A Quest through Interconnected Datasets: Research on Annotation Practices in Highly Cited Audio Machine Learning Work and Their Utilized Datasets

Annotation Practices in Datasets Utilized by The International Conference on Acoustics, Speech, and Signal Processing (ICASSP) Conferences: A Transparency Analysis

Doga Tascilar [1]

Supervisor(s): Cynthia Liem [1]**, Andrew Demetriou** [1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Doga Tascilar
Final project course: CSE3000 Research Project
Thesis committee: Cynthia Liem, Andrew Demetriou, Frank Broz

Abstract

This research examines transparency between ICASSP conference papers and the dataset documentations related to the datasets' annotation practices. Top-cited 5 papers and 51 unique resources in total were considered. All of the selected papers utilized at least one dataset. For every paper, an extensive metadata search has done to reach the initial datasource of those datasets. These searches happened both within the paper contents such as sections and references along with outside the paper contents through the way of extensive web queries. Analysis of the papers published from 2021 and 2022 and their relevant datasets revealed varying levels of transparency. Original dataset creators provide comprehensive information, while papers using modified datasets offer limited details on initial annotations. Emphasizing the need for accountability, this study suggests that papers utilizing datasets should trace back to the initial dataset and provide explicit comments. The findings underscore the importance of ensuring sufficient information in initial datasets and promoting transparency and traceability in dataset annotation practices within the ICASSP community.

1 Introduction

Audio machine learning is important because many people are integrating their lives through using audio services. More than 515 million people are using Spotify within their daily life [1]. Voice assistants are becoming popular [27] and mobile phones and computers are already using noise pollution reductions via the apps such as Zoom[2] and Google Meet[3].

For those services benefiting from machine learning to work properly, they are highly dependent on the dataset provided [29]. If the dataset contains substantial amount of errors, the algorithm would be doomed to produce incorrect results. Thus, the dataset provided has a significant impact on the success of the outcomes. Dataset quality can be divided into different categories one of which is related to the annotation practices. According to Geiger et al.'s research, it has been shown that annotations are poorly reported in one field, and the poor quality in annotations leads to misleading results [67]. The absence of information regarding the annotation methods of the dataset presents a significant issue as it hinders the reader's ability to promptly assess its quality. In addition, Gabelica et al. report in an extensive search that 93% of the authors who promised in the paper to provide their datasets upon request declined or did not respond to those requests [20].

According to our background research, the work examining the reporting practices and availability of audio data annotations has thus far not been conducted. In addition, a research analysing the timeline of datasets and datasets' documentation on their annotation practices has not found. Hence, the research question is: to what degree is audio machine learning data transparent and clearly reported in current, highly cited work and how the reporting practices in terms of their annotations differ in dataset's timeline?

Geiger et al.'s paper on annotation practices [68] serves as an inspiration for our research. In this current paper, we present the annotation practices within the context of a highly cited audio signal processing and machine learning venue, namely the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), for the years 2021 and 2022. The objective is to further evaluate the validity of Geiger et al.'s findings in the field of audio signal processing and machine learning on 5 top-cited papers, displaying the changes of reporting techniques in different resources and providing our insights on the metadata search strategies within different types of resources in tracing to the initial original datasets. ICASSP venue is chosen due to the fact that its popularity on audio signal processing and machine learning. In addition, its main language is English, therefore the papers and the related materials published would be in the English language which the authors of this paper are fluent in. The chosen year range is intended to encompass the venue's most recent works while the citation count is utilized as a criterion for selecting the most impactful studies. Specifically, the first stages of this current paper is inspired by a systematic review focusing on the chosen venue during the aforementioned time frame. The later stages are more related to metadata search in different types of resources until sufficient information is collected with regards to the initial original datasets. These metasearches happen through both within the resources(reading the sections and looking through figures, tables references) and outside of the sources through querying the terms, technologies and datasets utilized.

A systematic review analyzes and summarizes data from published studies to draw refined conclusions. In a typical systematic review, the collection of published studies are examined and documented. The search happens in those primarily selected set and it is reported based on the information given within those. The main idea within those is to provide an overview of the subject under investigation. However within this current search, the main objective is to understand the change of annotation practices between the initial creators and the academic users of those datasets. Therefore, it is significant to dive into the search of the initial dataset rather than broadly analyse the papers utilising datasets as an individual body of work representing those datasets. The current methodology this paper represents is to comprehensively identify datasets and the associated technologies that benefit from these datasets. Furthermore, the crucial point of this research is to obtain detailed information with regards to the

[1] https://www.reuters.com/technology/spotifys-monthly-active-users-rise-above-500-million-beats-estimate-2023-04-25/#: :text=The%20company%20forecast%20monthly%20active, expansion%20into%20podcasts%20and%20audiobooks

[2] https://support.zoom.us/hc/en-us/articles/360046244692-Configuring-professional-audio-settings-for-Zoom-Meetings

[3] https://support.google.com/meet/answer/9919960

primary sources of those datasets.

The paper is organized as follows: Section 2 presents the methodology utilized within the paper. It starts with informing the literature search for the main papers considered, then it continues with the questions that were answered throughout the investigation. It is concluded by diving deep into the dataset search strategy per paper by exhibiting and explaining a flowchart. Section 3 displays the results of the questionnaire in a statistical manner grouped by the relevant resource type. Section 4 is the discussion of the given results. Analysis along with the additional comments on the results are portrayed there. Section 5 presents the limitations. Later, the comments on the reproducibility are given in Section 6 as 'Responsible Research'. All of the process,findings,results and analysis related materials can be found in the supported material's link [65].

2 Methodology

This section describes the extensive investigation process utilized for the metadata mining on audio datasets. The first step of gathering the initial academic papers and answering the relevant survey questions is inspired by systematic review methods, while the extensive search is done through diving deep into the web queries.

The literature selection was through the Scopus[4] website. The search query was filtered to contain only the conference papers from ICASSP in 2021 and 2022. All of the papers were ordered in descending order of their citation numbers and the first 5 of them were chosen for this research. The complete list of the papers selected along with the resources that are connected can be found in Table 1.

The questions that inspired the process are explained in the 'Initial Question Set' subsection while the questions that arose from our process are explained in the 'Additional Questions' subsection. The investigation process as a flowchart is shown and explained in the 'Process' subsection.

Complete list of questions utilized can be found in Appendix A. And the complete list of answers can be found in the GitHub page [5].

2.1 Initial question set

Geiger et al.'s questions are utilized as an inspiration to our research due to their coverage in terms of information on annotation practices. Their questions are revealing on the annotation practice utilized and are good in terms of determining the quality of the annotation along with the dataset utilized. Some important questions are analyzed below.

Did they use human annotations as labels?
There are several annotation practices such as automatic annotation [21], rule-based annotation [13] and transfer learning [44] which does not involve human labelling. This question asks whether the dataset is labelled by humans or a machine. This question is crucial because the researcher must find the information that describes the initial dataset's annotation practices in order to be able to answer it.

Did they use original human annotations?
The 'original' adjective refers to a dataset being unmodified. If the dataset is edited, filtered, extra datapoints are added or combined with a different dataset at any point, it is not considered as a dataset using original human annotations. This question is significant in the sense that it reveals the dataset's timeline while guiding the researcher to search for the initial dataset. The additional dimension of time helps the research to further understand the process of this dataset such as how it started, what modification are done in which stages with which intentions. It is significant to note those findings when deciding on how transparent and how much quality does the dataset utilized represents. Furthermore, when the resource is marked 'no' to this question, it shows that it is not the initial dataset. Thus, the researcher needs to search further.

Who were the annotators?
This question asks the specification on the annotators. The annotators' profession or the level of knowledge is important when the annotated task requires significant amount of knowledge. To exemplify, if the speech corpus contains Japanese speech, the annotators need to be able to understand it perfectly. If the speech corpus contains the audio from a lecture from an international microbiology conference, the annotators need to be familiar with the jargon. Thus, this question reveals whether the annotators have sufficient qualifications to annotate the data.

