Audio Summaries

Expert Evaluation

Authors
Modan Tailleur
Mathieu Lagrange
Pierre Aumond
Vincent Tourre

July 22, 2025

1 Definitions

Audio summaries are short audio representations of longer audio recordings. Those longer audio recordings are called **full-length audio** in this document. Two key conflicting criteria characterize an audio summary:

1. **Temporal Consistency (TC)**: This criterion reflects how faithfully the summary represents the temporal structure of the full-length recording. Specifically, a high TC implies that the most present sound sources in each time interval of the full-length audio are included in the summary and presented in chronological order. Note that a high TC does not necessarily imply a linear compression of time.
2. **Source Diversity (SD)**: This criterion reflects how well the summary captures the diversity of sound sources present in the full-length audio.

In this perceptual evaluation, you will be asked to assess the levels of TC and SD of 32 summaries, with respect to the content of their full-length audio recordings. Full-length audio recordings are 24 hours long. You are not requested to listen to them entirely. We thus provide means to navigate through those files using an open-source sound editor with basic spectrogram visualization.

2 Preparation of the expert evaluation (10min)

2.1 Audacity Setup Instructions and Dataset Download

First, download the latest version of Audacity from `https://www.audacityteam.org/download/`, if you have not already installed it. Open Audacity and navigate to `Edit/Preferences/Track (Edition/Préférences/Piste)`. Change the default view to `Spectrogram`. Alternatively, you can switch views by clicking on the track name. Do not change spectrogram parameters.

For Audacity to support MP3 files, you must also install FFmpeg if it is not already installed. You can download it from `https://www.ffmpeg.org/download.html`.

Next, download the dataset from the following link: `https://tinyurl.com/audsummary`

Ensure that you have at least 50 GB of free disk space before unzipping the downloaded file. The unzipping process should takes about 4 minutes The extracted zip file contains five folders: `A`, `B`, `C`, `D`, and `X`.

2.2 Opening of X full-length Audio and its summaries

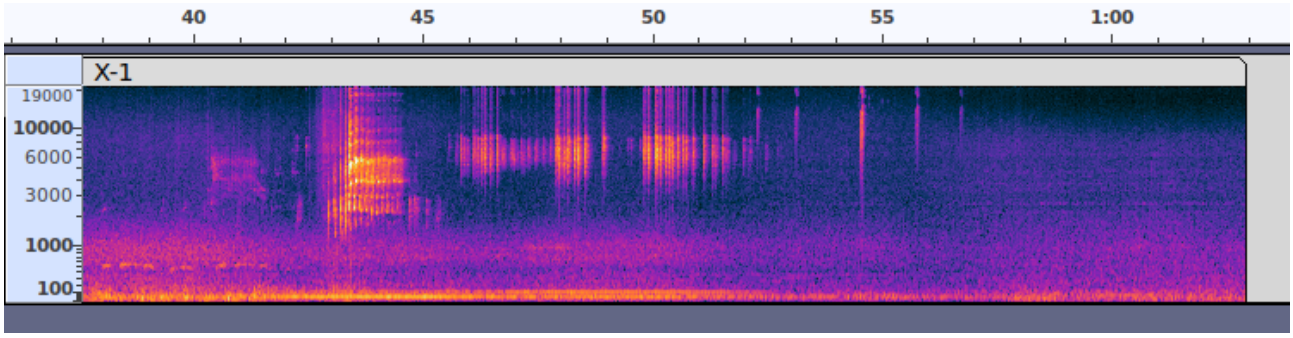


Figure 1: Summary X-1 in Audacity

The `X` folder contains example audio files that will be used for practice in Section 3.3.

Drag and drop the `X_full_length.mp3` file into a new Audacity session. The processing will take approximately 2 minutes.

In the meantime, open the `X_summaries.aup3` file. This file contains 7 summaries for the full-length audio `X`, separated by a one-minute silent gap in the timeline. Figure 1 illustrates how summary X-1 appears in Audacity.

At this point, you should have two Audacity windows opened: one containing the seven summaries and the other containing the full-length audio. Useful Audacity shortcuts for navigation can be found in Table 1. If needed, you can customize these shortcuts by navigating to `Edit/Preferences/Keyboard (Edition/Préférences/Clavier)`.

