>
<th>Kbits/sec (CODEC)</th>
<th>Total Data Sent**</th>
</tr>
</thead>
<tbody>
<tr>
<td>POI Search</td>
<td>2.5</td>
<td>28 (OPUS)* 256 (full PCM)</td>
<td>84 kilobits 768 kilobits</td>
</tr>
<tr>
<td>Music search</td>
<td>2.5</td>
<td>28 (OPUS) 256 (full PCM)</td>
<td>84 kilobits 768 kilobits</td>
</tr>
<tr>
<td>SMS</td>
<td>5</td>
<td>28 (OPUS) 256 (full PCM)</td>
<td>168 kilobits 1536 kilobits</td>
</tr>
<tr>
<td>Email</td>
<td>7.5 (multiple utterances)</td>
<td>28 (OPUS) 256 (full PCM)</td>
<td>252 kilobits 2304 kilobits</td>
</tr>
<tr>
<td>Web Search</td>
<td>3</td>
<td>28 (OPUS) 256 (full PCM)</td>
<td>100 kilobits 922 kilobits</td>
</tr>
</tbody>
</table>

* 16kHz CBR (complexity 10)

** 20% NCS protocol overhead

**SPEECH ACTIVATION**

In order to avoid clipping, battery drain or other issues Nuance Mobile solutions will typically not stream audio continuously to the receiving device but only send data when the user activates the wake-up word, presses a dedicated button (soft or hard push-to-talk key), or responds to a prompt in a dialog (programmatically).

Therefore the identification of the beginning and end of a user utterance on the device and NCS system needs to function hand in hand.

Detecting the start of the user utterance

Using a Wake-up-Word (e.g. “Hello Dragon”) to get the system’s attention is supported by the Nuance system requires a continuous audio stream from the microphone to the Nuance Wake-Up-Word Service (if enabled on the device).

Your application must provide the speaker with a clear and regular notification (audio or visual) that the system is listening. A slight delay (100 ms) in presenting this notification will improve recognition performance by providing a small amount of initial silence for the system’s channel adaptation to adapt to the user environment, and reduces the likelihood of clipping the beginning of speech.

Detecting the end of the user utterance
The Nuance system is capable of automatically detecting the end of the user utterance based on the signal characteristics. However this assumes a continuous stream of (non-zero) audio samples from the device.

At the same time the device needs a mechanism or behavior to stop streaming audio to the device.

Options:

1) **End-of-speech detection on device (CODEC or DSP based)**

   Concept: Once the user presses the SPEAK button, the application keeps transmitting audio to the device until it receives an "end-of-speech" event from the audio subsystem; typically handled by the SPEEX CODEC in mobile handsets.

   Risk: Poor performing EoS detectors lead to over aggressive end-pointing or audio streams much longer than the user's utterance. It is critical that the solution provide a user-activated EoS event.

2) **Keep button pressed while talking**

   Concept: The user needs to keep the SPEAK button pressed the whole time while he is speaking. The device stops transmitting audio only after the user has released the SPEAK button.

   Risk: User releases SPEAK button too early (\(\rightarrow\) continue sending audio for 200msec after user has released audio button).

   Recommendation: Preferred option

3) **Fixed timeout**

   Concept: After the user has pressed the SPEAK button, the device keeps transmitting audio for a fixed time even if the user releases the SPEAK button.

   Risk: The timeout needs to be carefully chosen and depends on the supported use cases (short command&control utterances vs long content queries).

   Recommendation: Apply only if the time window can be guaranteed to be long enough. (e.g. 20 seconds)