语音唤醒数据采集和交付规范

1、录音设备

- **采样深度**：16位
- **采样率**：16KHz或更高，必要时可后期进行降采样处理。
- **音量控制**：将麦克风音量设置为适当水平，避免录音时音量过低或过高。

如果使用手机或便携设备进行录音，确保设备的麦克风能够清晰捕捉声音，且不进行噪声抑制或其他处理，以避免产生过度失真。

2、录音环境

- **场地要求**：最佳选择为专业的录音室，如隔音室，其他室内环境也可以使用。
- **环境要求**：环境应安静，无明显回声，且应避免背景噪声的干扰，如电视声音、空调、风扇、街道交通等。不得录入鼠标、触摸板、桌面敲击声等类似噪音。
- **设备距离**：录音人员与录音设备之间的距离应根据产品的使用场景来确定。举例如下：

  手表产品，距离可以是20cm、50cm或70cm；

  手机产品，距离可以是0.5m或1m；

  音响产品，距离可以是0.5m、1m、3m或5m。

3、录音人员

为确保数据多样性，应尽可能录制更多人群，包括不同性别、年龄、口音和语速的声音。

- 人数要求：每个词汇录制350 ~ 600人。
- 样本数量：每人每个词汇录制30 ~ 50个样本（每个词汇的样本总数应为15,000 ~ 30,000个，更多数据可提升算法模型的性能）。
- 性别比例：男性和女性各占50%，波动不超过10%。
- 年龄分布：应包括符合产品目标人群的年龄段，

  举个例子如：

  儿童：7 ~ 12岁，10%。

  青少年：13 ~ 17岁，20%。

  年轻人：18 ~ 30岁，40%。

中年人：31～50岁，20%。

老年人：51岁及以上，10%。

- 地区要求：根据发音差异，数据采集应覆盖不同地区口音。
- 语速：语速应包括正常、慢速和快速，语速快慢的定义应基于录音人员的正常说话习惯。对于单一唤醒词，最快语速应不少于0.5秒，最慢语速应超过2.1秒，两个连续唤醒词之间应有约2秒的停顿。
- 音量：音量应有正常、轻声和大声的变化，音量大小的定义应基于录音人员的正常说话习惯。语音数据16位，数值范围-32768~32767。只看幅度正值，幅值值总体分布应该符合一个高斯分布，比如，主要集中在6000——14000，同时小幅值2000也会存一定比例，大幅值20000也会有一定比例。

录音人要求:

- 录制人发音正确：唤醒词发音清晰，可以有口音，但不能影响到正常人听识别。
- 避免发音错误：确保正确发音，不得有误读现象。
- 避免多余的声音：避免咳嗽、交谈或其他噪音。
- 呼吸和口腔声音：自然的呼吸和口腔声音是允许的，但不能在喘息时进行录音。
- 避免不自然的停顿：正常的呼吸停顿是允许的，但不应过长，停顿时间不应超过0.5秒。
- 避免口吃或中断：唤醒词应流畅发音，不应出现口吃或停顿。

如有干扰需重新录制：若录音过程中任何声音影响了音频质量，受影响的文件必须重新录制。

4、音频文件

- 格式：16位，16KHz，单声道，wav文件。
- 文件命名：文件名应该包含录音人ID、性别、省市地区、唤醒词、语速或音量信息以及音频文件编号

命名应该只包含：数字、大小字母、汉字、下划线“_”，文件命名要准确，尽量不要使用缩写，举例格式：Name_Gender_Region_Word_SpeechSpec_Number.wav