Were there formal instructions for annotators?
Formal instructions define how the annotation process happens and what are the parts that the annotator should be careful about. It also describes how edge cases are handled and has a section on frequently misunderstood/asked parts. It is a guide to the annotators and this question reveals if the instructors take any type of precaution to keep their annotation process consistent. Having a formal instructions increases the reliability of the annotations.

2.2 Additional Questions

Those questions are created and added to the questionnaire by the author of this paper. They are based on the necessities of deducing the results along with helping to increase the reproducibility of the research. These additional questions are necessary because they cover various types of resources while Geiger et al.'s questions only considers academic papers. Several important additional questions can be found below.

What are the website links representing or explaining the resource?
There might be more than one resource giving information on the annotation practices, with this question, the researcher is able to provide all of the relevant information sources that are looked at. This question is significant to help increasing the reproducibility of the research because it displays the specific resources the current researcher has read.

[4] https://www.scopus.com/home.uri

[5] https://github.com/DogaTascilar/AnnotationPracticesICASSP

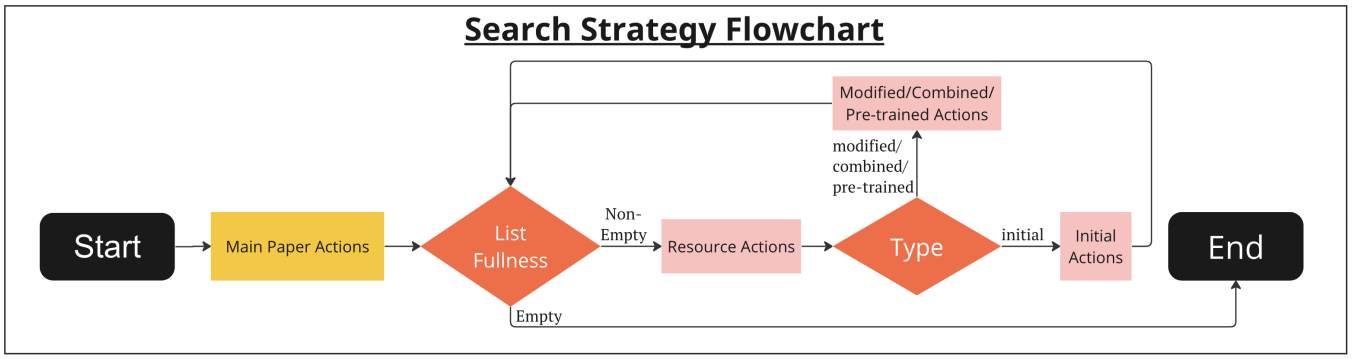

Figure 1: It represents the search strategy followed for one paper.

What is the type of the resource? (initial/modified/combined/model)
The resources are varied and the resource might be presenting an initial dataset, a dataset which is a modified version of another dataset, a dataset combining multiple datasets or a model which is pre-trained by a dataset. These encompass the possible representations a resource within this research. In this question, the researcher is expected to deduce the category the resource belongs to.

Shortly describe the resource.
This open ended question enables the researcher to concisely articulate the representation of the resource. This question helps the researcher to effectively preserve the information about the investigated resource for future reference. summarize the key points and while enabling the researcher to include supplementary information.

(Optional) Do you have any additional comments?
This question is left optional because the researcher is only expected to answer it only if there is anything significant to report in the research. The researcher can express her/his subjective opinion on the resource looked at. This question is also good for the analysis phase.

2.3 Process

The process of gathering the resources is recursive instead of iterative which indicates that based on the connections between the paper selected and the dataset(s), the research process increases exponentially. The 'recursive' indicates that each resource that is reached through the investigation is treated as an initial search point. Thus, the resources found in different depths are investigated the same way. To elaborate, in this search the paper is not the main focus but a starting point. Owing to the fact that the real search starts with the paper, the paper guides the researcher to look for the technologies which are human-labelled datasets or a technology utilising those such as a pre-trained model. Through there, the researcher tracks the timeline of the dataset to observe its creation along with the modifications, and how these might result into possible inconsistencies.

To better explain the structure of the search, Figure 1 represents the process flowchart. Different points in the chart are explained in their relevant subsections.

Main Paper Actions

Within this process point, the main paper is selected from the most cited ICASSP papers. Only this part is inspired by systematic surveys. The paper is read and the researcher answers the survey questions accordingly. The possible dataset containing resources (datasets or pre-trained models/algorithms) are recorded in a list by their names or titles. If the paper refers them explicitly, the references are also recorded.

Checks

This decision point represents the list's emptiness. If the list that was initially created in the above section is empty, the research process for that literature paper ends. Otherwise, it is forwarded to the 'Resource Actions'.

Resource Actions

Within this process point, one resource is selected from the list. The relevant links and search queries are done to gather information with regards to the resource. If the previous resource has already referenced this current resource, that reference has checked. After finding all of the relevant links representing that resource, the resource is forwarded to the 'Type' decision point.

Type

In this decision point, the resource is grouped into two different types: the initial dataset or a resource which represents a modified/combined/pre-trained model. According to its type, it is forwarded to relevant process points.

Modified/Combined/Pre-trained Actions

This process point stands for the resources representing a modified dataset (a version of another dataset which is edited, filtered or added values), a combined dataset (a dataset which combines several datasets into one) or a pre-trained model which benefits from a dataset. If the resource is one of them, it means that they are using at least one external dataset, therefore, the search should continue until the initial dataset(s) is reached. The questionnaire is answered according to this given resource. Then, the dataset(s) utilized by this resource is added to the list. Lastly, this current resource is removed from the list. This point is the recursive part of the investigation. Hence, it is forwarded to the decision point to continue the search.

Initial Actions
This process point describes the steps that needs to be taken for the resources representing the initial original datasets. The survey is answered based on this resource, and if there are parts where the resource does not explain in anywhere, the authors are searched and tried to be contacted through their provided emails and their personal LinkedIn accounts. Their informations provided are added when they answer. In the questionnaire there are some questions which needed initial dataset information, therefore, the answers of those are back-propagated. This point is also the recursive part of the investigation. Thus, it is forwarded to the decision point to continue the search.

3 Results

This section discusses the findings of our research. Initially, it is crucial to note that due to the high branching of the resources, we decided to generate graphs per literature paper to be able to keep track of the connections. An example can be seen in Figure 2. The work presented in this paper utilizes 17 datasets for train and test [24]. The arrow in the graphs represents that the predecessor is utilized by the successor. The ones in with the red frames represent the resources that are included in the results and the blue frame ones give information on how the initial dataset is colleced. All of the connection graphs per paper can be found in Appendix B. All of the unique resources that are looked at are presented in Table 1 along with their references. It is significant to point out that for some resources, it was necessary to look at more than one place because they were not providing sufficient information for our research.

After seeing those different types of resources, we decided to group our findings into 4 different categories which in fact represent the types of papers collected. Thus, those four different categories are: the papers presenting the initial original dataset, the papers presenting the modified dataset, the papers presenting the combined dataset and the papers solely utilizing dataset(s). The distribution of those are represented in the Figure 3 .

It is also significant to note the different types of references to the datasets. We divided into two categories: Explicit and implicit:

- **Explicit**ly stating the dataset within a paper refers making the dataset utilized abundantly clear such as stating it in the abstract or dedicating a section to the dataset(s) utilized. Within that dedicated section, the paper can address the qualities of the dataset such as how many datapoints it contains, who collected it, what does it contain, how the annotation was done, who were the initial collectors etc.
- **Implicit**ly stating the dataset within a paper refers giving the datasets' name under a figure/table or mentioning it in a section scarcely. This complicates the reader's task of locating the dataset since they must thoroughly examine the entire paper and be meticulous in identifying its utilization.