Navigation Recommendations:

- **Audio Summaries:** Select the clip of a summary, double-click on it, and press `Ctrl + E` to adjust the view to the scale of the summary.

Action	Shortcut
Play a clip	Double-click on a clip, then press `Space`
Zoom in	`Ctrl + 1 (or Ctrl + Shift + 1)`
Zoom out	`Ctrl + 3 (or Ctrl + Shift + 3)`
Move to next clip and select it	`Tab`
Zoom to selection	`Ctrl + E`

Table 1: Selection of Audacity shortcuts for navigation

- **Full-Length Audio:** Zoom to a 1-hour scale (each hour is delimited by a 5-second silence). Browse within the current hour before moving to the next.

3 Evaluation Instructions

3.1 Evaluating Sound Diversity

To evaluate SD, follow this two step process:

Step 1: Explore the full-length audio

Spend approximately 5 minutes browsing through the full 24-hour recording, and identify as many distinct sound sources as possible. The spectrogram visualization can assist you in detecting variations in sound content.

Step 2: Evaluate each summary

For each of the 8 different summaries:

1. Listen to the 1-minute audio summary.
2. Assess and rate the SD score by evaluating how well the summary captures the variety of sound sources present in the full-length audio.

You are encouraged to go back and forth between the summaries and the full-length audio to make a more informed evaluation.

3.2 Evaluating Temporal Consistency

To evaluate TC, follow this two-step process.

Step 1: Explore the full-length audio

For each 1-hour segment of the original recording

- Zoom into the timeline to view 1 hour at a time.
- Select 5 time points spread across the hour.
- At each point, listen to a 2-second audio excerpt.
- Note the most present sound source(s) heard across those 5 points for that hour.

Repeat this for all 24 hours.

Step 2: Evaluate each summary

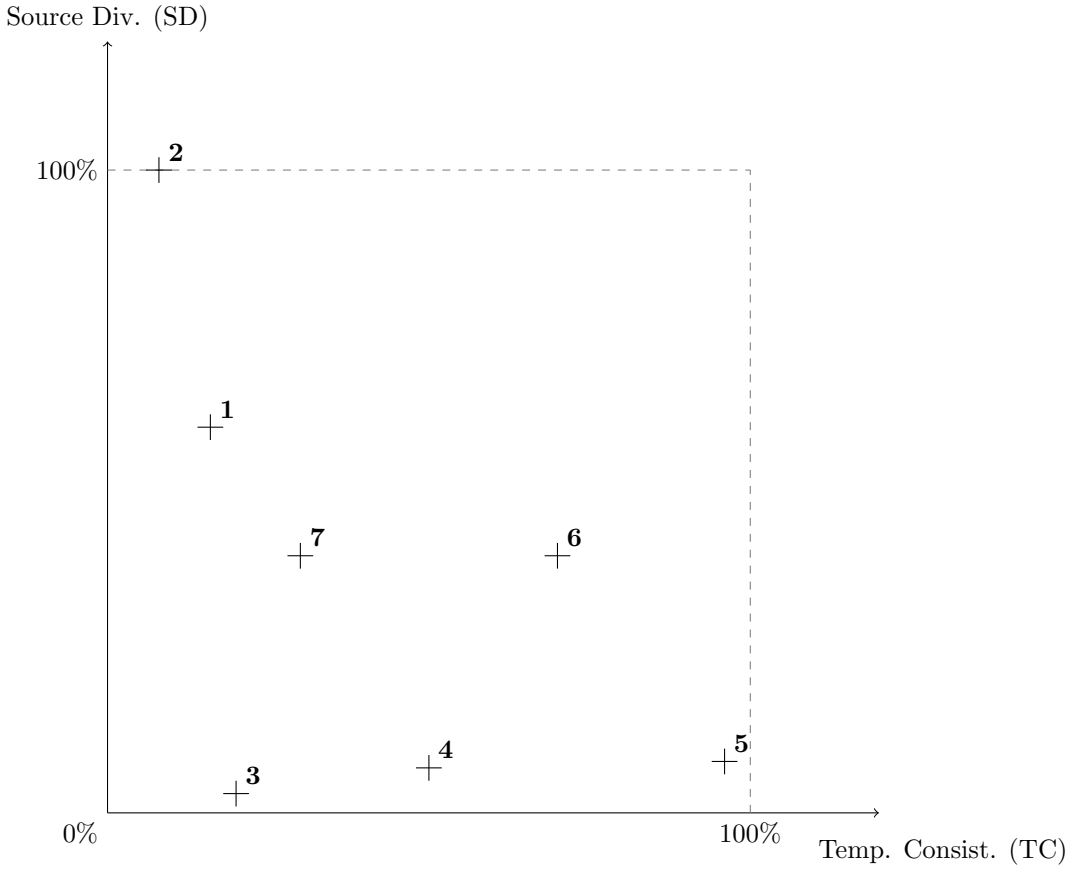
For each of the 8 different summaries:

- Listen to the 1-minute audio summary.
- Assess and rate TC by evaluating how well the summary reflects the temporal structure and most present sound sources identified across the original 1-hour segments. Note that a high TC score does not necessarily imply a linear compression of time, but rather a faithful representation of the overall temporal dynamics.

As with SD, you are encouraged to go back and forth between the summaries and the full-length audio to support your evaluation.

3.3 Example: evaluation of full-length audio for folder X

Below is an example evaluation of the 7 summaries of the full-length audio of folder X. You are not expected to evaluate the X summaries, they only serve as examples. The level of TC and SD for each summary is represented by a cross, with the corresponding index of each summary (e.g., 1 for X-1) placed above the cross. The maximum possible values for both TC and SD are indicated by dashed lines on the plane.



Please briefly browse through the `X_full_length.mp3` audio file in Audacity, while simultaneously browsing through each of the summaries (no need to listen to each one of them at this stage). As you do so, read through the following justifications for their respective placements on the TC/SD plane:

- **X-1:** Contains various bird species performing different actions (chirping, flapping wings, etc.), which gives it some SD. However, it lacks full diversity, as birds are not the only sound source present in the full-length audio. It does not have very high TC, as birds are prominent in the full-length audio, but don't have a high time of presence in every time interval, particularly at the beginning and end of the summary, which should represent nighttime.
- **X-2:** High SD, encompassing a diverse range of sound sources. However, it has slightly less TC than X-1, as birds are more present in each time interval of the full-length audio than other sound sources included in X-2.
- **X-3:** Low SD, as it contains only wind sounds. It also lacks TC, as the average content of the full-length audio is not predominantly windy in every time interval.
- **X-4:** Slightly more SD than X-3, as it contains both wind and background silence and thus more diversity. More TC than X-3, as the most present sound source in each time interval is closer to background silence than it is to wind.
- **X-5:** High TC, as it effectively conveys the sound source with the highest time of presence in each time interval and in chronological order. It includes more audio segments from the middle of the day, when the average sound activity is evolving, and fewer from the beginning and end of the day, when there is little variation in the average presence of sound sources.
- **X-6:** With moderate TC and SD, this summary showcases diversity while emphasizing sound sources with relatively high time of presence in each time interval. Overall, it leans slightly more towards TC than SD.

- **X-7:** Exactly the same SD as X-6, as it contains the same audio segments. However, it has less TC due to the altered chronological sequence of the audio segments. For instance, a car is arriving at the beginning instead of the expected nighttime silence that barely contains cars, and a rooster crows at the end of the summary, even though there are no roosters in the end of the full-length audio.

Notes:

- The audio summaries only contain audio segments that come from the full-length audio.
- The provided justifications serve only as examples to help you understand the logic behind the TC/SD plane. It will not be required to provide such justifications for any of your choices. However, an optional text box will be available below each plane if you wish to briefly explain your decisions.
- The full-length audio presented in this example is inherently uneventful. You may encounter full-length audio files that contain significantly more events. Pay close attention to the content of the full-length audio file, as the assessment of TC and SD is based on it. Thus, even if a summary contains diverse sound sources, it can still exhibit high TC and low SD depending on the content of the full-length audio.

4 Perceptual evaluation (1h20)

The perceptual evaluation being quite long, you are allowed to segment it into 4 different sessions: one for each folder A, B, C and D. Each session should take approximately 20min.

