Development of a Speech Quality Database Under Uncontrolled Conditions

Alessandro Ragano1,2,3,4, Emmanouil Benetos3,4, Andrew Hines1,2

1School of CS, University College Dublin, Ireland 2 Insight Centre for Data Analytics, Ireland 3School of EECS, Queen Mary University of London, UK 4The Alan Turing Institute, UK

alessandro.ragano@ucdconnect.ie, emmanouil.benetos@qmul.ac.uk, andrew.hines@ucd.ie

Abstract

Objective audio quality assessment is preferred to avoid time-consuming and costly listening tests. The development of objective quality metrics depends on the availability of datasets appropriate to the application under study. Currently, a suitable human-annotated dataset for developing quality metrics in the audio archives is missing. Given the online availability of archival recordings, we propose to develop a real-world audio quality dataset. We present a methodology used to curate a speech quality database using the archive recordings from the Apollo Space Program. The proposed procedure is based on two steps: a pilot listening test and an exploratory data analysis. The pilot listening test shows that we can extract audio clips through the control of speech-to-text performance metrics to prevent data repetition. Through unsupervised exploratory data analysis, we explore the characteristics of the degradations. We classify distinct degradations and we study spectral, intensity, tonality and overall quality properties of the data through clustering techniques. These results provide the necessary foundation to support the subsequent development of large-scale crowdsourced datasets for audio quality.

Index Terms: speech quality, speech intelligibility, Apollo space program, sound archives, dataset

1. Introduction

The audio quality of historical audio archives is regularly evaluated with inappropriate objective quality metrics or with individual judgements. Heterogeneity of large collections [1], lack of resources [2], and usage of inappropriate technology [3] are the main barriers to guarantee a careful quality assessment of audio archives which may cause the loss of cultural heritage [3]. Audio archives have been under investigation for different tasks such as broadband noise detection [4], impulse disturbance detection [5], digital restoration [6], impairment recognition [7] and preservation [2, 1]. However, no work to date has been done in terms of automatic audio quality assessment and control. Recent advances relate to the adaptation of the Quality of Experience (QoE) framework to evaluate perceived audio quality in audio archives [3]. However, in the current state of the art, a suitable dataset for assessing audio quality in historical audio archives is missing.

In this paper, we describe a methodology representing the initial phase needed to create a real-world speech dataset for predicting quality in audio archives. We use the archive recordings from the Apollo Space Program that constitutes one of mankind’s greatest achievements. We show how to conduct the extraction of meaningful audio data from a large collection full of silence, almost undetectable speech, and variable signal-to-noise ratio (SNR) [8]. When creating an artificial audio quality dataset, different algorithms are used to manipulate clean signals to obtain audio stimuli that will be rated by test participants [9]. However, audio archive applications cover broad acoustic scenarios [3] that would be difficult to simulate in a controlled environment. A review of the literature found no large real-world speech audio dataset with quality labels. As a consequence, there is no established methodology that we could apply. The proposed procedure is intended to be used as a general methodology to collect real-world data in similar uncontrolled conditions. Specifically, we aim at using this procedure to collect a “large enough” dataset that will allow the exploration of deep learning methods.

The Apollo audio dataset used in this work called the FEARLESS STEPS corpus has been recently curated for the challenge with the same name [10]. The FEARLESS STEPS corpus allows the exploration of different speech processing tasks but it is not labelled for speech quality assessment. However, we did not use the challenge data for the following reasons: 1) the sampling rate of the challenge data is at 8 kHz while the audio archives are stored at 48 kHz to preserve the fidelity of historical material [3]; 2) we want to explore each type of context available from the Apollo corpus (e.g., onboard, commentary, technical-air-to-ground etc.) to balance our data collection and this information was not annotated for the FEARLESS STEPS corpus; 3) we want to use data also from other Apollo missions to increase the speaker variability.

In this paper, we first describe the Apollo audio archive. We then describe the procedure and results of a pilot listening test which provided useful insights to prevent data repetitions. The data used for the pilot listening test is available online with quality ratings included 1. Next, we conduct an unsupervised exploration aimed at classifying different degradations in the data under study. Results from both steps are the foundation to create a human-annotated large-scale speech quality dataset and are necessary in order to minimise the introduction of biases [11] that could result in data mislabeling.