On another note, some research papers integrated a pre-trained model into their own implementation. When this is the case, they do not mention that the model is pre-trained, therefore, the paper benefiting from a dataset is only found out after the model was examined. This is also considered as an implicit referencing.

The distribution of the reference types based on their category can be seen in the Figure 4

3.1 Selected Literature

As shown in Figure 4, most of the selected literature mentions the dataset utilized implicitly with the only exception being Tak's paper [64].

The average number of unique resources looked at per paper is 10 with the maximum of 31 and minimum of 3 5. The maximum is an outlier thus when we remove it from the calculation, the average connections of resources is 6.5 .

Unfortunately, no correlation between the literature paper and the number of necessary resources to look at is observed. Hence according to our findings, there is no means to predict how long the search would take to reach the information of the initial dataset(s) of the selected literature.

3.2 Resources Presenting Original Dataset(s)

In our experiment, this is the most prevalent unique resource category with 19 resources. It is expected because all of the other categories uses them. The only thing that we should point out is that the results are given as the unique datasets and only WSJ0 and LibriVox is seen twice within our research. Given that 80% of the papers were focused on audio speech, the presence of only 2 datasets can confirm that there is not one speech dataset that is used often among the highly cited work. This group of the papers are also the most transparent ones. The authors were highly willing to inform the reader on both the collection and the annotation process with regards to their datasets. They generally mention all the precautions taken while and after the data collections. The only limiting factor in their transparency seen was related to the copyright issues. They sometimes used public content and they did not want to deal with copyright laws.
In addition, the authors were not always reachable. 3 out of 7 contacted authors answered our questions while one of them did not respond us after we present our questions. The ones that answered were very informative with regards to their annotation practices.

3.3 Resources Presenting Modified Dataset(s)

Within our survey, 12 resources belong in this category. These datasets generally explicitly stated their dataset utilized by dedicating a section also shown in Figure 4. Even though they were not as transparent in terms of the annotation practices as the original ones, they are quite transparent on what the dataset contains and when and where the data was collected. They always emphasized heavily on their changes to the data.

Type of Resource	Style of Reporting	Resource
Selected Literature	Explicit	Attention is all you need in speech separation [63]
Selected Literature	Implicit	SA-Net: Shuffle attention for deep convolutional neural networks [74] Recent developments on ESPNeT toolkit boosted by conformer [24] FastPitch: Parallel text-to-speech with pitch prediction [33] End-to-end anti-spoofing with rawnet2 [64]
Initial Dataset	Explicit	WSJ0 [51] Datatang [6] [8] [62] [7] AISHELL-2 [17] CSJ [59] [45] HKUST Mandarin Telephone Speech [19] [49] HKUST Mandarin Telephone Transcript Data [46] LDC Fisher Spanish Speech [15] LDC Fisher Spanish - Transcripts [14] The CALLHOME Spanish Speech [71] The CALLHOME Spanish Transcripts [72] [71] JSUT [61] VoxForge English [40] [69] AISHELL-ASR0009 [9] LibriVox [41] SWITCHBOARD [23] TED Talks [38] FreeSound sound library [66]
Initial Dataset	Implicit	WordNet [42] LibriVox data of Linda Johnson [35] MagnaTune Song Dataset [11]
Modified Dataset	Explicit	AMT (for checking ImageNet) [1] WSJ0-2/3mix [26] Switchboard-1 Release 2 [22] [30] TED-LIUMv2 [4] [58] AMT for evaluating the FastPitch algorithm [1]
Modified Dataset	Implicit	Aurora-4 [50] LibriSpeech ASR corpus [48] AMT for translation (Fisher and CALLHOME) [1] WSJCAM0 [57] MC-WSJ-AV [37] [18] LJSpeech [28] MagnaTagATune dataset [11]
Combined Dataset	Explicit	ImageNet-1k [2] [16] Fisher and CALLHOME [55] [54] [53] TED-LIUM Release 3 [25]
Combined Dataset	Implicit	MS COCO [36] AISHELL-1 [10] 4th CHiME [12] The REVERB [32] [31] ASVspoof2019 [70] [73] [43]
Pre-trained Model	Explicit	ResNet50 [3] Deep Voice 3 [52]
Pre-trained Model	Implicit	Kaldi framwork for WSJ0 [56] Kaldi HKUST recipe [39] Tacotron 2 [60] WaveNet [47] [5] Tag A Tune [34]

Table 1: The table displays all of the resources looked at. For some resources, more than one reference is present because generally, more than one place has looked at per resource.

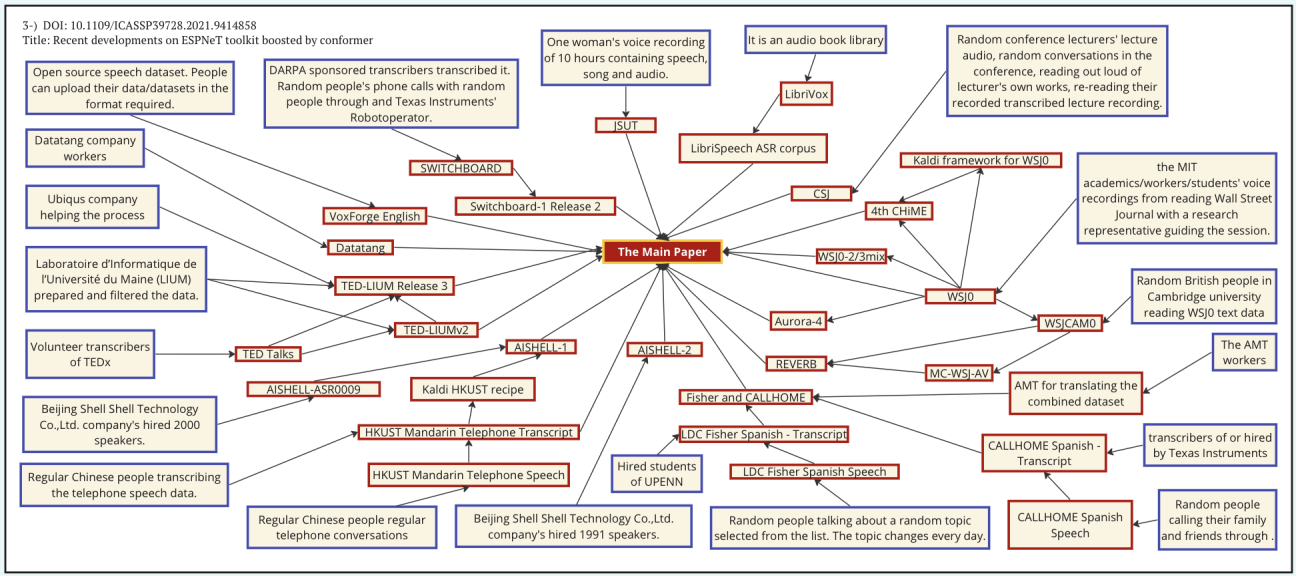

Figure 2: An example graph for one literature paper displaying the dataset connections.

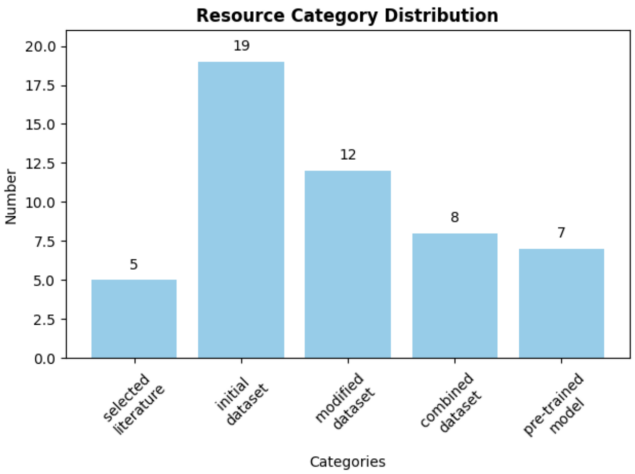

Figure 3: It represents how many resources belong to those categories.