- 文件夹组织：以录音人ID作为文件夹，里面存放录音人员的音频文件；如果有多个唤醒词，则建立子文件夹，以唤醒词命名，每个唤醒词的音频文件存放在对应的子文件夹下。

举例如下：

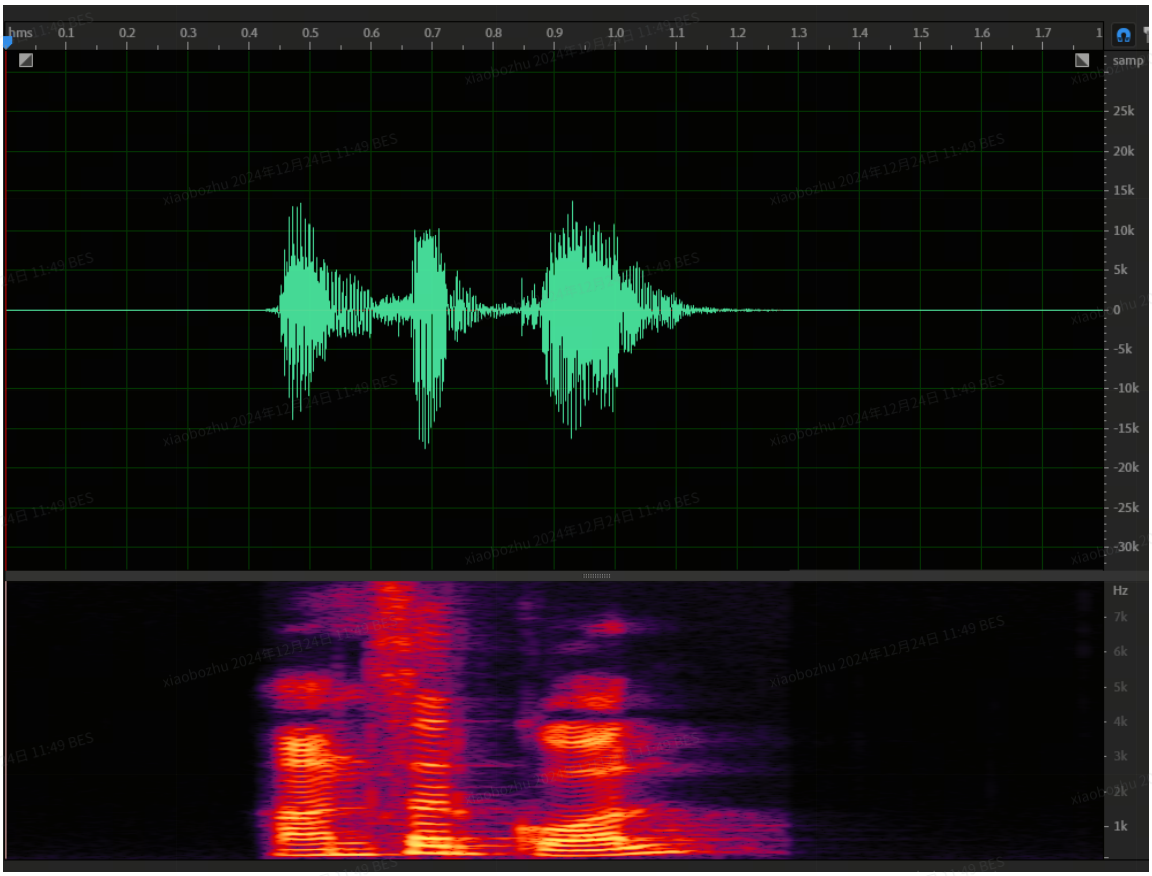
```
SPK001
SPK002
...
SPK090
        /HiSiri
```

```
        /PreviousTrack
        /NextTrack
        /PlayMusic
                /SPK009_M_NewYork_PlayMusic_Normal_001.wav
                /SPK009_M_NewYork_PlayMusic_Loud_002.wav
                /SPK009_M_NewYork_PlayMusic_Small_003.wav
        /StopMusic
                ...
SPK010
        /HiSiri
        /PreviousTrack
        /NextTrack
        /PlayMusic
        /StopMusic
                /SPK010_F_ShangHai_StopMusic_Normal_001.wav
                /SPK010_F_ShangHai_StopMusic_Loud_002.wav
                /SPK010_F_ShangHai_StopMusic_Small_003.wav
                ...
```

其他注意事项：

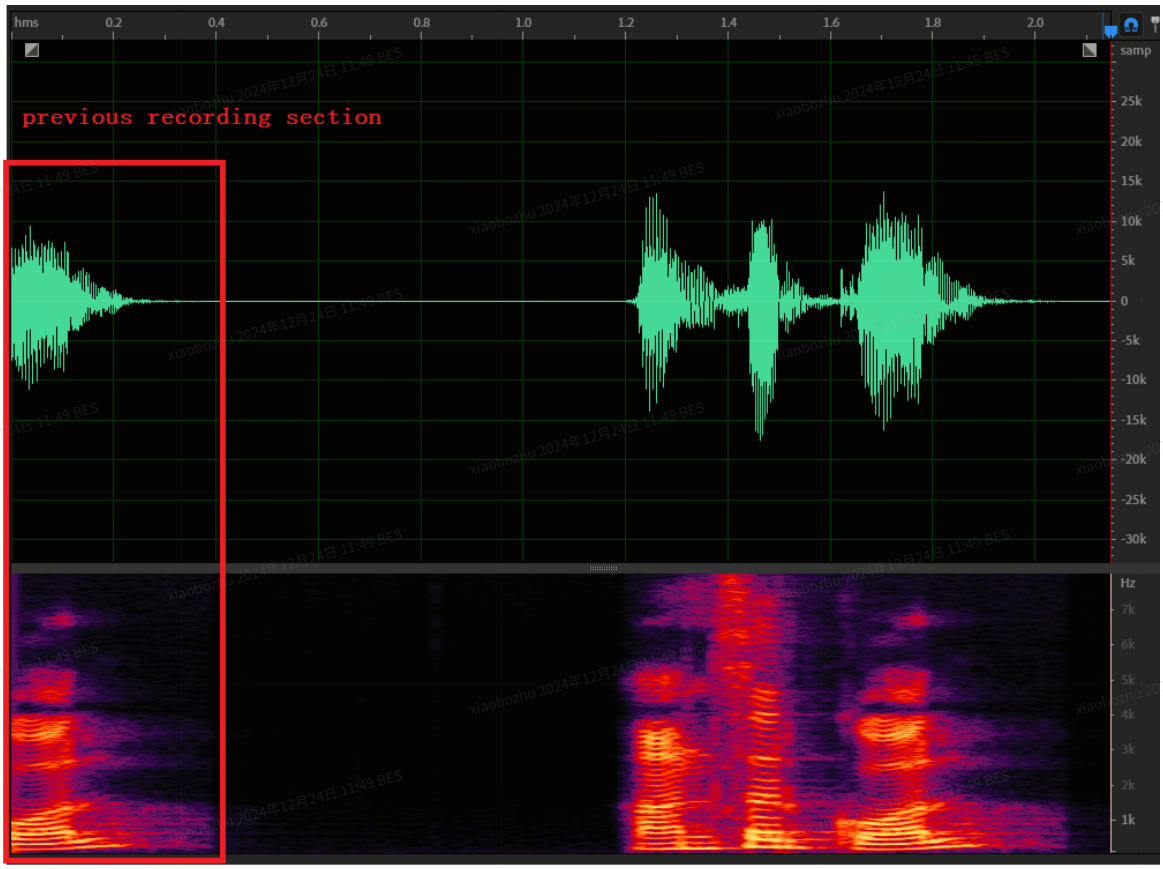
- 每个音频文件应只包含一个唤醒词。
- 禁止空白音频文件。
- 避免语音截断：比如将“Hi Siri”剪辑为“Hi Si”。