2. The Apollo Audio Archive Description

The Apollo Space Program is documented with pictures, telemetry data, conversation transcripts, video and audio recordings. Audio recordings capture interactions and conversations between astronauts, crew members and backroom staff at the NASA Mission Control Center (MCC). Some transcripts and audio recordings are available online [12] and they can be divided into different categories: onboard, commentary, technical-air-to-ground, MCC recordings, before and post-mission recordings. Onboard recordings include all the conversations between the astronauts on the two spacecrafts, the lunar module (LM) and the Command Service Module (CSM). Onboard recordings are mainly characterised by very low-quality audio due to the harsh acoustic conditions of the spacecraft (e.g., engine-like noise).

1https://doi.org/10.5281/zenodo.3969507
Commentary and technical air-to-ground recordings include conversations between astronauts and the capsule communicator (CAPCOM), the only person who was communicating with the astronauts. These two datasets are the same but the commentary has public affairs officers (PAOs) comments overdubbed. These two datasets represent a broad acoustic scenario as they are affected by the usage of different voice channel implementations that was changing according to the mission status [13].

MCC audio includes conversations between the staff members who were located at the MCC during the mission e.g., communications between flight controllers and backroom specialists [14]. Before and post-mission recordings include audio unrelated to the actual missions such as interviews and press conferences.

3. Pilot Listening Test

The massive heterogeneity of the Apollo corpus, constituted by a large number of acoustic conditions, extended periods of almost undetectable speech and long periods of non-speech activity, makes the random extraction of audio clips infeasible. As our ground truth target labels are quality ratings, we conducted a pilot listening test to answer two questions: 1) How can we avoid an excessive amount of repetitive audio clips that have almost imperceptible quality differences? 2) How are existing objective quality metrics correlated with subjective ratings? The first question is crucial to avoid a narrow or skewed quality distribution in the final large-scale dataset which might cause a crucial fault which is data repetition. This normally would happen if we randomly select audio clips, given the natural characteristics of archive recordings. The second question applies the current state of the art metrics to the Apollo data to validate the hypothesis that more appropriate quality metrics are needed to evaluate speech quality for this dataset.

The experiment is as follows: we choose intelligibility as the quality factor to be evaluated and we assess the correlation between subjective ratings and objective metrics. We presume that intelligibility has higher relevance than other quality factors in this application, given the importance of space communications and the historical significance of the Apollo missions. 17 participants aged 25 to 33 years with mixed gender, 8 males and 9 females, took part in the pilot test. Six participants are English native speakers and 11 participants are fluent non-native English speakers who were located at the MCC during the mission e.g., communications between flight controllers and backroom specialists who were located at the MCC during the mission e.g., communications between flight controllers and backroom specialists [14]. Before and post-mission recordings include audio unrelated to the actual missions such as interviews and press conferences.

The experiment is as follows: we choose intelligibility as the quality factor to be evaluated and we assess the correlation between subjective ratings and objective metrics. We presume that intelligibility has higher relevance than other quality factors in this application, given the importance of space communications and the historical significance of the Apollo missions. 17 participants aged 25 to 33 years with mixed gender, 8 males and 9 females, took part in the pilot test. Six participants are English native speakers and 11 participants are fluent non-native English speakers. They self-declared no hearing disorders and that they were active. We also included speaker labels in each audio clips which are also taken from the mission transcripts. We collected onboard and commentary data from Apollo 11, and commentary data from Apollo 17, as indicated in Table 2.

The next step is about expanding the dataset. Data collection was done by selecting audio clips with a duration from 4 seconds to max 13 seconds, which is an appropriate duration to judge the overall quality and intelligibility [9, 22]. The audio clips must include active speech and we need to extract the corresponding text from the NASA transcripts so that we are able to assess speech intelligibility. We solved the problem of long periods of silence by exploiting the Mission Elapsed Time (MET) reported in the mission transcripts which tells us when speech is active. We also included speaker labels in each audio clips which are also taken from the mission transcripts. We collected onboard and commentary data from Apollo 11, and commentary data from Apollo 17, as indicated in Table 2.

To detect the likely excessive amount of repetitive audio clips i.e., clips that likely will be equally rated, and to explore mission context quality differences we computed Google STT WER on the expanded dataset. We also explored other objective metrics to confirm pilot study findings on the expanded dataset. In Figure 1 we show the histograms of Google STT WER and each non-intrusive objective metric distinguishing the two contexts: commentary and onboard. Google STT WER histograms present the distribution of WERs computed for the samples from the corpus. The histograms show clear quality differences between onboard and commentary. 103 audio clips of onboard
recordings show Google STT WER equal to one, which is a sign of repetitive data. A random sampling of audio clips from onboard recordings could generate a quality distribution skewed towards very low quality which is undesirable. This situation might naturally happen if we extract more clips, from other missions and mission contexts as well. Therefore, by using Google STT WER we prevent data repetition. For instance, we can remove a certain amount of audio clips with similar WER so to get a more uniform-like distribution. In the same plot (Figure 1) we show the distributions for the outputs computed using the above-mentioned non-intrusive objective metrics. We observe that SRMR and MOSNet do not capture a difference between commentary and onboard recordings, a distinction that is obvious in Google STT WER histograms. ITU-T P.563 shows results that are closer to Google STT WER which is aligned with the correlation analysis described in Table 1 where ITU-T P.563 shows a stronger correlation with subjective WER compared to the correlation between subjective WER and both MOSNet and SRMR prediction scores.

