Chapter 3

Data Collection

In this chapter, we first present the detailed process of constructing MS21: Data Gathering 3.1.1, Semi-automatic Annotation Generation 3.1.2. We also introduce Hierarchical Structure of Multi-track Dataset 3.1.3, which was utilized to implement Automatic Stem Generation 3.1.4. In Section 3.2, We perform quantitative analysis within the three subsets of MS21 and among a collection of dataset used in Music Source Separation (MSS).

3.1 Dataset Construction

3.1.1 Data Gathering from *The 'Mixing Secrets' Free Multitrack Download Library*

The 'Mixing Secrets' Free Multitrack Download Library [1] is an additional source for the book *Mixing Secrets For The Small Studio*, written by Mike Senior. There are more than 500 multi-track projects which can be freely downloaded for mixing practice purposes. All projects contain uncompressed WAV files (24-bit or 16-bit and 44.1kHz or 48kHz sampling rate) and mp3 mixture for reference. A subset of projects contain un-mastered wav for mastering practise. To maximize the mixing

[1] https://cambridge-mt.com/ms/mtk/

flexibility, the contributors follow a guideline to provide audio with well-defined filename and aligned.

Script geturl.py in the repository was used to gather the necessary information from the website, such as *Track Name, Artist Name, Broad Genre, Sub Genre, Archive URL, Preview MP3.* Then, a unzipping process was carried out. After getting rid of broken archives, we have more than 500 songs in total.

3.1.2 Semi-automatic Annotation Generation

An automatically instrument classification at the lowest hierarchical level was conducted according to our observation of the naming convention of multi-track with several archives. Specifically, we created text filtering conditions to group the instruments from the same level of ontology together. For example, we group *Hammond* to *Electric Organ*; *Rhode*, *Wurlizer* and *Keys* to *Electric Piano*. Results were documented in a CSV (Comma Separated Value) file. Then, manually correction was carried out to modify the automatically generated labels, based on the actually recorded music content within the multi-track. Besides, with special attention on the leakage situation of vocal tracks, a label called *Vocal Quality* was annotated. It should be noted that, because of recording setup such as live recording, multi-tracks of certain instruments (each instrument of drum set) may contain leakage. Such contaminated tracks makes it impossible to separate that individual instrument (e.g. Tom-Tom, Hi-Hat).

While we sill labelled those drum instrument as long as their leakage is from other drum instrument, we created a category named *Ambient Microphone* to get rid of those contaminated tracks containing tracks from totally different instrument family members such as *Room, Mainpair, Ruffmix.* For the "outsider" tracks of the current instrument ontology, such as *Rain, Kick Trigger*, we discarded them by labelling them as *Unused.* Note that we still keep those contaminated vocal track in use with additional *Vocal Quality* label waiting for future solution of training with leakage data. *Lead Back both* and *Guitar Both* are generated automatically by checking the corresponding instrument columns. *Broad Genre, Sub Genre* metadata was aggregated from the previous data gathering process. Finally, over 500 songs were annotated using more than 70 instrument labels, 2 genre labels and 1 leakage label.

3.1.3 Hierarchical Structure of Multi-track Dataset

As briefly mentioned in [16], there is a hierarchy of the audio files for each song, namely *Raw*, *Stems* and *Mix*. This hierarchy leads to the folder structure of MedleyDB, where RAW folder contains unprocessed multi-tracks and STEMS folder contains professionally mixed stems, each stem corresponding a specific sound source. Since stems are stereo audio components of the final mix and include all effects processing, gain control, and panning, researchers in MSS tend to view stems as the smallest units to work on. The mixing or grouping standard were not documented in [16] and no analysis was done to present the track-to-stem process.