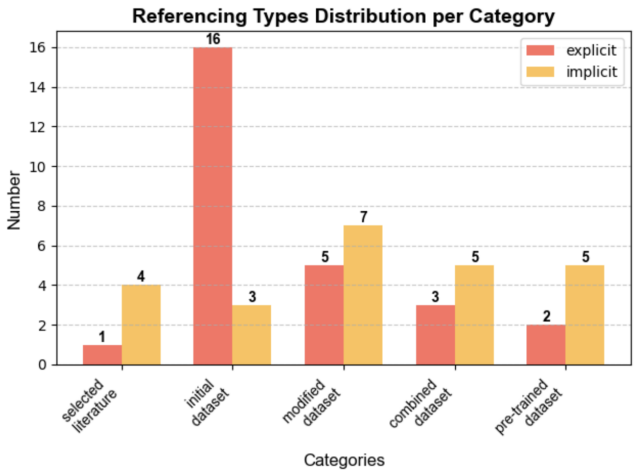

Figure 4: Pink and yellow represent explicit and implicit references to the data/dataset respectively.

3.4 Resources Presenting Combined Dataset(s)

Out of 51 unique resources, 8 of them belong in this category. According to our research, this technique is utilized when the separate datasets did not contain enough datapoints to be used as a training set. 3 out of 8 times ([36], [25], [55]), it was explicitly stated as a reason to do so.

When different datasets utilized, the datasets' sizes and contents generally differ. For example in Guo et al.'s paper, JSUT dataset [61] represents one woman's voice recording of 10 hours while LibriSpeech ASR corpus [48] represents the sound data gathered from an audio book website having more than 100 speakers with long content. When the model is tested and trained, this unequality between the sources should be pointed out since they can create bias [29]. However, none of those combined datasets mentioned any type of normalization to represent the joining datasets equally even though it was necessary in specific cases.

Additionally,we found out some experimental integrity problems associated with this issue on several datasets. Some multi-lingual speech datasets were representing English language far more than they represent any other language [24]. In addition, it seemed to us that the main concern of those combining datasets is to be consistent with the formatting of the joining datasets. This emphasis on formatting was evident in their documentation. They were explicitly mentioning the datasets' contents and the data formatting style but they rarely gave information on the annotation practices.

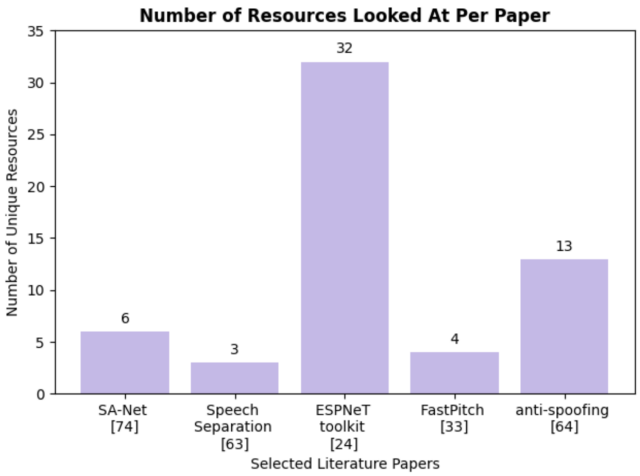

Figure 5: It displays how many unique resources are looked at per paper.

3.5 Pre-Trained Models Utilizing Datasets

This was the most time-consuming category to search for in a referenced paper/resource. Due to our lack of information with each paper subject, we needed to search for all of the methodologies/models/external machine learning algorithms to be sure that they do not contain or use any dataset. At the end of our research, we found out 7 resources looked at belonged to this category. Within the model documentation, only two of them explained the dataset utilized in details. For the others, it was harder to find the relevant information as they rarely mention their datasets.

4 Discussion

This section provides the analysis and explanation of the results, discussing the reasons and factors contributing to the observed outcomes. It is significant to look at few selected papers in order to summarize our findings. The first and the second paper was selected to be explained because of their good practices. They were the top-cited fourth and the fifth paper respectively. Then, the top-cited third paper is explained to display how the lack of metadata search of the datasets can be misleading to the experiments. The most cited first two papers are not explained because they were fairly transparent and did not portray any valuable feature to be mentioned. Finally, our general opinion is presented.

Łańcucki's paper and dataset were different in terms of the dataset content [33]. It was containing only one person's (Linda Johnson's) speech data and one person (Keith Ito) annotated all of the dataset. He was very transparent in the documentation and he answered our additional questions through email.

Tak et al.'s paper referenced her dataset very explicitly [64]. They give information on the content of the dataset and their reasoning behind their choice on this dataset. How it differs from its previous versions and what they would like to achieve with the usage of this dataset. They were very transparent on their usage. However the original dataset owners were not very transparent. We could not reach out to the authors to ask for clarification nor a document stating their annotation practices clearly. They used vague statements on where their data came from and we needed to look for unofficial documentation and found one Reddit post explaining the metadata of the dataset.

Guo et al.'s paper was noteworthy. They used 17 datasets to both test and train [24]. We have realized some inconsistencies with their reportings. They were using several datasets to test and train their systems namely 'Aurora-4', '4th CHiME' and 'WSJ0-2/3mix'. All of those datasets data came from WSJ0 dataset and they were also using WSJ0 as another dataset. They were reporting all of those 4 datasets ('Aurora-4', '4th CHiME', 'WSJ0-2/3mix' and 'WSJ0') as 'different datasets' even though their source of data was exactly the same. This was not the only case. They were using TED-LIUMv2 and TED-LIUM Release 3 as two different datasets even though half of TED-LIUM Release 3 is the same as all of TED-LIUMv2. Furthermore they were using 'AISHELL-1' and 'HKUST Mandarin Telephone Transcript Data'. It is sinificant to note that 'AISHELL-1' is a dataset created by modifying 'HKUST Mandarin Telephone Transcript Data' dataset. It is found out that 12 out of those 17 are unique datasets and the paper did not mention this information. We think that the naming of those datasets might have been misleading because finding this information requires meticulous work of digging deep into dataset metadata. They also misspelled 10 out of 17 dataset utilized which made it even harder for the researcher to find the specific datasets they have benefitted from.

Considering all of the 51 unique resources looked at, the general pattern that we recognized is that the further away from the creation of the dataset, the less transparent the resources become.

When it was considered that every resource has a limit on length, it is sensible that the resources would focus on their own contributions more than the individual components utilized. To exemplify, a dataset which modified another dataset would focus more on the modifications that are done than the unmodified initial dataset's features in their documentation. This is due to the fact that their contribution is their modification, not the component -the initial dataset- utilized.

In addition, the datasets and the pre-trained models were seen as a consistently dependable tool even though the initial dataset creators stated their concerns with regards to their data integrity and correctness in their documentations. Audio machine learning papers not giving the information and the concerns on their utilized dataset's annotation practices implies that the quality of the dataset is often overlooked.

As a matter of fact, when more than one dataset is used, the authors mentions the number of different datasets as a sense of quality. This approach might be relevant in some cases. However, it is also significant to point out the differences between those datasets and why a number of datasets is preferred to be utilized. Because the quality is not determined by the quantity of the datasets but the quality of the data and

its annotation itself.

The language differences, the representative equality, the bias in the speech data are all neglected. These are all very important aspects of machine learning because as we stated at the start, the data is as important as the algorithm and no algorithm without meticulous unbiasing work can generate a correct and a fair result. Thus, we recommend that every researcher to check the dataset utilized thoroughly.

