Interspeech 2022
18-22 September 2022, Incheon, Korea

Challenges in Metadata Creation for Massive Naturalistic Team-Based Audio Data

Chelzy Belitz[1], *John H.L. Hansen*[1]

[1]Center for Robust Speech Systems (CRSS), Robust Speech Technologies Lab (RSTL), Erik Jonsson School of Engineering & Computer Science, University of Texas at Dallas, Richardson, Texas, U.S.A.

chelzy.belitz@utdallas.edu, john.hansen@utdallas.edu

Abstract

A broad range of research fields benefit from the information extracted from naturalistic audio data. Speech research typically relies on the availability of human-generated metadata tags to comprise a set of "ground truth" labels for the development of speech processing algorithms. While the manual generation of metadata tags may be feasible on a small scale, unique problems arise when creating speech resources for massive, naturalistic audio data. This paper presents a general discussion on these challenges and highlights suggestions when creating metadata for speech resources that are intended to be useful both in speech research and in other fields. Further, it provides an overview of how the task of creating a speech resource for various communities has been and is continuing to be approached for the massive corpus of audio from the historic NASA Apollo missions, which includes tens of thousands of hours of naturalistic, team-based audio data featuring numerous speakers across multiple points in history.

Index Terms: metadata, naturalistic audio

1. Creating a Speech Resource: an Issue of Scale

State-of-the-art speech processing algorithms have shown high levels of performance on numerous applications however, there remains room for improvement when considering naturalistic data. Furthermore, there is interest in exploration of the higher-level information about how people communicate which can be extracted from their interactions and conversations. These tasks and the push toward the use of machine learning-based algorithms, however, presents two challenges – first, obtaining large-scale, naturalistic audio and second, generating sufficient metadata tags to facilitate research and development of more advanced or more specialized tools.

Traditionally, ground-truth labels are manually generated by human transcribers when creating speech resources. However, various transcription services estimate that the generation of transcripts alone can take up to four times real time, with additional difficulties being presented by the presence of background noise, accents, and technical/field-specific or uncommon vocabulary [1, 2, 3]. Further, additional metadata tags (e.g., speaker tags) are generally of interest in addition to word-level transcripts for many areas of research.

When considering the creation of a speech resource that can be useful to a variety of fields, the issue of scale becomes readily apparent as the amount of available data is increased. Manual transcription generation for an hour of audio can, hypothetically, be considered to take a transcriber around four hours. This is task which can easily be accomplished by a single individual within a day. Transcription of ten hours, however, would roughly scale to taking the same individual a full 40-hour work week to complete. Considering 100 hours might be estimated to take one transcriber roughly 10 work weeks (about two and a third months) or otherwise require multiple transcribers and the additional resources that go into their training and compensation.

While, at this level, use of all manual transcriptions remains feasible albeit time-consuming, consider instead the creation of metadata for a massive audio data set, including over 100,000 hours. Using the same estimation, manual transcription generation would require a single transcriber over 191 years of full-time work, assuming they took no breaks and there was no interest in generating other metadata tags. Alternatively, it would require upfront resources to train and compensate a team of 191 transcribers to complete the task within a year.

However, it remains that such an audio set could be an invaluable resource both in the interest of speech research as well as in numerous other fields. For example, interest in the use of automatically-generated metadata has been suggested for use in psychology, with the acknowledgement that such large amounts of data may contain invaluable information but that manually-generated transcripts become infeasible to create at a certain point due to time and resource limitations.

2. Generating Metadata for Massive Audio Data

2.1. Crowdsourcing

Crowdsourcing is an increasingly popular solution to large-scale and complex metadata generation tasks using platforms such as MTurk, UpFlower, etc. [4]. LanguageARC, in particular, an online platform on which volunteers are able to contribute language data, is being used in the expansion of available manually-generated metadata for the UTD-CRSS Fearless Steps project [5].

This method is advantageous because it allows for multiple people to tag the same portions of data, allowing for more consistent labelling of relatively subjective tags (e.g., emotion, sentiment).