After closely inspecting the relation between the instrument label of raw audio files and the one of stem audio files in MedleyDB, we found that: 1) Contradicted labels of raw tracks exist in one stem, e.g. raw tracks labelled as *clean electric guitar* are grouped in a stem labelled as *distorted electric guitar* or vice versa, but they should be separated; 2) Different levels of instrument label exists: Some labels overlap with other in terms of ontology, e.g. *auxiliary percussion* and *drum set* share many common instruments, while some label is a subset of another, e.g. *drum set* and *drum machine*[2].

One possible explanation is that, when creating labels for stems, annotators did not follow an instrument ontology. Raw multi-tracks were mixed to create stems mostly because they should be viewed as an independent source according to the situation of that specific song. For example, for music genre such as electronic, *drum machine* is an independent source while in other genre of popular music, *drum machine* is a complementary instrument along with *drum set*. For world/folk music, stems which includes *cymbal* and *tom-toms* are labelled as *auxiliary percussion* rather than *drum set* because the concept of *drum set* is less used in that genre. As a result, there are

[2] The detailed documentation can be seen in Appendix

82 instrument labels for stems, but there are not 82 distinct instruments.

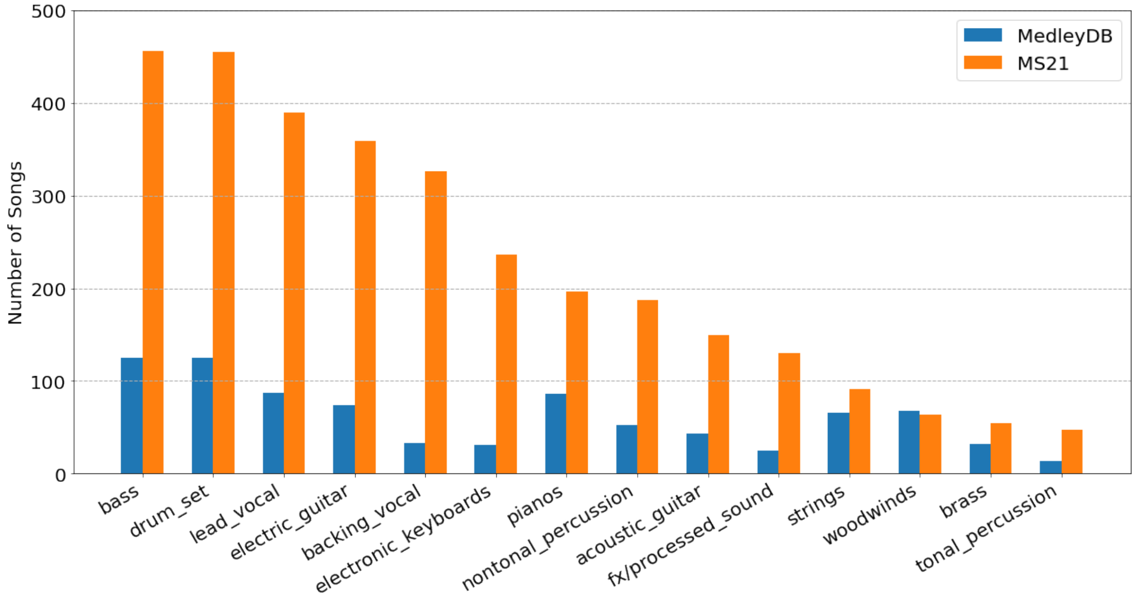

Figure 3: Instrument Distribution of MS21 and MedleyDB

In order to compare the instrument distribution between MedleyDB and MS21, we created a hierarchy to define the relations of different level instrument labels, grouping the lower level labels to higher-level labels. The audio structure is a three-level hierarchy, namely *track-to-inst.*, *inst.-to-stem*, *stem-to-stem*. In *track-to-inst.*, we group the most fine-grained instrument labels into 14 instruments[3], as shown in Figure 3. In *inst.-to-stem* stage, we follow the four-stem tradition, and group the 14 instruments into *vocal*, *bass*, *drums* and *other*, in the hope that traditional MSS research can also leverage this dataset. In *stem-to-stem* phase, we further create *accompaniment* by grouping *bass*, *drums* and *other* together, leaving *vocal* as another source in this level. This hierarchy serves as the hierarchy in this thesis. Note that this hierarchy is not the only hierarchy and can be studied its affects in HMSS in the future.

[3] Note that for MedleyDB, we regard the vocal tracks with "melody" label as "lead vocal" and those vocal track without "melody" label as "backing vocal"

3.1.4 Automatic Stem Generation

Considering memory overflow issues when mixing the instrument stems from the raw multi-tracks, we provide three levels of pre-computed stems: Instrument stems, MUSDB-ish stems and the mixture file, following the hierarchy mentioned in the previous subsection. In the first stage *track-to-inst.*, multi-tracks that belong to the same instrument class are mixed into one instrument stem. Loudness normalization of ITU-R BS.1770-4 standard implemented in pyloudnorm repository was applied when creating stems. According to [51], a relatively simple ITU-R BS.1770 integrated loudness was preferred over other complex psycho-acoustic model. Integrated Loudness was used for non-percussive instrument stem and target loudness was set to -25 LUFS. Whereas peak normalization was applied to percussive instrument stem such as drum set and non-tonal percussion. The target loudness level for peak normalization is set to -1 LUFS. In the second stage *inst.-to-stem*, MUSDB-ish stems were also generated specially for traditional MSS research. The same loudness normalization process was carried out. By directly loading the MUSDB-ish stems rather than mixing the relavant multi-tracks on the fly, computational resources can be significantly reduced and the training process can speed up. Finally, the mixture file is created by mixing the MUSDB-ish stems. We neglect the third stage *stem-to-stem* because memory cannnot not be a big problem at this stage.

Note that only linear mixing was applied after loudness normalization, which means no equalization, dynamic range compression and distortion was applied during mixing. As to our best knowledge, a baseline of auto-mixing system does not exist. All generated stems are kept from being professionally mixed by mixing engineer, which was left for future optimization or served as a baseline for auto-mixing related research.

Following the standard multi-track dataset folder structure proposed in MedleyDB, we put multi-tracks in RAW folder, the automatically generated stems in STEM folder, the metadata file and mixture audio file were left in the root directory of each song.

A new multi-track wise data augmentation was possible in multi-track dataset. In traditional MSS, data augmentation involves pitch shifting, gain adjusting and channel swapping can only applied to the stems such as *vocal, bass, drum, other.* However, for multi-track dataset, stems can be re-mixed by applying different mixing parameters to its corresponding multi-tracks. In this way, a stem with different content can by generated, for example, we can generate different versions of drum set stem where each drum instrument can be re-balanced. We believe that this data augmentation can be extremely helpful for music source separation, especially when the dataset in this field are not big enough in terms of total mixture duration. However, if we consider the number of multi-track, the size of the potential generated dataset is **unlimited**.

3.2 Dataset Statistics

3.2.1 Split of the dataset

We did the split of the dataset following several principles:

- Same artists would not be in both train and test sets, otherwise the system would overfit on the artist, as mentioned in [52]. The detailed split of the dataset can be found in Appendix.

- Different sub sets should share similar content distribution, e.g. genre and instrument. We here present the genre distribution of MS21, MUSDB18 and MedleyDB and the instrument distribution of MS21, as illustrated in Figure 4 and Figure 5. Note that for some artists, their music genre are hard to define so in The 'Mixing Secrets' Free Multitrack Download Library [4], their Sub Genre label is "Various Styles" while sharing different Broad Genre labels. The detailed report of those songs can be seen in Table **??** in Appendix. Besides, as shown in Table 6, the raw annotations of Sub Genre is too detailed, in order to compare the genre distribution across different datasets, we clustered the

[4] https://cambridge-mt.com/ms/mtk/

the sub genre into a broader genre based on the text information. Finally, distributions over 16 genres was computed, for those genres that do not fall into the category of the 16 genres, we labeled them as "Other" in the final column of Figure 4 and Figure 7.

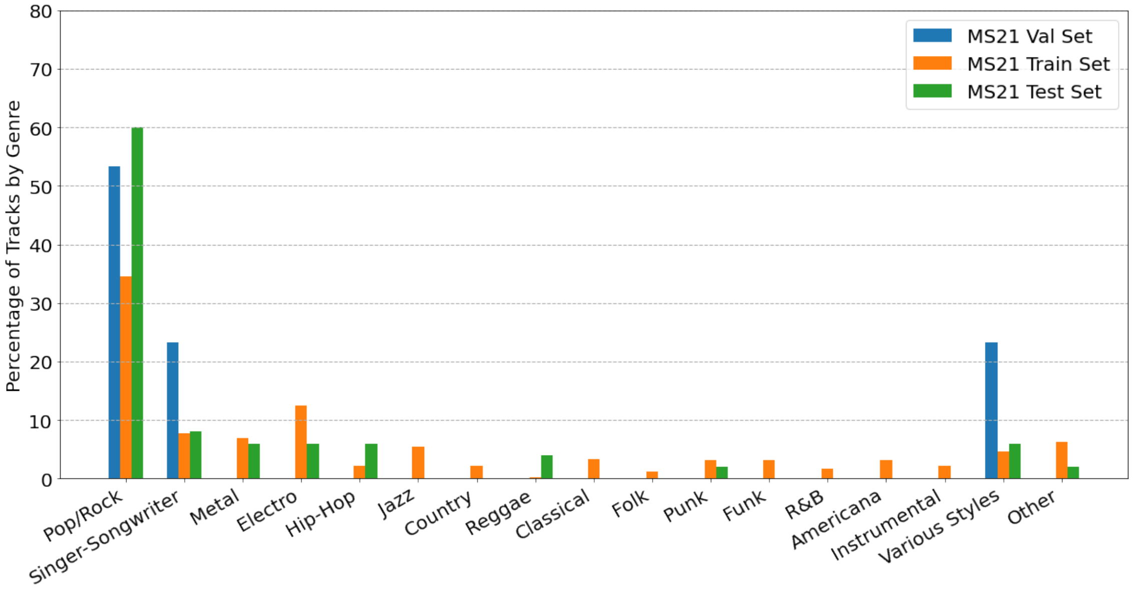

Figure 4: Genre distribution of each sub set of MS21

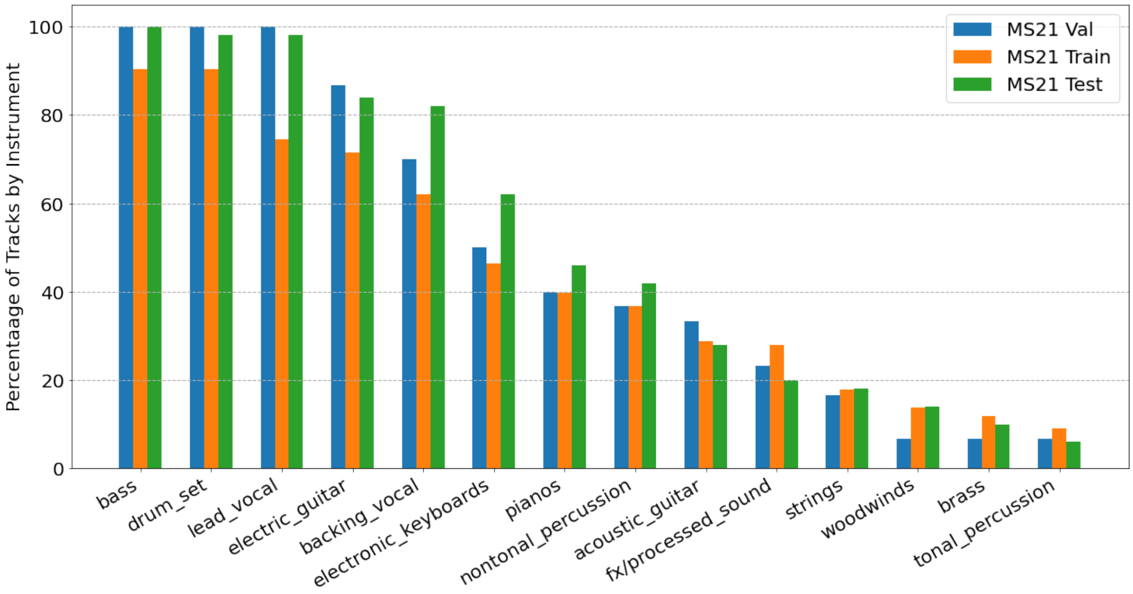

Figure 5: Instrument distribution of each sub set of MS21

- When creating test set, we tried to include all the songs in the test set of MUSDB18, since these two datasets share one data source and the unified test set can be useful when comparing results. In the end, only 7 of 50 from the test set of MUSDB are missing from MS21 dataset, as shown in table:

3.2.2 Summary: MS21 in Cross Datasets Comparison

Comparison across different dataset should follow specific criteria. By studying the statistics of the dataset, research know the characteristics and bias of the dataset. In [16], criteria such as *size*, *duration*, *quality*, *content*, *annotation*, *audio* are addressed when creating MedleyDB multi-track dataset. Besides,according to [53], in addition to total training duration, *separation quality* and *diversity* were mentioned as important criteria of datasets in MSS. We here follow this framework when evaluating our proposed dataset, also with our further development of the theory framework. Here we present our criteria when building our new dataset along with further discussion.

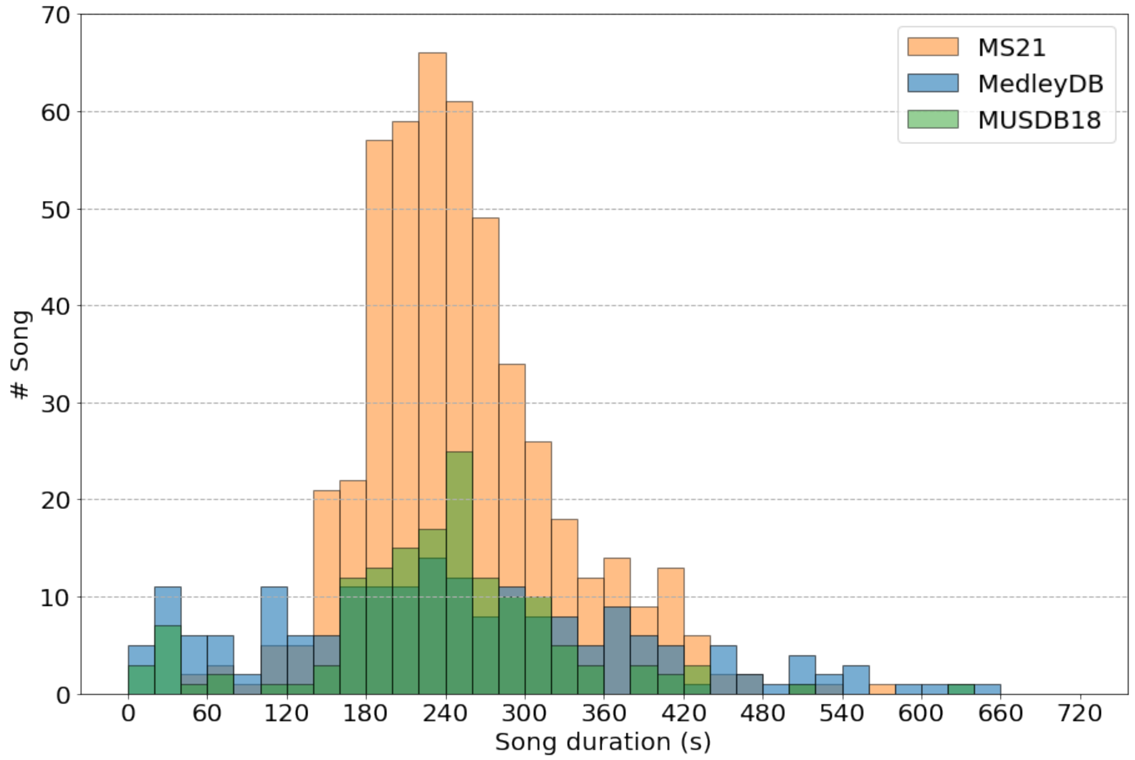

Figure 6: Length Distribution of the dataset

- **Size**: the dataset should be at least one order of magnitude greater than previous dataset.

Dataset	Size	Duration(h)	Num. Tracks	Instr. Cat.	Vocal %	Leakage %	Stereo/Mono
MS21	500	34.2	20±13	70	84	20	Stereo
MedleyDB	196	12.7	18±17	82	44	41	Stereo
MUSDB18	150	9.8	4	4	0	0	Stereo
Slakh2100	2100	145.0	11±3	34	0	0	Mono

Table 3: Statistics across multi-track dataset in Music Source Separation

- **Duration**: the dataset should primarily consist of full length songs. Besides, total training duration matters.

- **Quality**: the audio should be of professional or near-professional quality. Quality can be divided into **Performance Quality**, **Recording Quality** and **Production Quality**. For performance quality, dataset should be consisted of music composed and performed by professional or near professional artists. For recording quality, the recording session should be carried out in a recording studio with proper acoustic treatment(e.g. Noise ground should be lower than 40dBA(?) in order to prevent **non-musical leakage**) and adequate audio engineering devices. The audio format of the dataset should be at least 44.1kHz, 16bits and stereo, which is the industrial standard for commercial music recording. Note that, musical leakage or **bleeding** can be found in multi-track due to live-recording schemes. This phenomenon should be carefully addressed especially for multi-track dataset. For production quality, the stems and mixture should be produced by professional or near professional sound engineers in a recording studio.

- **Content Diversity**: the dataset should consist of songs from a variety of genres. A diverse genre distribution of the dataset can lead to a relatively balanced instrument distribution, which is of great concerned in the task of source separation.

- **Annotation**: the annotations must be accurate and well-documented.

- **Audio**: each song and corresponding multi-track session must be available and distributable for research purposes.

We here further present the distinct features of MS21 following the criteria proposed above. Statistics of MS21, MedleyDB, MUSDB18 and Slakh2100 were shown in