一个正确音频文件示例如下，

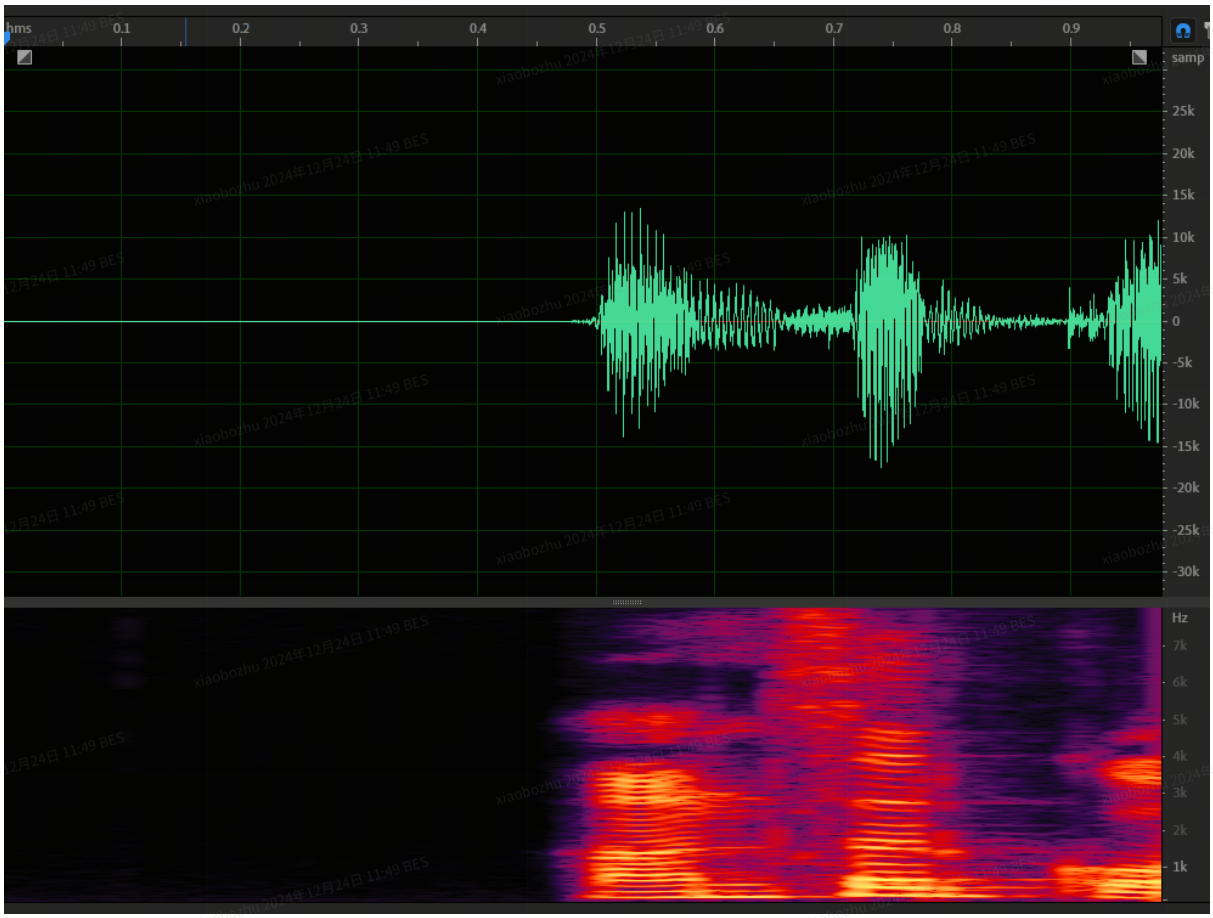


如下是不正确的音频文件，

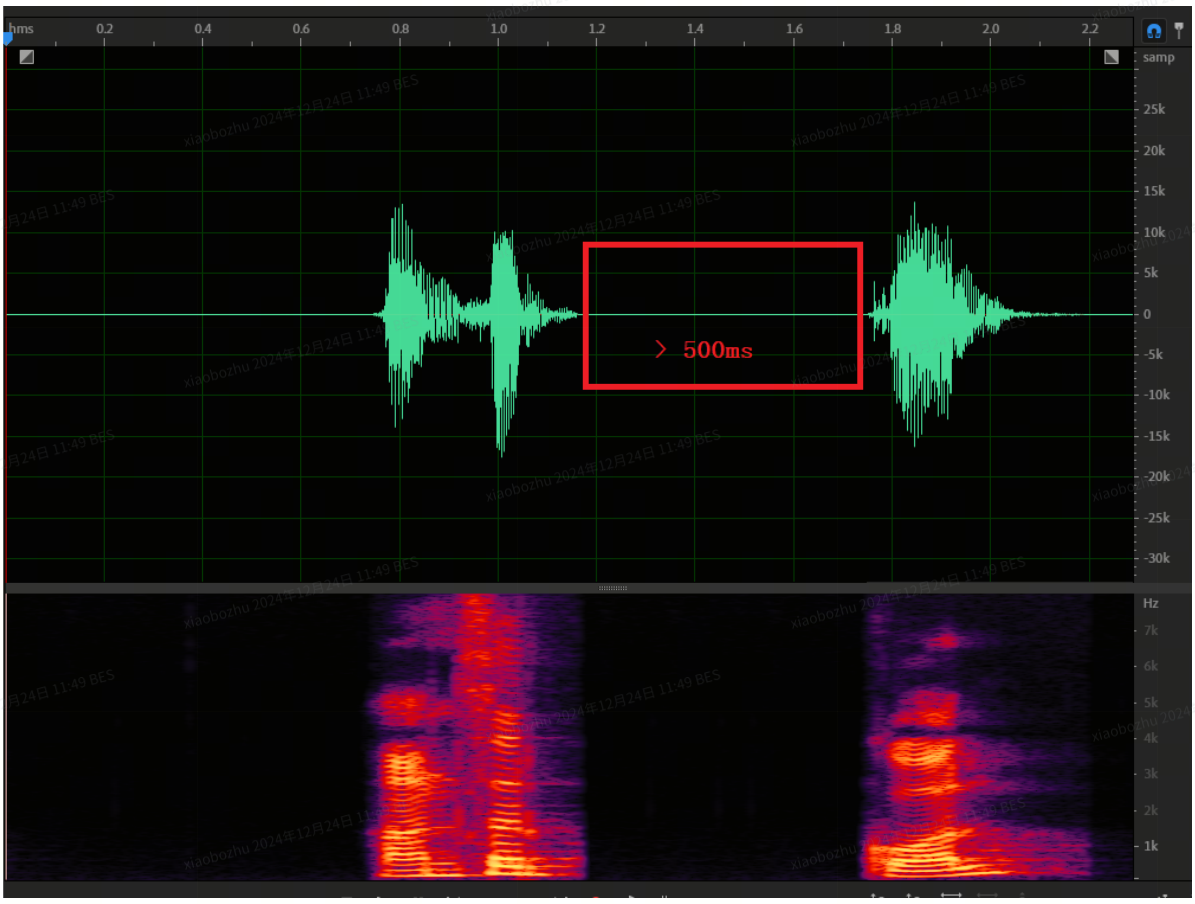
（1）录音文件在截切的时候，错误的剪切位置，导致包含了前面的词的部分。



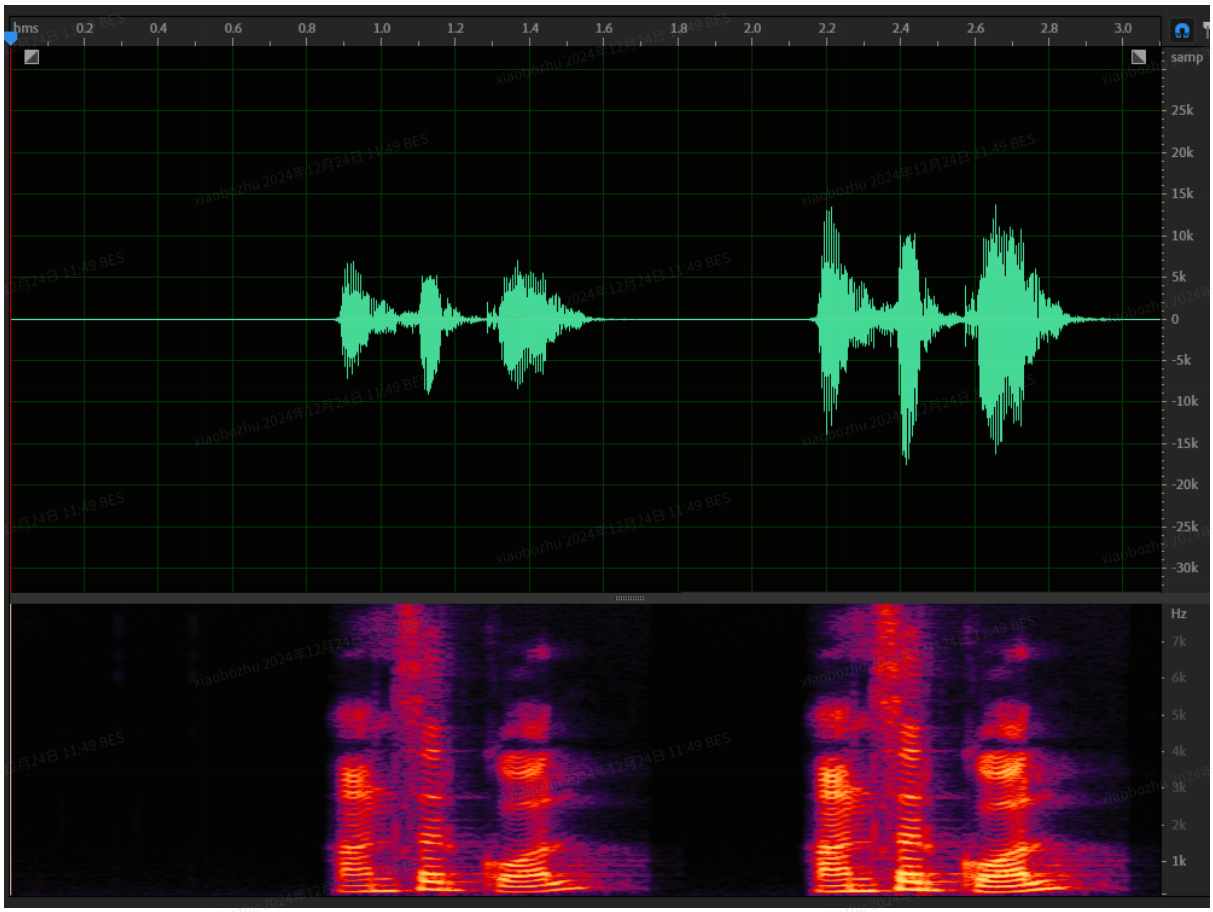
（2）唤醒词尾部被截断或丢失，唤醒词音频不完整



（3）单个唤醒词说话停顿时间过长，明显大于500ms



（4）单个音频文件包含多个唤醒词语音



1. Recording Equipment

- **Sampling depth**: 16-bit
- **Sampling rate**: 16KHz or higher, with downsampling applied later if necessary.
- **Volume control**: Set the microphone volume to an appropriate level to avoid recording with either too low or too high volume.

If using a mobile phone or portable device for recording, ensure that the device's microphone supports clear voice capture without any noise reduction or other processing, to avoid excessive distortion.