However, there remain limitations in the amount of data which can be feasibly shared. Further, it must be assured that data is structured and uploaded in such a way that it is both easy and time-efficient for participants to complete tasks. This, in turn, brings about the question of how to best balance the amount of time for utterances in such a way that the selected time both contains enough information for a listener to make a decision while ensuring the task doesn't become overly complicated or take too much time so as to avoid deterring volunteers.

10.21437/Interspeech.2022-11243

2.2. Automatic Metadata Generation

The most intuitive and likely only feasible solution when considering a massive data set is creating manual metadata tags for only a relatively small amount of the audio available. This comparatively small subset of tagged data can be used the development of supervised and semi-supervised solutions in order to create automatic labels for the rest of the audio data, thereby creating additional resources that may be of interest in various research fields.

This does, however, bring into question the degree of accuracy an automatically generated tag may need in order to be considered useful. This expectation, however, is largely dependent on the type of metadata and the associated audio and research intent. For example, it is reasonable to anticipate a higher degree of accuracy when using an automatic speech recognition solution for adult speech than one would anticipate for child speech.

3. Metadata Tags

As such, it becomes relevant to discuss the types of metadata tags which may be viewed as most broadly useful to the various communities with interest in the Apollo mission audio when considering the creation of a speech resource. Speech activity, speaker labels, text content, and general speaker traits are considered.

3.1. Speech Activity

Speech activity tags are an important starting place in the creation of a speech resource. In addition to the obvious benefit to knowing where speech occurs, in naturalistic team-based data, silence or speech/silence patterns has been suggested to potentially contain information on both the speaker and, particularly relevant to a team-setting, how speakers are perceived (and, therefore, potentially giving insight into team-dynamics) [6, 7].

3.2. Speaker Tag

Knowledge of who is speaking is relevant in identifying individual speakers and, again, plays an integral role in the overall understanding of communication by allowing for the attribution of specific utterances to particular individuals. In the case of the Apollo audio data discussed in Section 4, it also plays an important role in highlighting the efforts and contributions of the many people who contributed to a historic task.

3.3. Text Content

It is evident that much of the information conveyed in conversation is present in the actual words that are used, making extraction of text content clearly useful in many fields of study. Additionally, generation of text from audio allows for the implementation of natural language processing in understanding conversation.

3.4. Speaker Traits

Still, an individual's voice has the potential to convey a lot more than just the text content of their speech. Speaker trait labels allow for a higher-level examination of speech content that, again, provide invaluable information for academics and researchers. Something like gender, for instance, not only provides an interesting task in speech processing, but gives information on group composition or may hold important historical context (e.g., the comparative lack of female voices present in historic Apollo audio). Other relevant tags may include stress, emotion, sentiment, accent/dialect, among others.

4. Apollo Audio Recovery and Distribution

Beginning in 2017, UTD-CRSS has been working with NASA in an effort to recover and digitize Apollo Program mission audio previously only available on the 30-track analog tapes on which the audio was originally recorded between 1963 and 1972 [8]. This audio is unique not only in its historical significance but, from the perspective of speech research, it provides a uniquely massive set of naturalistic audio featuring large team-based communications.

The number of hours of naturalistic audio included in this data set is already large when accounting for just the total duration of each of the crewed Earth orbiting and lunar missions, as summarized in Table 1.

Table 1: *Summary of Apollo program missions, the mission time rounded to the nearest hour, and the estimated total hours of audio data per mission*

Mission	Mission Hours	Total Hours
Apollo 7	260	15,080
Apollo 8	147	8,526
Apollo 9	241	13,978
Apollo 10	192	11,136
Apollo 11	195	11,310
Apollo 12	245	14,210
Apollo 13	143	8,294
Apollo 14	216	12,528
Apollo 15	295	17,110
Apollo 16	266	15,428
Apollo 17	302	17,516

However, the anticipated total hours of available audio data is multiplied when considering that each tape recorded 30-tracks and further doubled when considering that, historically, as the team communications were recorded, they were recorded using two separate recorders, each recording a different set of audio tracks. For an estimate on the potential total available hours of audio data, consider the sum of the above mission hours doubled then multiplied again by 29 (with one of the 30 channels on each tape being reserved for Inter-range instrumentation group or IRIG timecodes).