5 Limitations

The general limitations are due to the time limit, the limited computer science knowledge of the author in terms of the topis of the literature papers looked at and the limited foreign language knowledge of the author. These are presented within this section.

In terms of the time limit, the duration of this project was 9 weeks and after realizing some inconsistencies due to the initial datasets, the focus of the research has changed to go deeper into the datasets instead of broader with the papers. Most of the links and references did not forward the researcher to the correct documents, therefore there was a lot of time lost while trying to reach to the documents providing sufficient information. Moreover, some authors of the resources has tried to be contacted but only one of them returned back within a week. The other returned back after 3 weeks while 3 of them did not reply. The data retrieval stage took more time than it was expected.

In terms of the current computer science audio machine learning knowledge of this paper's author, she was a Computer Science and Engineering student therefore, she was not very familiar with the literative work looked at. The author used extensive web queries to understand and search the technologies utilized. There still might have been some parts that were misunderstood, thus misrecorded.

In terms of the language differences, ICASSP was chosen also due to the fact of being an English venue. However, some resources were in Chinese and Japanese which the author is not fluent in. The author used Google Translate for those.

6 Responsible Research

This section describes the validity of the findings in line with the research methodology.

In terms of the reproducibility of the results, two categories can be presented: the reproducibility of the questionnaire and the reprodubility of the results that are deducted from the questionnaire. While filling out the questionnaire, all of the links that have read before answering for the resource are added in the results. Thus, the researcher who would like to repeat the investigation can read the exact document that were read and answer the questionnaire accordingly. In terms of the deductions, all of the calculation metrics are recorded int eh questionnaire. To exemplify, *whether the resource is referencing the dataset explicitly or implicity* is a question that is answered for all of the resources. The categories are also seenable through the additional questions. In addition, the findings and the opinions of the author per paper are recorded so that another researcher can compare hers to the author in a faster way.

Furthermore, the understanding level of the researchers were the undergraduate computer science and engineering student level which might affected the additional information presented in the academic papers.

Within this systemic survey, we examined the most cited 5 papers of ICASSP between the years of 2021 and 2022 to find to what extent the literature and the related datasources are transparent with regards to their annotation practices. We have found out that the initial dataset creators are very transparent while the papers utilizing those as an external dataset or an external pre-trained model are not. This was expected considering that every authors' focus was on their own contributions. Thus if they did not contribute to the annotation, they would not mention. However after finding out some experimental integrity violations, it is highly crucial for the authors to check their datasets especially when they are using more than one. We have looked at the most cited 5 papers to still be able to generalize our findings due to the smaller scope of this paper. In addition, we recommend the future authors to include and check their datasource links and the versions of their utilized datasets. Giving information on the annotation practices is crucial to decide on the quality of the dataset which has a high impact on the machine learning algorithm utilized.

References

[1] Amazon Mechanical Turk — mturk.com. https://www.mturk.com/. [Accessed 25-Jun-2023].

[2] imagenet-1k · Datasets at Hugging Face — huggingface.co. https://huggingface.co/datasets/imagenet-1k. [Accessed 25-Jun-2023].

[3] ResNet-50 convolutional neural network - MATLAB resnet50 - MathWorks Benelux — nl.mathworks.com. https://nl.mathworks.com/help/deeplearning/ref/resnet50.html. [Accessed 25-Jun-2023].

[4] P. Deléglise A. Rousseau and Y. Estève. Tedliumv2. https://openslr.magicdatatech.com/19/, 2014. [Accessed 25-Jun-2023].

[5] Sander Dieleman Aäron van den Oord. WaveNet: A generative model for raw audio — deepmind.com. https://www.deepmind.com/blog/wavenet-a-generative-model-for-raw-audio, 2016. [Accessed 25-Jun-2023].

[6] Ltd Beijing DataTang Technology Co. AI data annotation data customization Datatang — datatang.ai. https://www.datatang.ai/annotation. [Accessed 25-Jun-2023].

[7] Ltd Beijing DataTang Technology Co. aidatatang 200zh. http://openslr.org/62/. [Accessed 25-Jun-2023].

[8] Ltd Beijing DataTang Technology Co. Papers with Code - aidatatang 200zh Dataset paperswithcode.com. https://paperswithcode.com/dataset/aidatatang-200zh, 2010. [Accessed 25-Jun-2023].

[9] Ltd Beijing Shell Shell Technology Co. Open source mandarin speech corpus [aishell-asr0009-os1] training and test data. online, Nov 2018. documentation.

[10] Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*, pages 1–5. IEEE, 2017.

[11] John Buckman. Information: about Magnatune — magnatune.com. http://magnatune.com/info/, 2003. [Accessed 25-Jun-2023].

[12] Kean Chin. The 4th CHiME Speech Separation and Recognition Challenge — spandh.dcs.shef.ac.uk. https://spandh.dcs.shef.ac.uk/chime_challenge/CHiME4/, 2016. [Accessed 25-Jun-2023].

[13] Laura Chiticariu, Rajasekar Krishnamurthy, Yunyao Li, Frederick Reiss, and Shivakumar Vaithyanathan. Domain adaptation of rule-based annotators for named-entity recognition tasks. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1002–1012, 2010.

[14] Ingrid Cartagena Kevin Walker Christopher Cieri David Graff, Shudong Huang. Fisher spanish - transcripts. https://catalog.ldc.upenn.edu/LDC2010T04, 2010. [Accessed 25-Jun-2023].

[15] Ingrid Cartagena Kevin Walker Christopher Cieri David Graff, Shudong Huang. Fisher Spanish Speech - Linguistic Data Consortium — catalog.ldc.upenn.edu. https://catalog.ldc.upenn.edu/LDC2010S01, 2010. [Accessed 25-Jun-2023].

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[17] Jiayu Du, Xingyu Na, Xuechen Liu, and Hui Bu. Aishell-2: Transforming mandarin asr research into industrial scale. *arXiv preprint arXiv:1808.10583*, 2018.

[18] Mike Lincoln Erich Zwyssig. Mc-wsj-av corpus — catalog.ldc.upenn.edu. https://catalog.ldc.upenn.edu/docs/LDC2014S03/README.pdf, 2012. [Accessed 25-Jun-2023].

[19] P Fung, S Huang, and D Graff. Hkust mandarin telephone speech, part 1. *LDC2005S15. Web download*, 2005.

[20] Mirko Gabelica, Ružica Bojčić, and Livia Puljak. Many researchers were not compliant with their published data sharing statement: a mixed-methods study. *Journal of Clinical Epidemiology*, 150:33–41, 2022.

[21] Luit Gazendam, Christian Wartena, Véronique Malaisé, Guus Schreiber, Annemieke De Jong, and Hennie Brugman. Automatic annotation suggestions for audiovisual archives: Evaluation aspects. *Interdisciplinary Science Reviews*, 34(2-3):172–188, 2009.

[22] John J Godfrey and Edward Holliman. Switchboard-1 release 2. *Linguistic Data Consortium, Philadelphia*, 926:927, 1997.

[23] John J Godfrey, Edward C Holliman, and Jane McDaniel. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520. IEEE Computer Society, 1992.

[24] Pengcheng Guo, Florian Boyer, Xuankai Chang, Tomoki Hayashi, Yosuke Higuchi, Hirofumi Inaguma, Naoyuki Kamo, Chenda Li, Daniel Garcia-Romero, Jiatong Shi, et al. Recent developments on espnet toolkit boosted by conformer. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5874–5878. IEEE, 2021.

[25] François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve. Ted-lium 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *Speech and Computer: 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18–22, 2018, Proceedings 20*, pages 198–208. Springer, 2018.

[26] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35. IEEE, 2016.

[27] Justine Humphry and Chris Chesher. Preparing for smart voice assistants: Cultural histories and media innovations. *New media & society*, 23(7):1971–1988, 2021.