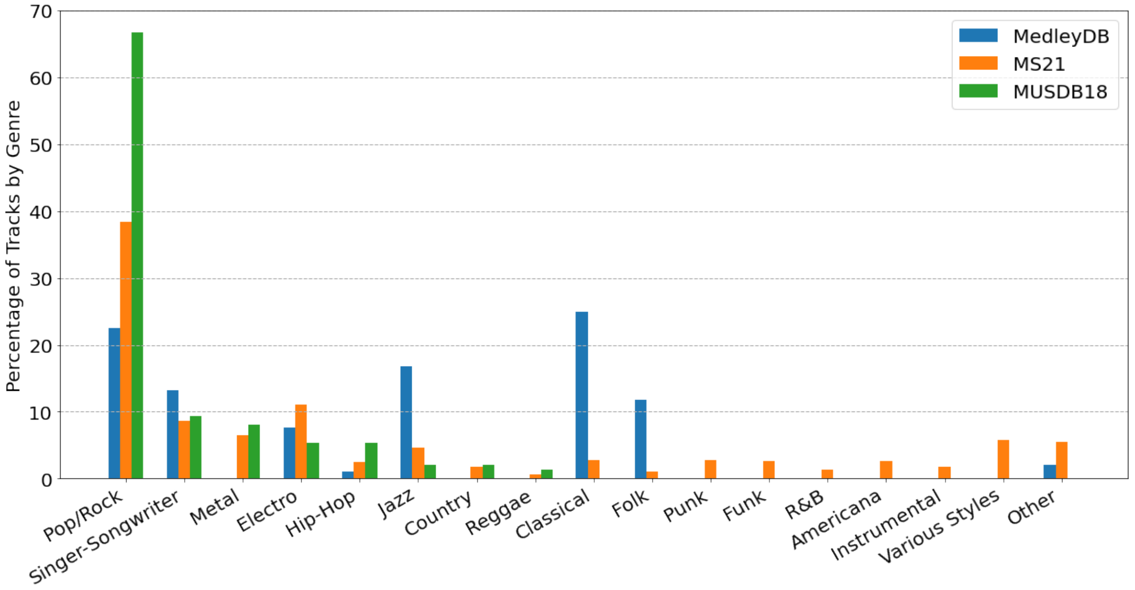

Figure 7: Genre distribution comparison across dataset

Table 3. In terms of **total mixture duration** of the dataset, MS21 is the largest one compared to MedleyDB and MUSDB18. However, if we consider the number of multi-tracks per song, the size of MS21 is significantly larger than MUSDB18. As shown in Figure 6, the **length distribution** of MS21 and MUSDB18 is more clustered to the mean value (240 seconds) while MedleyDB shows a flat and scattered distribution. What's more, MedleyDB and MUSDB18 even contains songs under 40 seconds. For **content diversity**, as seen in Figure 7, although *Pop/Rock* genre remains a predominant status in MS21, a variety of genre such as *Punk, Funk, RB, Americana* do exists. In contrast, MUSDB18 shows huge bias in *Pop/Rock* genre (over 65%). MedleyDB shows a distinctive focus on *Folk* (11%), *Jazz* (17%) and *Classical* (25%), while other two dataset contains very few songs in these three genres. Genre distribution leads to different Instrument distribution. As shown in Figure3, compared with MedleyDB, MS21 contains more number of instrument in every category except for woodwinds, which is a distinctive instrument family in *Classical* and *Folk* genre.

Derived from Mixing Secrets Website, one characteristics of MS21 is that it contains songs with **human performance** and **professional recording quality**. Besides, for different recording setup, the raw multi-tracks may contain different levels of leakage or *Bleeding*. In MedleyDB, a song-level label of bleeding was provided and it shows 41% of songs contain contaminated stems. One possible reason is due to the live-recording setup, especially MedleyDB contain many songs in genre of *Classical* and *Jazz*. In MS21, leakage situation is currently only annotated for vocal source, not for all multi-tracks of a song. The annotation was done by a single annotator (the author), who listened to all the vocal multi-tracks of all the songs. If any bleeding was heard, then this song would be labeled as "0" in *Vocal Quality*. As shown in Table 3, it means for those songs that contain vocal tracks, 20% of them contain leakage from other instruments in MS21.

For **annotation**, we provide detailed annotation as mentioned in subsection 3.1.2. This annotation was kept in the similar style of MedleyDB. Compared to MUSDB18, which does not provide more information other than the four well labelled stems, one advantage of MS21 is that instrument information about each stem was provided. With these information, a researcher can know what instrument each stem contains, which would be helpful when dealing with *Other* stem in traditional MSS.

For **audio**, the audio files and annotations of MS21 is freely available for research purpose. No commercial purpose was allowed for using this datset. For other purpose, one should contact the artist directly.

(I want to argue that, not professional produced dataset also works, unless the content was professionally performed and recorded. I need to carry out some experiment to demonstrate it.)