Considering this method of estimation, the number of hours of audio data available from these missions can be approximated to be nearly 150k hours, with over 50k hours from 8 different missions already having been digitized. Two particularly well-known missions, Apollo 11 and Apollo 13, have already been preserved and digitized in their entirety with the former having already, in part, been distributed to various communities interested in the available data.

4.1. Digitization

Previously, the analog tapes containing the historic recordings could only play the tapes one channel at a time using one of NASA's Soundscriber playback systems. Early CRSS efforts included the redesign and implementation of a 30-channel read head to facilitate the simultaneous digitization of all channels,

thereby drastically reducing the total amount of time necessary to create digital copies of the recordings [9]. The resulting read head is pictured in Fig. 1.

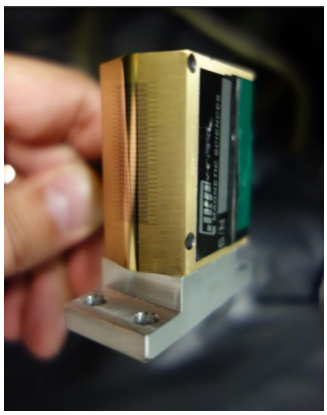

Figure 1: *The UTD-designed read-head for simultaneous 30-track digitization.*

A photo of the original Soundscriber playback system located in NASA's Johnson Space Center used in digitizing the tapes is shown in Fig. 2.

Figure 2: *The Soundscriber playback system located in NASA's Johnson Space Center. The fully installed UT-Dallas Patch Panel, bearing the UTD logo, can be seen in the bottom left corner.*

Fig. 3 shows an overview of all hardware used. A PreSonus (RM32-AI) multi-track recorder is used for this process, connected to the Soundscriber playback system via a patch-panel installed by UTD-CRSS.

In addition to retaining a copy of the resulting digitized audio without any pre-processing, copies of the audio are created following the data preparation pipeline shown in Fig. 4 for more convenient distribution.

Cuts are made where the audio picked up the spikes caused by physically stopping and starting the tape on the playback system. For better audibility, peak normalization is then applied. The audio is then downsampled from its initial 44.2 kHz sample rate to 8 kHz. From there, the audio is split into 30-minute streams, with each file being named following the naming convention shown in Fig. 5.

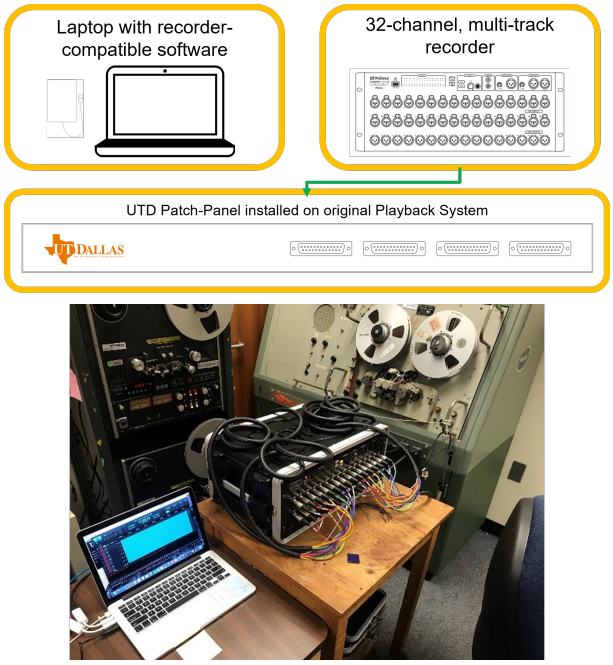

Figure 3: *Above: An overview setup used in digitizing Below: A picture of the setup in use at Johnson Space Center*

4.2. Transcription Effort

The original digitization effort focused primarily on the digitization of all recorded Apollo 11 audio data, given the mission's well-known historical significance. This mission alone included over 10k hours of audio data. Because the manual creation of various metadata for this amount of data is not practically possible, transcribers focused instead on manually tagging metadata for just 100-hours of the total available data from different important points in the mission and focusing on just five of the 30 total channels.