[28] Keith Ito and Linda Johnson. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/, 2017.

[29] Abhinav Jain, Hima Patel, Lokesh Nagalapatti, Nitin Gupta, Sameep Mehta, Shanmukha Guttula, Shashank Mujumdar, Shazia Afzal, Ruhi Sharma Mittal, and Vitobha Munigala. Overview and importance of data quality for machine learning tasks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3561–3562, 2020.

[30] Edward Holliman John J. Godfrey. Switchboard-1 Release 2 - Linguistic Data Consortium — catalog.ldc.upenn.edu. https://catalog.ldc.upenn.edu/LDC97S62, 1997. [Accessed 25-Jun-2023].

[31] Takuya Yoshioka Tomohiro Nakatani Keisuke Kinoshita, Marc Delcroix. The REVERB challenge - Evaluating de-reverberation and ASR techniques in reverberant environments — reverb2014.dereverberation.com. https://reverb2014.dereverberation.com, 2014. [Accessed 25-Jun-2023].

[32] Keisuke Kinoshita, Marc Delcroix, Sharon Gannot, Emanu¨el A P. Habets, Reinhold Haeb-Umbach, Walter Kellermann, Volker Leutnant, Roland Maas, Tomohiro

Nakatani, Bhiksha Raj, et al. A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research. *EURASIP Journal on Advances in Signal Processing*, 2016:1–19, 2016.

[33] Adrian Łańcucki. Fastpitch: Parallel text-to-speech with pitch prediction. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6588–6592. IEEE, 2021.

[34] Edith Law and Luis Von Ahn. Input-agreement: a new mechanism for collecting data using human computation games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1197–1206, 2009.

[35] Linda Johnsonn LibriVox. Linda johnsonn LibriVox — librivox.org. https://librivox.org/sections/readers/11049, 2016. [Accessed 25-Jun-2023].

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[37] Mike Lincoln, Iain McCowan, Jithendra Vepa, and Hari Krishna Maganti. The multi-channel wall street journal audio visual corpus (mc-wsj-av): Specification and initial experiments. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*, pages 357–362. IEEE, 2005.

[38] TED Conferences LLC. translate transcribe — ted.com. https://www.ted.com/participate/translate/transcribe, 2022. [Accessed 25-Jun-2023].

[39] ma08. Kaldi hkust. https://github.com/kaldi-asr/kaldi/tree/master/egs/hkust, 2016. [Accessed 25-Jun-2023].

[40] Ken MacLean. Free Speech... Recognition (Linux, Windows and Mac) - voxforge.org. https://www.voxforge.org/. [Accessed 25-Jun-2023].

[41] Hugh McGuire. LibriVox — free public domain audiobooks — librivox.org. https://librivox.org/, 2005. [Accessed 25-Jun-2023].

[42] George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244, 1990.

[43] Andreas Nautsch, Xin Wang, Nicholas Evans, Tomi H Kinnunen, Ville Vestman, Massimiliano Todisco, Héctor Delgado, Md Sahidullah, Junichi Yamagishi, and Kong Aik Lee. Asvspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(2):252–265, 2021.

[44] Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? *Advances in neural information processing systems*, 33:512–523, 2020.

[45] National Institute of Japanese Language (). *How to build a corpus of spoken Japanese ()*. National Institute for Japanese Language (), Mar 2006.

[46] Hong Kong University of Science and Technology. Linguistic Data Consortium Catalog. https://catalog.ldc.upenn.edu/docs/LDC2005S15/, 2007. [Accessed 25-Jun-2023].

[47] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[48] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

[49] David Graff Pascale Fung, Shudong Huang. Mandarin chinese conversational telephone speech transcripts, part 1. https://catalog.ldc.upenn.edu/docs/LDC2005S15/, 2005. [Accessed 25-Jun-2023].

[50] David Pearce and J Picone. Aurora working group: Dsr front end lvcsr evaluation au/384/02. *Inst. for Signal & Inform. Process., Mississippi State Univ., Tech. Rep*, 2002.

[51] Michael Phillips, James Glass, Joseph Polifroni, and Victor Zue. Collection and analyses of wsj-csr data at mit. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

[52] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep voice 3: 2000-speaker neural text-to-speech. *proc. ICLR*, pages 214–217, 2018.

[53] Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur. Linguistic Data Consortium Catalog — catalog.ldc.upenn.edu. https://catalog.ldc.upenn.edu/docs/LDC2014T23/. [Accessed 25-Jun-2023].

[54] Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur. Improved speech-to-text translation with the Fisher and Callhome Spanish–English speech translation corpus. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany, December 2013.

[55] Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur. Fisher and callhome spanish–english speech translation. *LDC2014T23. Web Download. Philadelphia: Linguistic Data Consortium*, 2014.

[56] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The kaldi speech recognition toolkit. In *IEEE 2011*

workshop on automatic speech recognition and understanding, number CONF. IEEE Signal Processing Society, 2011.

[57] Tony Robinson, Jeroen Fransen, David Pye, Jonathan Foote, and Steve Renals. Wsjcamo: a british english speech corpus for large vocabulary continuous speech recognition. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 81–84. IEEE, 1995.

[58] Anthony Rousseau, Paul Deléglise, Yannick Esteve, et al. Enhancing the ted-lium corpus with selected data for language modeling and more ted talks. In *LREC*, pages 3935–3939, 2014.

[59] Hanae Koiso Kenya Nishikawa Yoko Mabuchi Seiju Sugito, Kikuo Maekawa. Documents - Corpus of Spontaneous Japanese — clrd.ninjal.ac.jp. https://clrd.ninjal.ac.jp/csj/en/document.html, 2006. [Accessed 25-Jun-2023].

[60] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4779–4783. IEEE, 2018.

[61] Ryosuke Sonobe, Shinnosuke Takamichi, and Hiroshi Saruwatari. Jsut corpus: free large-scale japanese speech corpus for end-to-end speech synthesis. *arXiv preprint arXiv:1711.00354*, 2017.

[62] SpeechOcean. The largest open source Chinese corpus and another five speech recognition datasets — en.speechocean.com. https://en.speechocean.com/Cy/778.html. [Accessed 25-Jun-2023].

[63] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. Attention is all you need in speech separation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 21–25. IEEE, 2021.

[64] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. End-to-end anti-spoofing with rawnet2. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6369–6373. IEEE, 2021.

[65] Doga Tascilar. Dogatascilar/annotationpracticesicassp: This is the repository where i shared the relevant data for my bachelor thesis in tu delft., Jun 2023.

[66] the Music Technology Group of Universitat Pompeu Fabra. Freesound — freesound.org. https://freesound.org/, 2005. [Accessed 25-Jun-2023].

[67] Aditya Thyagarajan, Elías Snorrason, Curtis Northcutt, and Jonas Mueller. Identifying incorrect annotations in multi-label classification data. *arXiv preprint arXiv:2211.13895*, 2022.

[68] Bertie Vidgen and Leon Derczynski. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300, 2020.

[69] VoxForge. Papers with Code - VoxForge Dataset. https://paperswithcode.com/dataset/voxforge, 2019. [Accessed 25-Jun-2023].

[70] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, et al. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 64:101114, 2020.

[71] Barbara Wheatley. CALLHOME Spanish Transcripts - Linguistic Data Consortium — catalog.ldc.upenn.edu. https://catalog.ldc.upenn.edu/docs/LDC96T17/, 1997. [Accessed 25-Jun-2023].

[72] Barbara Wheatley. CALLHOME Spanish Transcripts - Linguistic Data Consortium — catalog.ldc.upenn.edu. https://catalog.ldc.upenn.edu/LDC96T17, 1997. [Accessed 25-Jun-2023].

[73] Junichi Yamagishi. — ASVspoof 2019 — asvspoof.org. https://www.asvspoof.org/index2019.html, 2019. [Accessed 25-Jun-2023].