Instructions: evaluation of TC and SD

For a given full-length audio (e.g., `A_full-length.mp3`), you will position each summary on a TC/SD plane. To conduct the evaluation, please follow this procedure:

- Listen to one summary
- Browse through the full-length audio to compare its content with the summary.
- Mark with a cross its levels of TC and SD using a wooden pencil. Above the cross, indicate the corresponding index of the audio summary. For instance, for the full-length audio X, for the summary X-1 (represented on Audacity in figure 1) write the number "1" above the cross. Optionally, you can cut out the numbers provided at the bottom of this page using scissors and place them on the plane before finalizing your decision with a pencil.

You may repeat and alternate between these three steps as needed. Repeat this process for all 8 summaries of the folder. You can revise your evaluation at any time by erasing and repositioning your marks. You are free to listen to the summaries in any order and as many times as needed.

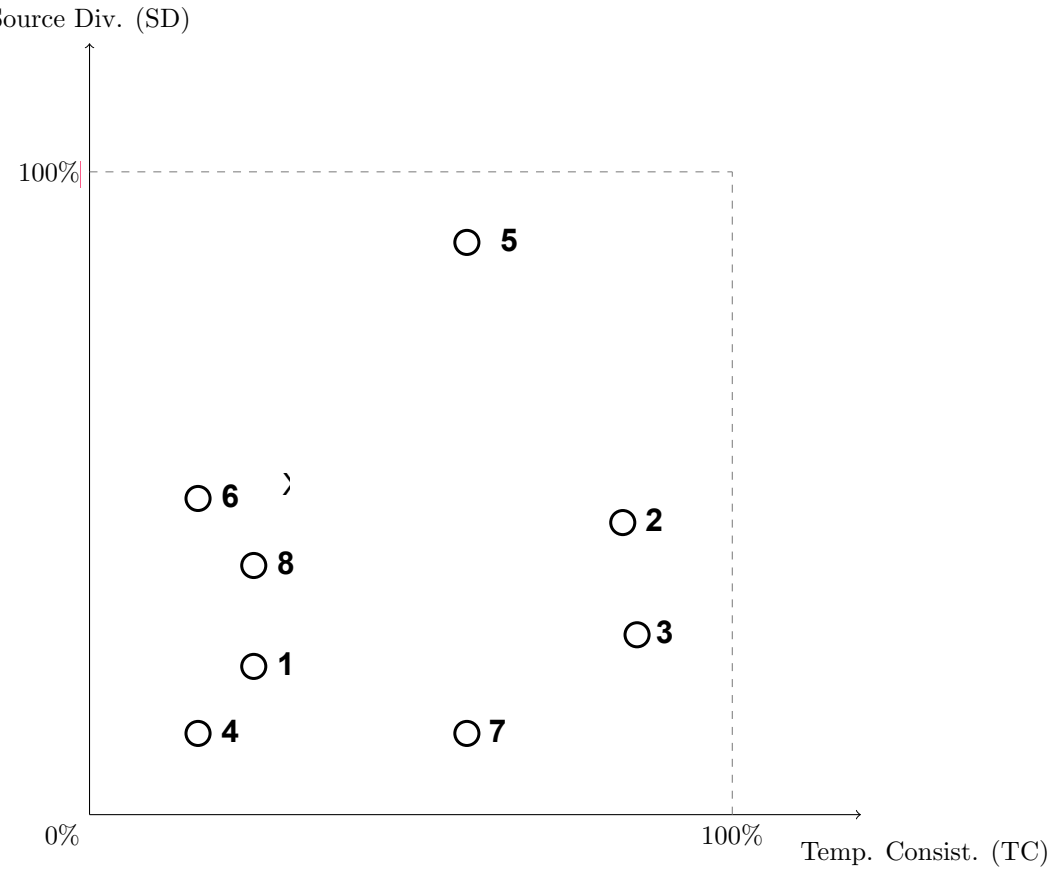
Note: You may also work with the digital version of this document and modify the PDF by adding crosses and inserting text boxes using PDF editing tools such as `Sejda` (`https://www.sejda.com/fr/desktop`).