2. Recording Environment

- **Venue Requirements**: The ideal choice is a professional recording room, such as a soundproof room, though other indoor environments can also be used.
- **Environmental Requirements**: The environment should be quiet, with no noticeable echoes, and free from background noise such as TV sound, air conditioning, fans, street traffic, etc. Do not record sounds from devices such as a mouse, touchpad, desk tapping, or similar noises.

- **Device Distance**: The distance between the recording personnel and the recording equipment should be determined based on the product's usage scenario.For example，

  For watch products, the distance can be 20cm, 50cm, or 70cm;

  For mobile phone products, the distance can be 0.5m or 1m;

  For speaker products, the distance can be 0.5m, 1m, 3m, or 5m.

3. Recording Personnel

To ensure data diversity, as many people as possible should be recorded, including voices of different genders, ages, accents, and speaking speeds.

- **Number of People**: 350 ~ 600 people per word.
- **Number of Samples**: 30 ~ 50 samples per person per word (the total number of samples per word should be 15,000 ~ 30,000, as more data can improve the performance of the algorithm model).
- **Gender Ratio**: 50% male and 50% female, with a fluctuation of no more than 10%.
- **Age Distribution**: The data should include age groups that match the target audience of the product. For example:

  Children: 7 ~ 12 years old, 10%.

  Teenagers: 13 ~ 17 years old, 20%.

  Young Adults: 18 ~ 30 years old, 40%.

  Middle-aged Adults: 31 ~ 50 years old, 20%.

  Seniors: 51 years and older, 10%.
- **Regional Requirements**: Data collection should cover various regional accents based on pronunciation differences.
- **Speech Speed**: The speech speed should include normal, slow, and fast speeds. The definitions of fast and slow speech should be based on the recorder's normal speaking habits. For a single wake-up word, the fastest speed should be no less than 0.5 seconds, and the slowest speed should exceed 2.1 seconds. There should be approximately a 2-second pause between two consecutive wake-up words.
- **Volume**: The volume should vary between normal, soft, and loud. The definition of volume levels should be based on the recorder's normal speaking habits. The speech data should be 16-bit with a range of -32768 to 32767. When considering only the positive amplitude values, the overall amplitude distribution should resemble a Gaussian distribution, with most values concentrated between 6000 and 14000, while smaller values around 2000 and larger values around 20000 should also be present in a certain proportion.

Requirements for the Recorder:

- **Correct Pronunciation**: The wake-up word should be clearly pronounced. An accent is allowed, but it should not affect the intelligibility of the word for normal listeners.
- **Avoid Pronunciation Errors**: Ensure correct pronunciation without misreading.
- **Avoid Unnecessary Sounds**: Avoid coughing, talking, or other background noises.
- **Breathing and Mouth Sounds**: Natural breathing and mouth sounds are allowed, but recording should not occur while gasping for air.
- **Avoid Unnatural Pauses**: Normal breathing pauses are allowed, but they should not be excessively long. The pause duration should not exceed 0.5 seconds.
- **Avoid Stuttering or Interruptions**: The wake-up word should be pronounced smoothly, without stuttering or interruptions.

Re-record if Affected: If any sound during the recording process affects the audio quality, the affected file must be re-recorded.

4. Audio Files

Format: 16-bit, 16KHz, mono channel, WAV file.

File Naming: The file name should include the recorder's ID, gender, province or city region, wake-up word, speech speed or volume information, and the audio file number.
The naming should only contain: numbers, uppercase and lowercase letters, Chinese characters, and underscores ("_"). The file name should be accurate and avoid abbreviations.

Example format: Name_Gender_Region_Word_SpeechSpec_Number.wav

Folder Organization: Use the recorder's ID as the folder name, and place the recorder's audio files inside that folder. If there are multiple wake-up words, create subfolders named after the wake-up words, with each wake-up word's audio files stored in the corresponding subfolder.
Example format:

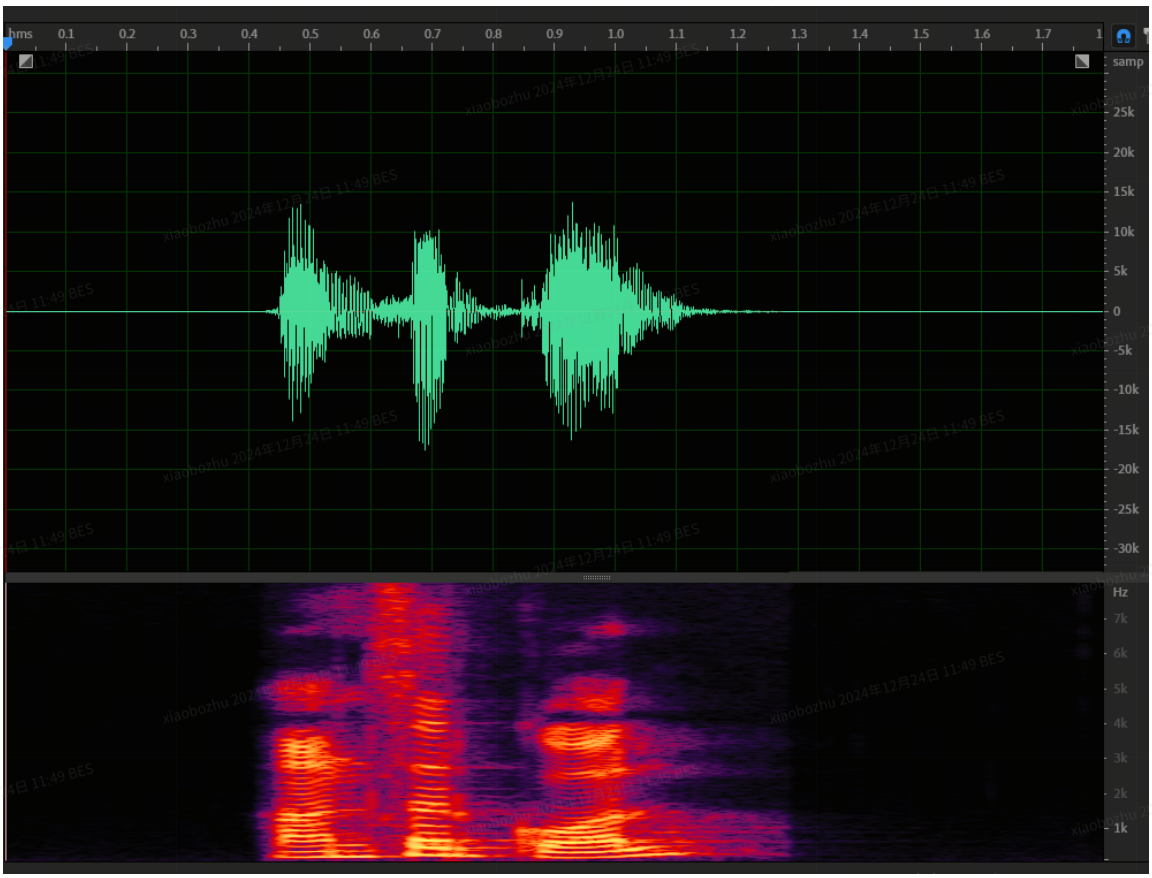
```
SPK001
SPK002
...
SPK090
        /HiSiri
        /PreviousTrack
        /NextTrack
        /PlayMusic
```

```
                /SPK009_M_NewYork_PlayMusic_Normal_001.wav
                /SPK009_M_NewYork_PlayMusic_Loud_002.wav
                /SPK009_M_NewYork_PlayMusic_Small_003.wav
        /StopMusic
                ...
SPK010
        /HiSiri
        /PreviousTrack
        /NextTrack
        /PlayMusic
        /StopMusic
                /SPK010_F_ShangHai_StopMusic_Normal_001.wav
                /SPK010_F_ShangHai_StopMusic_Loud_002.wav
                /SPK010_F_ShangHai_StopMusic_Small_003.wav
                ...
```

Other Notes:

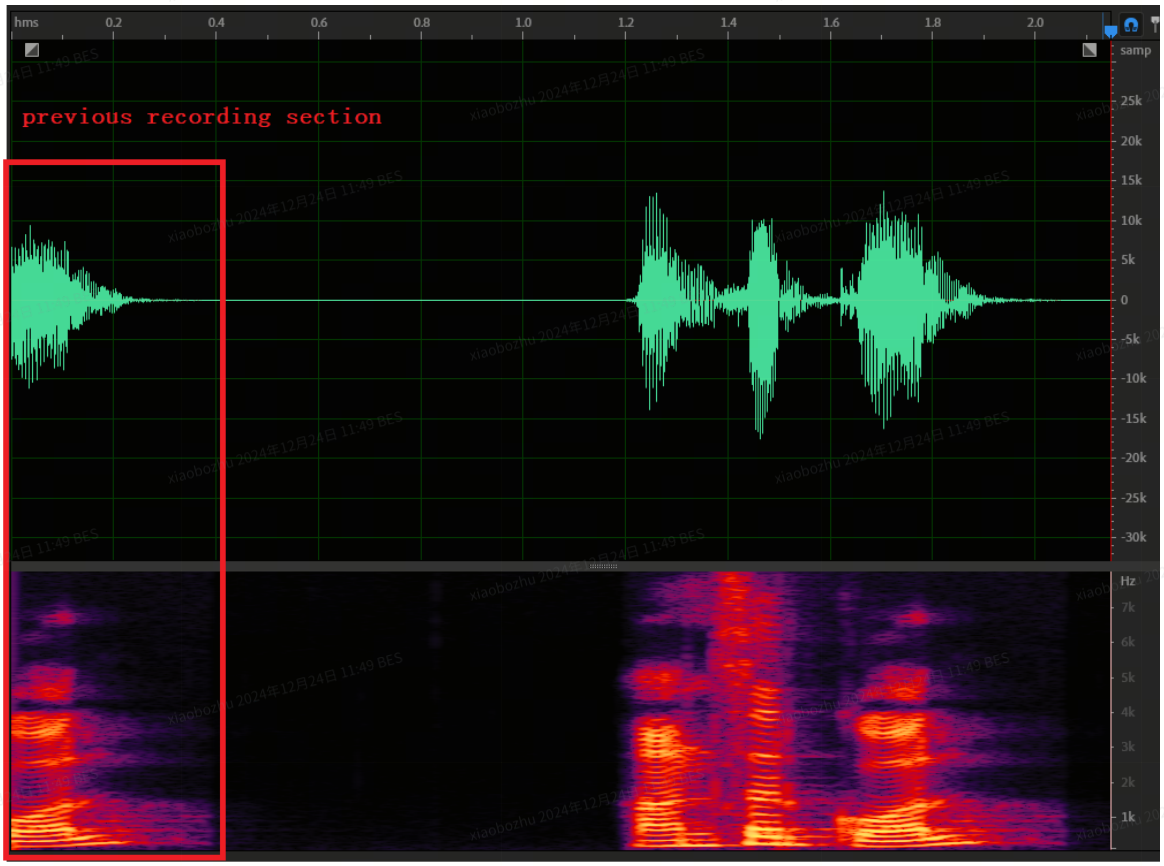
- Each audio file should only contain a single wake-up word.
- Empty audio files are not allowed.
- Avoid audio clipping: For example, do not cut "Hi Siri" to "Hi Si".