[74] Qing-Long Zhang and Yu-Bin Yang. Sa-net: Shuffle attention for deep convolutional neural networks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2235–2239. IEEE, 2021.

Appendix

A Questions

A.1 Questions that are Strongly Inspired by Geiger et al.'s Research

These are the questions representing the highlights of Geiger et al.'s research [68]. Their investigation is the primary source of inspiration of those questions and they reflect identical notions.

1. Is the work an original task?
2. Does the work use human annotations as labels?
3. Does the work use original human annotations?
4. Does the work use external human annotations?
5. Who were the annotators?
6. Is the number of annotators specified?
7. Is the number of annotators they would need estimated beforehand?
8. Were there formal instructions for the annotators?
9. Was there training for the annotators?
10. Was there prescreening on the crowdwork platforms?
11. Was there multiple annotator overlap?

12. Is an inter-annotator agreement reported?
13. Is there any other metric of label quality utilized?
14. did they link to the dataset?

A.2 Additional Questions

Those questions are created due their significance in the analysis phase. Their answers were found valuable, therefore, it is decided to include those to the questionnaire as well.

1. What are the website links representing or explaining the resource?
2. Shortly describe the resource.
3. What is the type of the resource? (Initial/modified/combined/pre-trained model)
4. How did the resource state the dataset utilized? (Imlicit or Explicit)
5. (Optional) Do you have any additional comments?

B Result Graphs for Selected Literature Papers

To be able to keep track of the connections between the resources and the selected literature papers, graphs are created.

B.1 SA-Net: Shuffle attention for deep convolutional neural networks

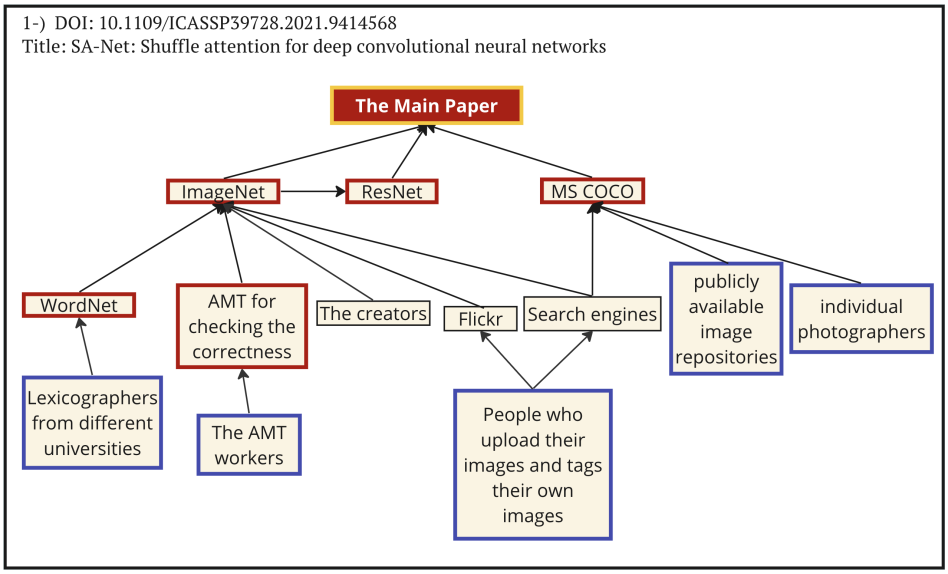

Figure 6: It represents the connections of the main selected literature and the resources. The red framed boxes represent the resources that are included in the research while the blue framed boxes informs the reader on the initial dataset providers/workers. The thin black boxes represent the datasources that cannot be tracked down further due to the lack of information.

B.2 Attention is all you need in speech separation

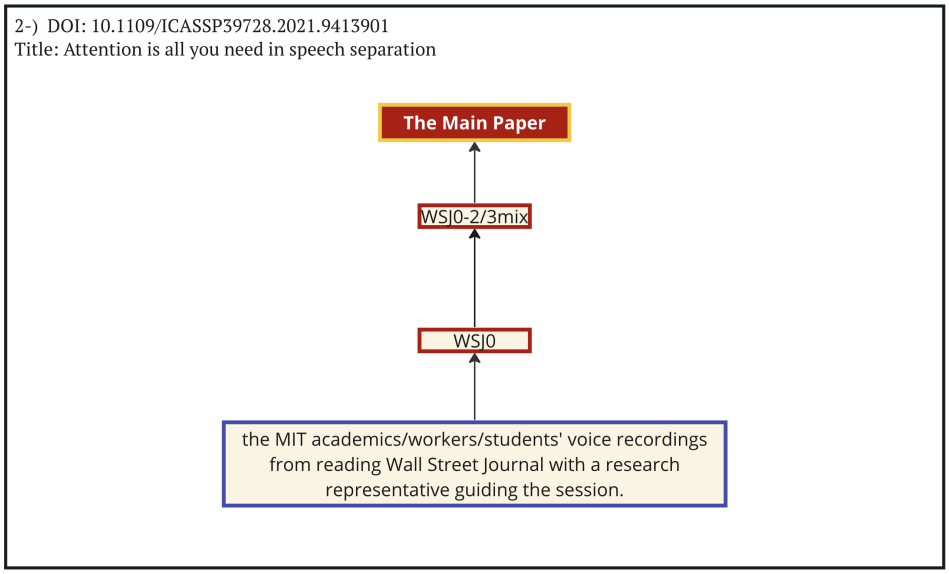

Figure 7: It represents the connections of the main selected literature and the resources. The red framed boxes represent the resources that are included in the research while the blue framed boxes informs the reader on the initial dataset providers/workers. This graph represents that they used one dataset (WSJ0-2/3mix) whose initial data come from WSJ0. The blue box informs the content of the initial dataset.

B.3 Recent developments on ESPNeT toolkit boosted by conformer

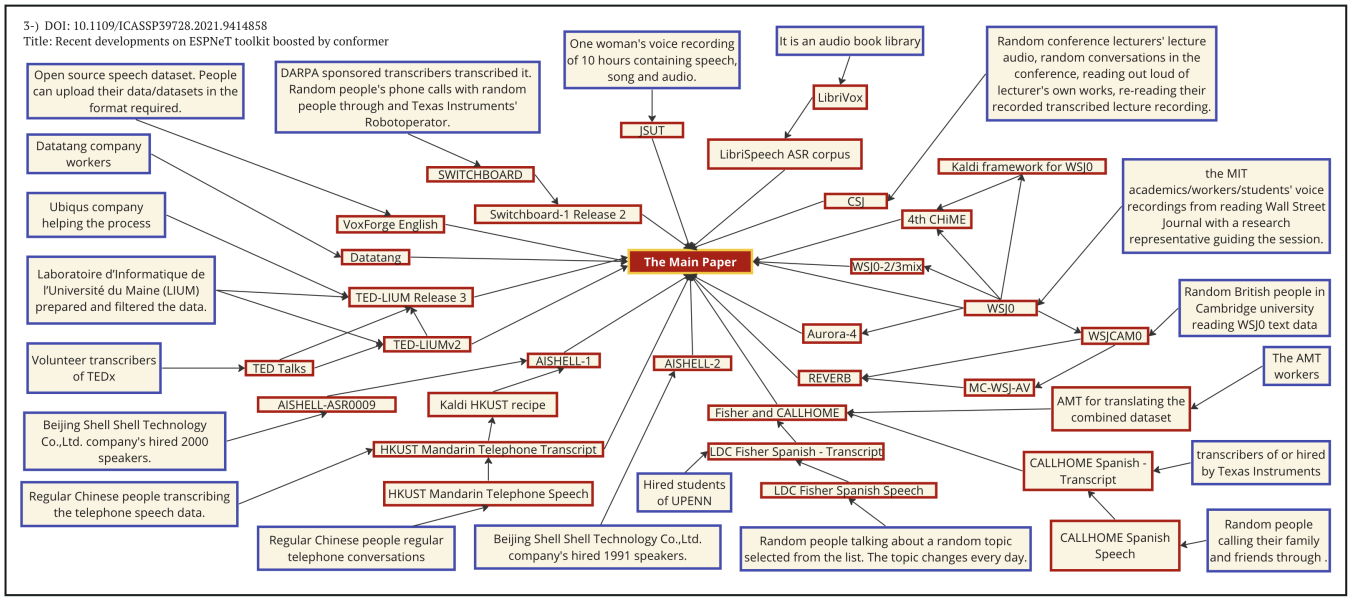

Figure 8: It represents the connections of the main selected literature and the resources. The red framed boxes represent the resources that are included in the research while the blue framed boxes informs the reader on the initial dataset providers/workers. This graph represents that they used 17 datasets. The most significant issue to point out in thise graph is that some of the initial datasets are common among those 17 datasets such as TED-LIUMv2 and WSJ0.