Next, we conduct an exploratory data analysis on the expanded dataset. It must be noted that at this stage we do not remove audio clips using Google STT WER because we want to also explore features of repetitive data. Data pruning will be performed before annotating the large-scale dataset through crowdsourcing. In this experiment we want to answer the following questions: 1) Can we find different degradations between different contexts (onboard and commentary) and within the same context? 2) To what extent are existing objective metrics capable of capturing the different degradations? The first question is crucial for selecting audio stimuli for the third step of the dataset creation which is not covered in this paper i.e., large-scale crowdsourced data annotation. By knowing signal differences we can expose participants to purposely selected stimuli, as it would happen in a controlled artificial quality dataset. It must be noted that even if we assess speech intelligibility in this study, these observations are meant to overall quality as well. We believe that in the third step it will be necessary to evaluate the overall quality and compare it with intelligibility, so to have a more accurate knowledge about quality on these data. The second question is to gain a deeper knowledge about the performance of existing objective metrics on this dataset, that will be added to the discoveries made in the pilot listening test.

Exploratory data analysis is made by computing the audio features shown in Table 3. We computed the actual values and the first-order difference of each feature except for Log Mel Filterbank Energies and Loudness. For each feature, we took the mean and the variance to integrate from frame-level to clip-level. Overall, we collected 253 features. Cluster analysis is performed using the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm [23] in the 2-dimensional space obtained with t-Distributed Stochastic Neighbor Embedding (t-SNE) [24] as shown in Fig. 2. We can see 7 clusters occurring. Commentary data is well separated from the onboard data and the Apollo 17 commentary is...
Figure 3: HDBSCAN clustering for each feature type. Each colour represents a different cluster with noise shown in dark grey. Each marker represents: ■ Apollo 11 onboard audio; ▼ Apollo 11 commentary audio; □ Apollo 17 commentary audio.

Figure 4: Clustered degradations found by HDBSCAN.

Figure 5: Speech quality database development timeline. Transparent steps are not covered in this paper.

which will be addressed before conducting the large-scale listening test.

5. Conclusions and Future Work

In this paper, we proposed a procedure to curate a real-world audio quality dataset and presented the outcomes from executing the first two stages, shown in Fig. 5. We used intelligibility as a proxy for quality. The methodology could be applied to any audio corpus within the context of archival recordings, given the similar issues of audio archives. The first part is the pilot listening test from which we show that Google STT WER can be used to avoid sampling repetitive data from the audio archive. The second is about the unsupervised data exploration from which we discovered 7 distinct degradations as shown in Figure 4. These findings are used to avoid mislabeling as a result of biases [11] caused by poor audio stimulus preparation in the large-scale listening test. The two stages are solidly connected. Google STT WER can be used to filter out repetitive audio clips. Then, clustering results applied to the filtered clips can be used to control each degradation individually, allowing a controlled listening test stimulus preparation aimed at avoiding mislabeling the data. From both experiments, we confirmed our expectation that existing objective quality metrics designed for other speech quality tasks fail to predict subjective quality ratings for the audio degradations tested. This reinforced our opinion that new quality metrics would be beneficial for audio archive speech quality prediction.

In the future, we will expand the dataset by adding audio clips from other Apollo missions and increasing the number of speakers. We will label the expanded dataset through crowdsourcing with both overall quality and intelligibility ratings by speakers. We will label the expanded dataset through crowdsourcing with both overall quality and intelligibility ratings by filling the last stage shown in Figure 5. We will use the findings described in this paper to avoid data repetition and mislabeling and we will study the Google STT WER distribution in each cluster to identify data that might cause model overfitting.

6. Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number 17/RG/483 and 17/RG/2289_P2 and was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1. The work of EB was supported by RAEng Research Fellowship RF/128 and a Turing Fellowship.
7. References


[14] A. Sangwan, L. Kaushik, C. Yu, J. H. Hansen, and D. W. Oard, “Houston, we have a solution”: using NASA Apollo program to advance speech and language processing technology,” in INTERSPEECH, 2013, pp. 1135–1139.