A correct audio file example is as follows:

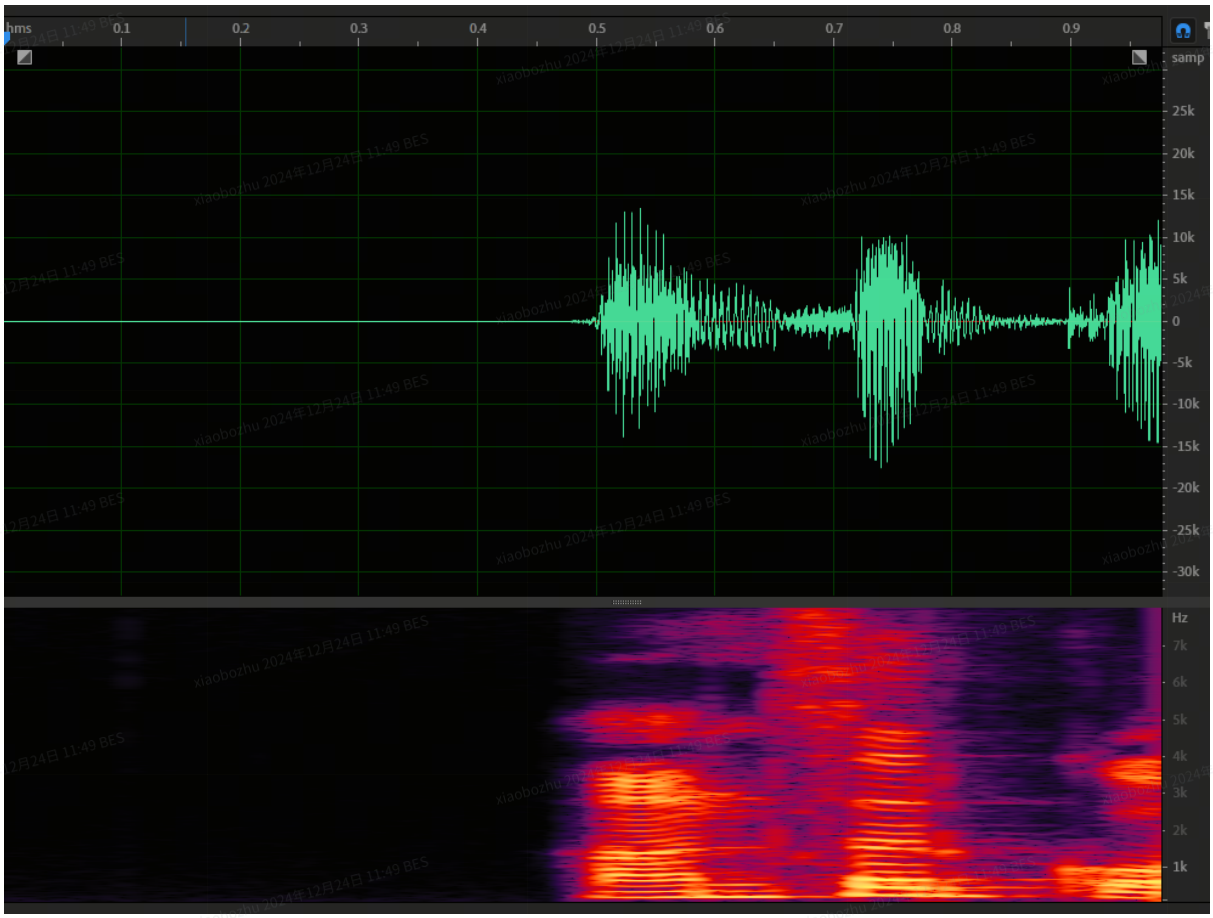


The following is an incorrect audio file example:

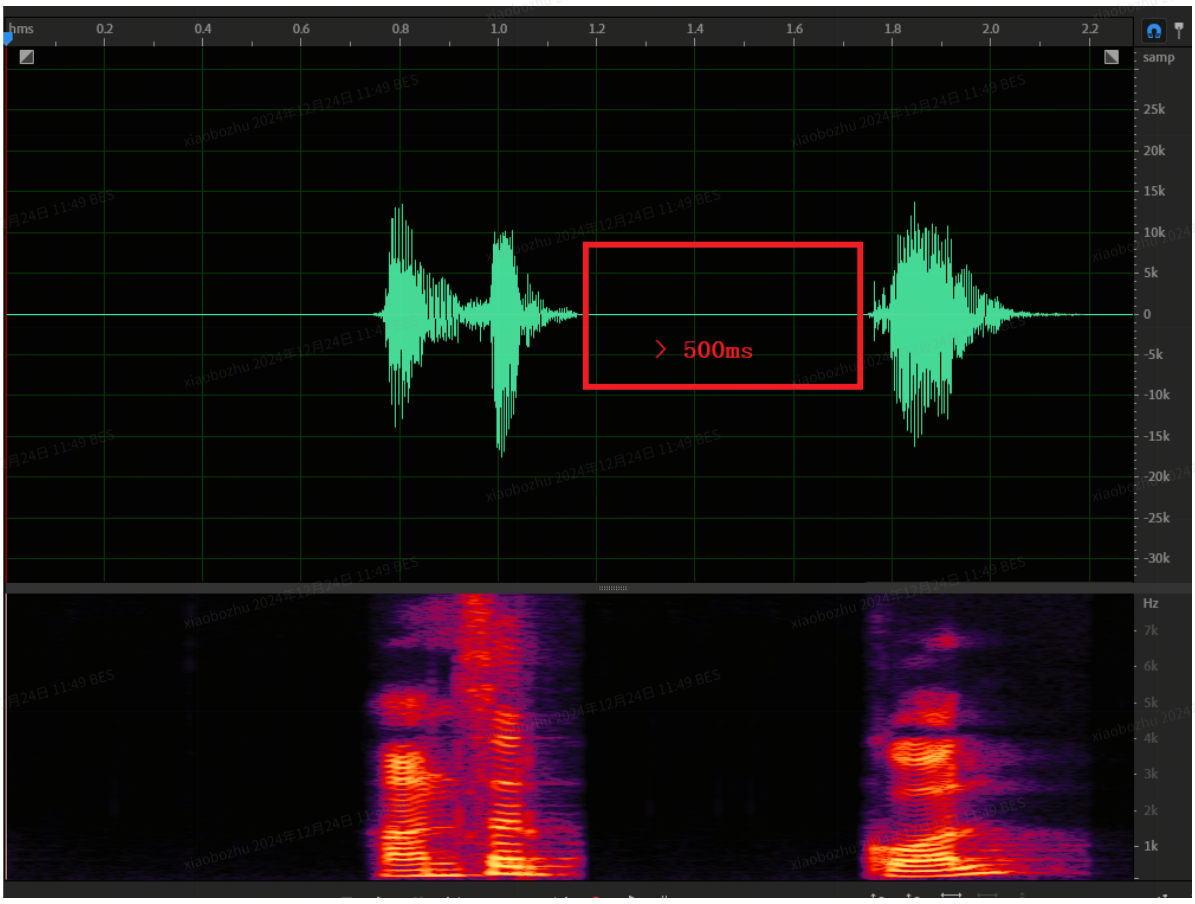
(1) The recording file has been incorrectly cut, with the wrong cutting point, resulting in part of the previous word being included.