B.4 FastPitch: Parallel text-to-speech with pitch prediction

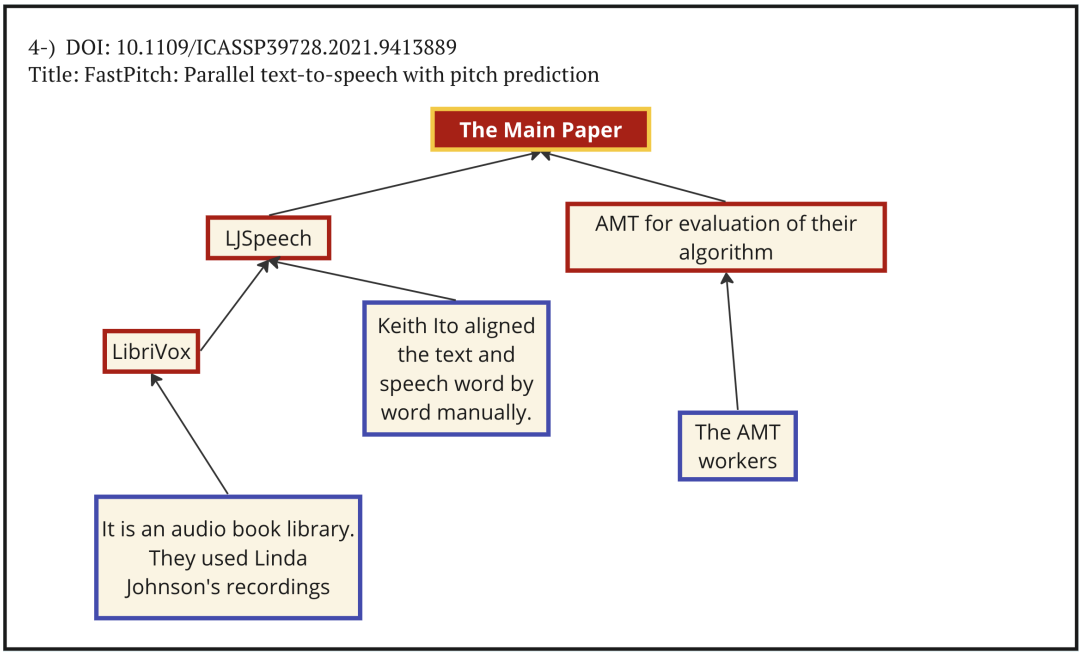

Figure 9: It represents the connections of the main selected literature and the resources. The red framed boxes represent the resources that are included in the research while the blue framed boxes informs the reader on the initial dataset providers/workers. This graph represents that they used one dataset of Linda Johnson's recordings and Keith Ito combined and aligned the text. He used Amazon Mechanical Turk to evaluate his algorithm.

B.5 End-to-end anti-spoofing with rawnet2

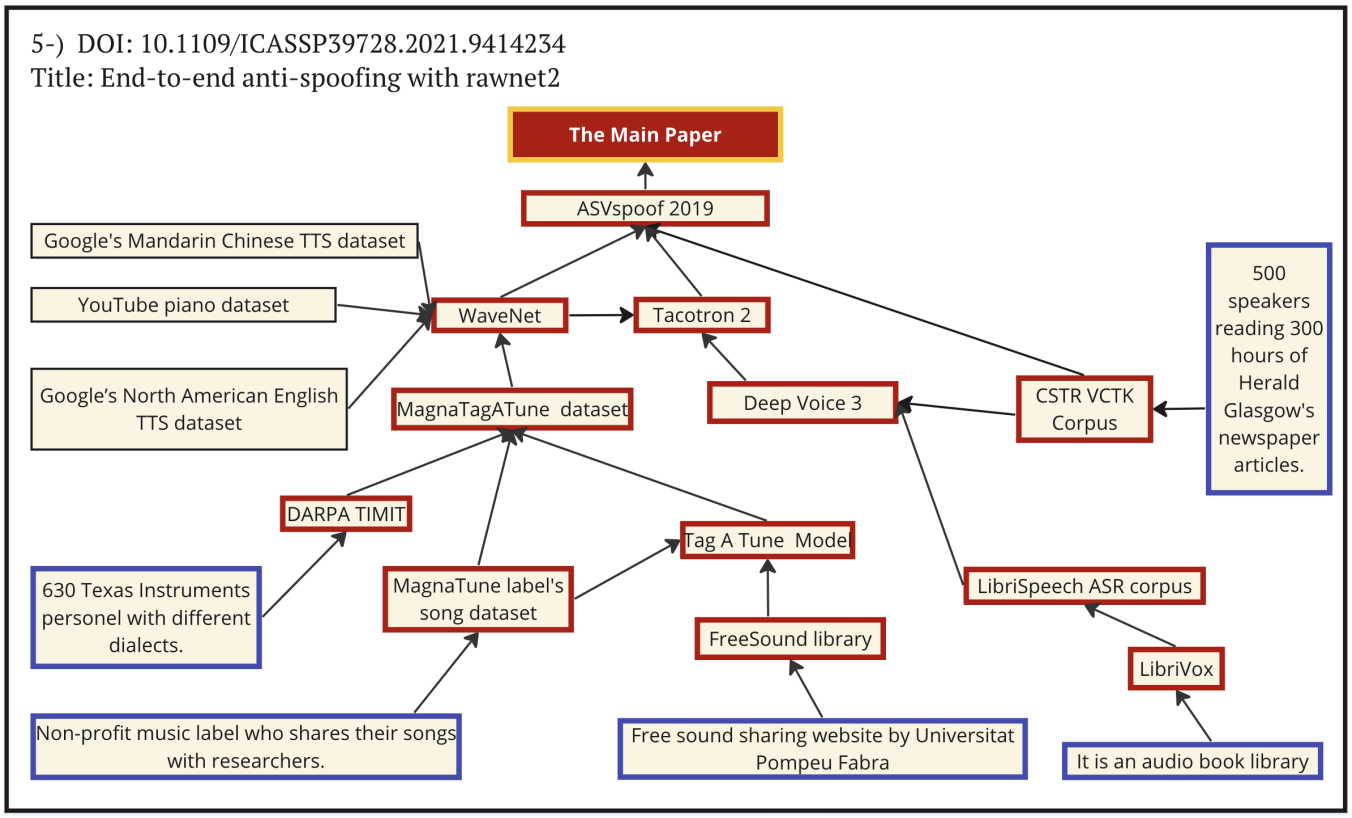

Figure 10: It represents the connections of the main selected literature and the resources. The red framed boxes represent the resources that are included in the research while the blue framed boxes informs the reader on the initial dataset providers/workers. This graph represents that they used one dataset from ASVspoof 2019 challenge which utilized several technologies. The thin black boxes represent the datasets that are not reachable due to the lack of information given in the documentation.