1	2	3	4	5	6	7	8

4.1 Evaluation - A

Warning: To avoid mistakes, ensure that all previously opened Audacity windows are closed before starting the perceptual evaluation of this folder.

As described in Section 2.2, open the `A_folder`, drag and drop the `A_full-length.mp3` file into a new Audacity session, and then open the `A_summaries.aup3` file. You should thus have 2 Audacity windows opened. Evaluate here the 8 summaries of `A_summaries.aup3`:



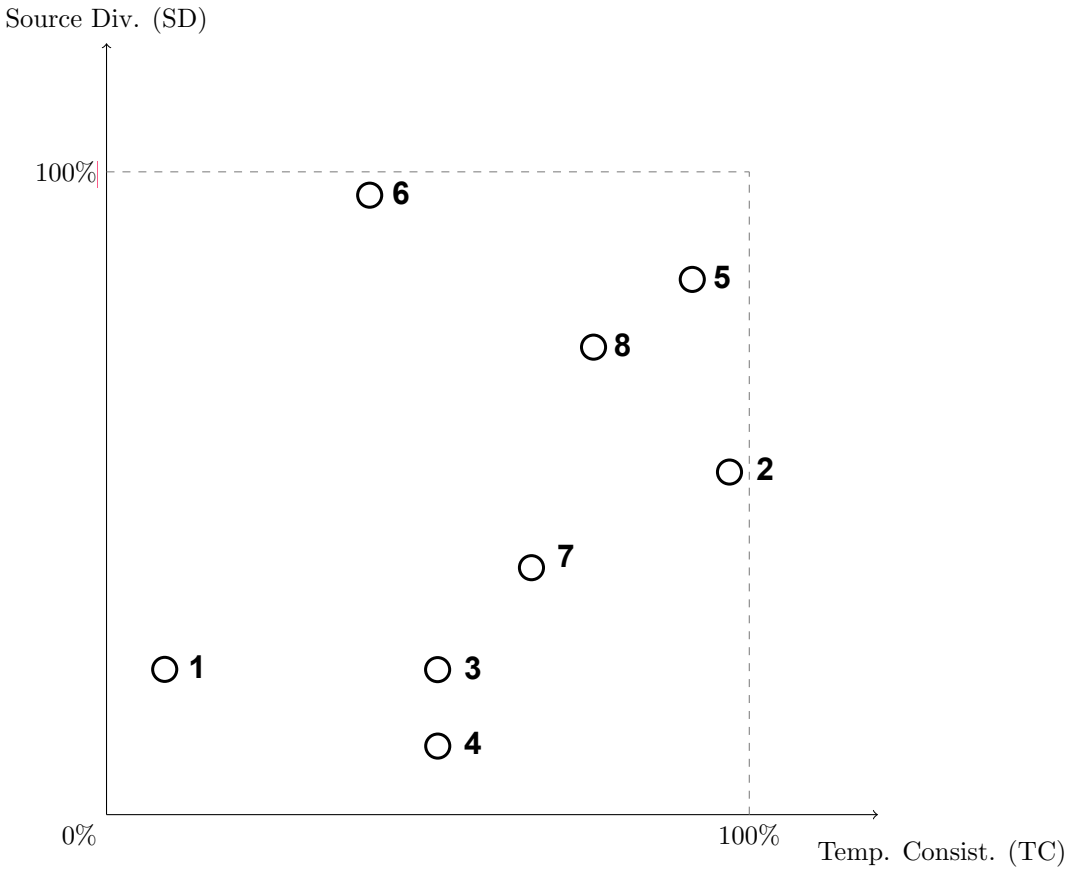
Please write a few words to explain your choices:

A-1: Contains rain sounds, tapping noises, and clock chimes, which gives it some SD.It does not have very high TC, as construction sounds appear in the latter part of the summary, which are not consistent with the full-length audio. The second appear time of the chime is too late.
A-2:Has a slightly higher SD score compared to A-1, with the presence of human speech and car driving sounds contributing to greater sound diversity. TC is also higher, as these sound events are more consistently present throughout the summary.
A-3: Has a similar level of SD to A-1. Although it lacks the clock chimes found in A-1, it includes ambulance sirens and car driving sounds. TC is slightly better than A-2.
A-4: Has low SD, as only human speech and tapping sounds are heard. TC is also low, and construction sounds in the latter part should not be present.
A-5: Has the highest SD score among all summaries, with a very rich variety of sound types. TC is also relatively high. While the timing of each sound generally aligns with the full audio, some segments are overly long—for example, the shouting at the beginning takes up too much time and does not accurately represent the overall soundscape of the early morning period.
A-6: Has higher SD than TC, with the presence of an ambulance siren at the beginning and construction sounds appear in the latter part, which do not align with the actual full-length audio.
A-7: Has limited variety in sound types. TC is higher than SD.

4.2 Evaluation - B

Warning: To avoid mistakes, ensure that all previously opened Audacity windows are closed before starting the perceptual evaluation of this folder.

As described in Section 2.2, open the `B_folder`, drag and drop the `B_full-length.mp3` file into a new Audacity session, and then open the `B_summaries.aup3` file. You should thus have 2 Audacity windows opened. Evaluate here the 8 summaries of `B_summaries.aup3`:

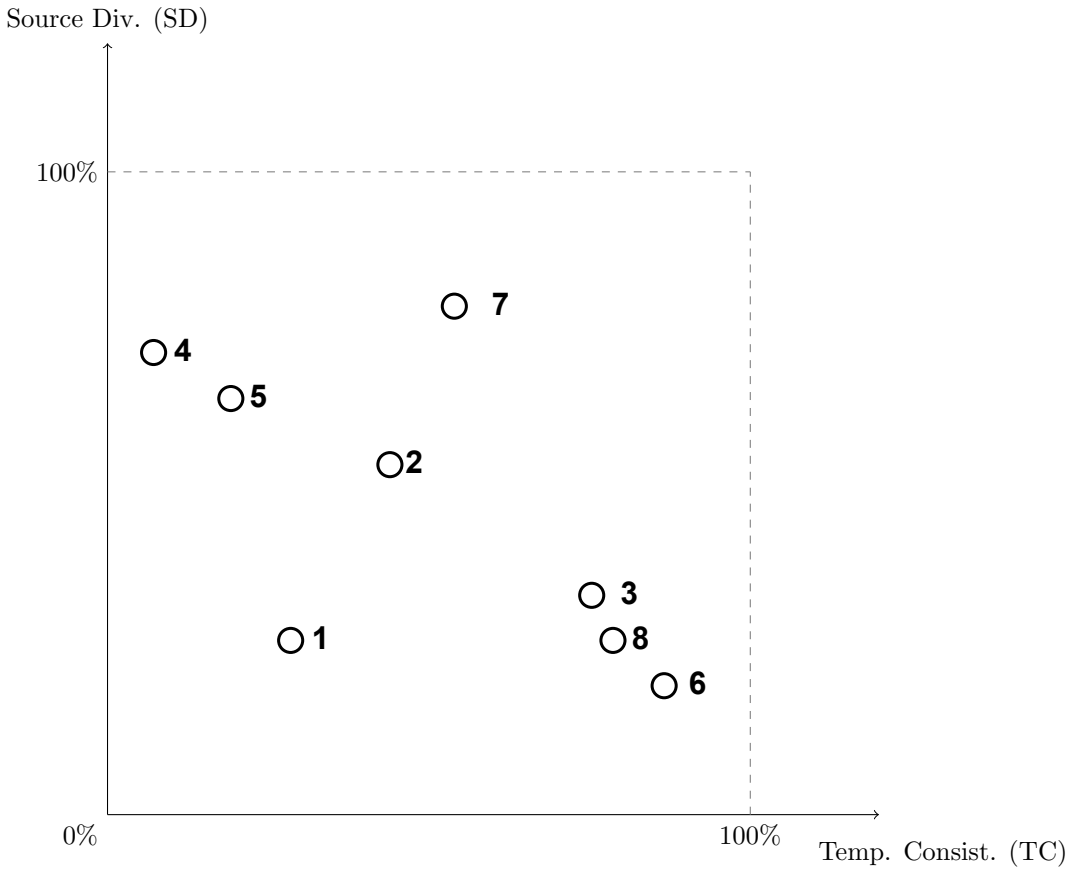


Please write a few words to explain your choices:

4.3 Evaluation - C

Warning: To avoid mistakes, ensure that all previously opened Audacity windows are closed before starting the perceptual evaluation of this folder.