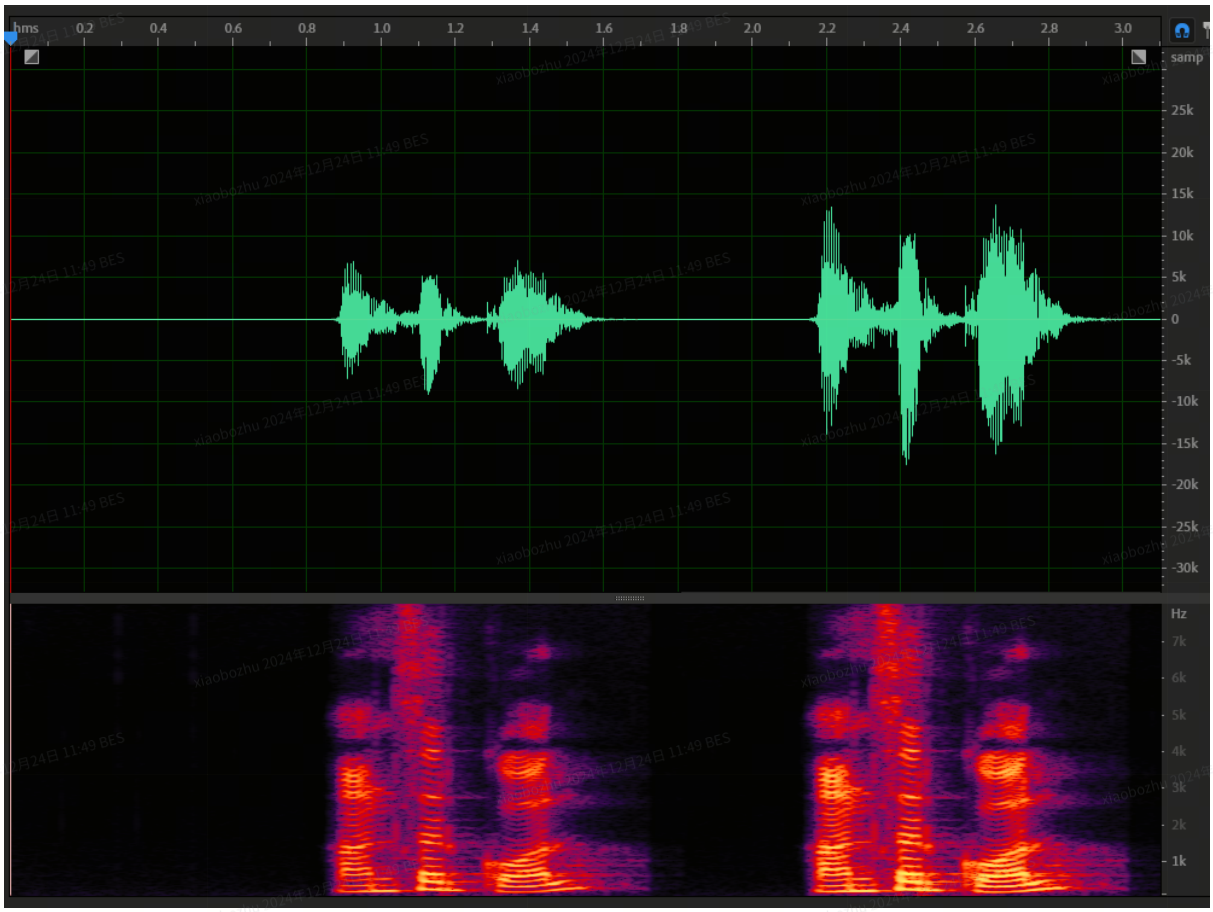
(2) The end of the wake-up word is cut off or lost, resulting in an incomplete wake-up word audio file.



(3) The pause between the single wake-up word is too long, clearly exceeding 500ms.



(4) A single audio file contains multiple wake-up words.


hms
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4
2.6
2.8
3.0
samp
25k
20k
15k
10k
5k
0
-5k
-10k
-15k
-20k
-25k
-30k
Hz
7k
6k
5k
4k
3k
2k
1k