5. Outreach and Inviting Community Participation

The Apollo mission audio carries information that is relevant not only to the speech community as a massive, naturalistic data set, but also preserves information that is of interest to numerous other academic fields. Through the Fearless Steps Challenge and through meetings with an advisory board representing a wide array of academic interests, UTD-CRSS has made an effort to invite further community participation and study.

5.1. The Fearless Steps Challenges

To date, three phases of the Fearless Steps Initiative led by UTD-CRSS have encouraged algorithm development for the extraction of information from the resulting digitized naturalistic multi-channel audio data, beginning with focus on development of algorithms for single-channel data in Phase I, incorporating more difficult/noisy data and focusing on supervised strategies for Phase II, and moving on to emphasis on testing generalizability between channels in Phase III as well as between missions through the inclusion of Apollo 13 audio samples [10, 11, 12].

Specific challenge tasks have included speech activity de-

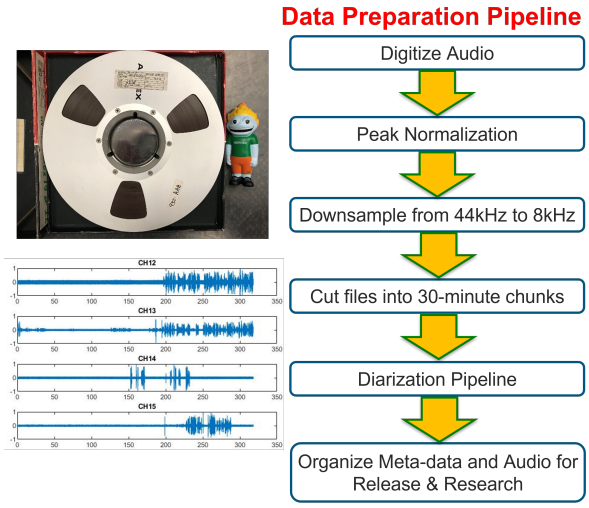

Figure 4: *An overview of the data preparation pipeline*

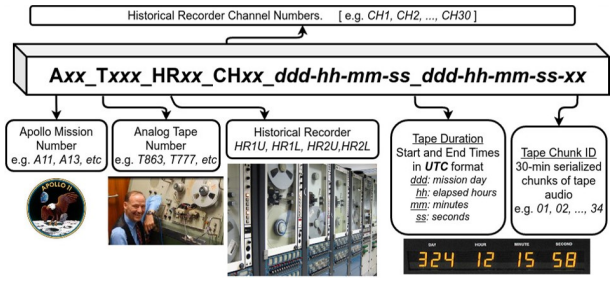

Figure 5: *An overview of the naming convention used in storing and organizing Apollo mission audio for distribution.*

tection, speaker identification, diarization, automatic speech recognition, and conversational analysis, encouraging the development of systems which can target a wide array of potentially valuable metadata tags. Notably, between Phases II and III, best system performance was noted to have improved in the categories of speaker diarization and speech activity detection despite the potential introduction of additional variability in the inclusion of an additional mission [12].

These results further hint at the viability of the implementation of automatic solutions in generating metadata for even massive, variable naturalistic data sets.

5.2. Community Outreach and Feedback

Interest in Apollo mission audio extends far beyond the speech and language community, with the audio archive providing valuable data for various researchers, engineers, scientists, and historians, with interests in fields such as team-based task learning and analysis, audio preservation, and education among others.

Accounting for this, UTD-CRSS has hosted open workshops for parties with an interest in the Apollo mission audio and has maintained contact and met with an advisory board comprised of representatives from diverse fields.

The represented fields are categorized as the following:

- Speech & Language Technology
- Audio Preservation & Archiving
- History & Education
- Speech Science & Psychology (with an emphasis on interest in human/group engagement)

6. Beyond Apollo 11

Previous efforts focusing on Apollo 11 lead by UTD-CRSS have yielded thousands of hours of automatically generated metadata tags, with over 19,000 hours of audio having been put through the diarization pipeline [13]. This has already enabled many hours this massive amount of data to be released to the public – both for use by the speech and language community as well as the many others who share an interest in this historic audio and the wide array of information it contains.

Moving to the inclusion of missions beyond Apollo 11, it is clear that the abundance of data necessitates the use of system to automatically generate metadata tags in order to best create a speech resource with wide-reaching utility in a broad number of fields. Effort has already been put forth in encouraging and facilitating the development of systems which can generalize between channels as well as between missions.

Further efforts for inclusion of the remaining Apollo mission audio are being made with the early labelling of metadata tags for comparatively small samples of audio. Presently, data is being organized such that an hour of time from six selected channels can be uploaded for crowd-sourced labelling. The selected channels include:

- Missions Operations Control Room (MOCR)
- Flight Director (FD)
- Network Controller (NETWORK)
- Guidance, Navigation, and Control (GNC)
- Electrical, Environmental and Communication systems (EECOM)
- Public Affairs Officer (PAO)

Of the six channels selected for manual transcription, Public Affairs Officer (PAO) is a new consideration following feedback from community members suggesting concern that other channels may focus too heavily on application-specific vocabulary and communication. This channel, containing the recordings of the channel in charge of communicating with media representatives, includes more common, conversational vernacular.

The creation of metadata tags for just small samples of the remaining missions will enable the testing of systems across missions to see how they perform. Additional hours of data from the missions are intended to be added as progress is made in order to aid in the development of systems with increasingly improved generalizability which are anticipated to generate more accurate tags for the massive amounts of data which cannot be feasibly transcribed manually.

7. References

[1] D. Chazen. (2019, Jul.) How long it really takes to transcribe (accurate) audio. [Online]. Available: https://verbit.ai/how-long-it-really-takes-to-transcribe-accurate-audio

[2] How long does it take to transcribe one hour of audio or video? [Online]. Available: https://www.rev.com/blog/resources/how-long-does-it-take-to-transcribe-audio-video

[3] Transcription time per audio hour: How long does transcribing really take? [Online]. Available: https://www.opaltranscriptionservices.com/transcription-time-per-audio-hour/

[4] H. Garcia-Molina, M. Joglekar, A. Marcus, A. Parameswaran, and V. Verroios, "Challenges in data crowdsourcing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 4, pp. 901–911, 2016.

[5] J. Fiumara, C. Cieri, J. Wright, and M. Y. Liberman, "Languagearc: Developing language resources through citizen linguistics," in *CLLRD*, 2020.

[6] D. Duez, "Perception of silent pauses in continuous speech," *Language and speech*, vol. 28, no. 4, pp. 377–389, 1985.

[7] D. D. MacGregor L.J, Corley M, "Listening to the sound of silence: disfluent silent pauses in speech have consequences for listeners." *Neuropsychologia*, vol. 48, no. 14, pp. 3982–3992, Dec. 2010.

[8] D. Williams. (2013, Sep.) The apollo program. [Online]. Available: https://nssdc.gsfc.nasa.gov/planetary/lunar/apollo.html

[9] O. of Media Relations. (2017, Dec.) Researchers launch moon mission audio site. [Online]. Available: https://news.utdallas.edu/science-technology/researchers-launch-moon-mission-audio-site/

[10] J. H. Hansen, A. Joglekar, M. C. Shekhar, V. Kothapally, C. Yu, L. Kaushik, and A. Sangwan, "The 2019 Inaugural Fearless Steps Challenge: A Giant Leap for Naturalistic Audio," in *Proc. Interspeech 2019*, 2019, pp. 1851–1855.

[11] A. Joglekar, J. H. L. Hansen, M. C. Shekar, and A. Sangwan, "Fearless steps challenge (fs-2): Supervised learning with massive naturalistic apollo data," *ISCA Interspeech 2020*, 2020.

[12] A. Joglekar, S. O. Sadjadi, M. Chandra-Shekar, C. Cieri, and J. H. Hansen, "Fearless steps challenge phase-3 (fsc p3): Advancing slt for unseen channel and mission data across nasa apollo audio," *ISCA INTERSPEECH-2021*, 2021.

[13] J. H. L. Hansen, A. Sangwan, A. Joglekar, A. E. Bulut, L. Kaushik, and C. Yu, "Fearless steps: Apollo-11 corpus advancements for speech technologies from earth to the moon," in *INTERSPEECH*, 2018.