As described in Section 2.2, open the `C_folder`, drag and drop the `C_full-length.mp3` file into a new Audacity session, and then open the `C_summaries.aup3` file. You should thus have 2 Audacity windows opened. Evaluate here the 8 summaries of `C_summaries.aup3`:

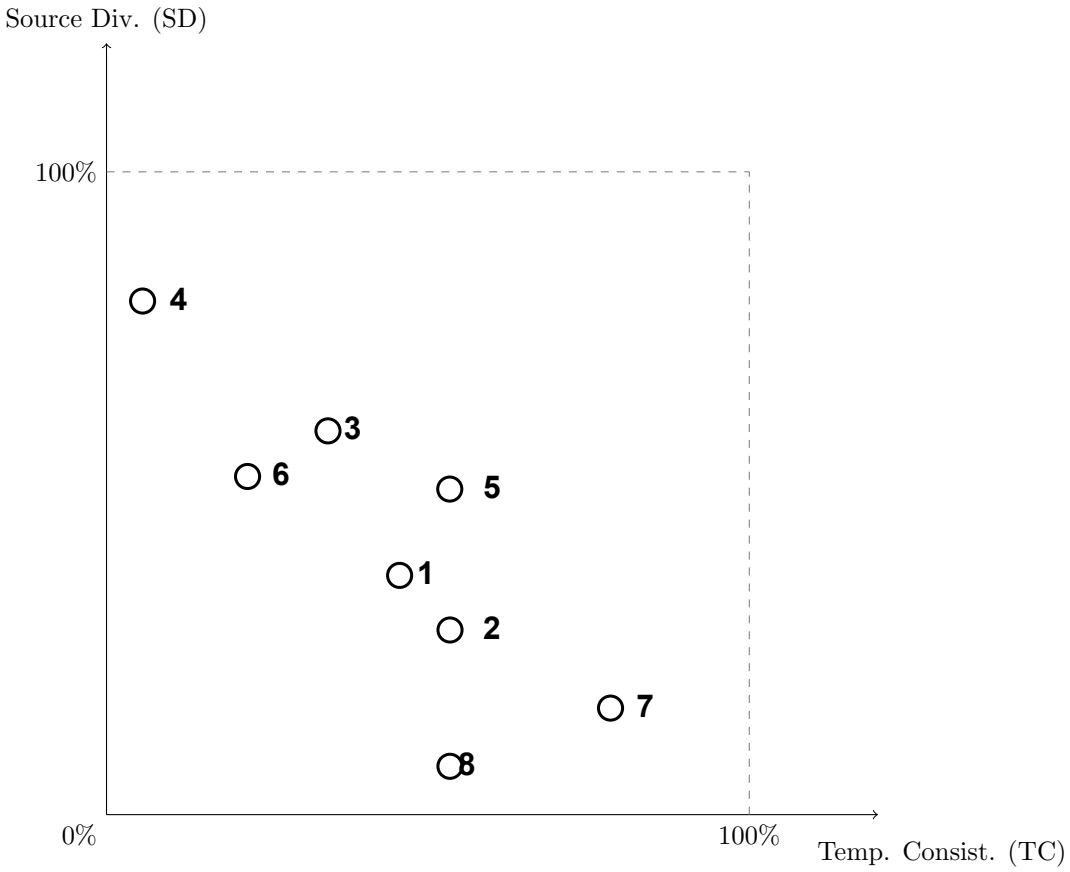


Please write a few words to explain your choices:

4.4 Evaluation - D

Warning: To avoid mistakes, ensure that all previously opened Audacity windows are closed before starting the perceptual evaluation of this folder.

As described in Section 2.2, open the `D_folder`, drag and drop the `D_full-length.mp3` file into a new Audacity session, and then open the `D_summaries.aup3` file. You should thus have 2 Audacity windows opened. Evaluate here the 8 summaries of `D_summaries.aup3`:



Please write a few words to explain your choices: