Development of a Speech Quality Database Under Uncontrolled Conditions

Alessandro Ragano[1,2,3,4], *Emmanouil Benetos*[3,4], *Andrew Hines*[1,2]

[1]School of CS, University College Dublin, Ireland [2] Insight Centre for Data Analytics, Ireland
[3]School of EECS, Queen Mary University of London, UK [4]The 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 archive audio 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], impulsive 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).

[1]https://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. They self-declared no hearing disorders and that they had never listened to Apollo recordings before. They were informed about the topic of the recordings and they did a training session to gain familiarity with the data and to adjust the volume to a comfortable listening level before starting the test. The test was conducted in a silent room using professional studio headphones. Each participant was asked to repeat the words heard in the audio clips and their responses were noted.

We used 32 audio clips from the Apollo 11 mission distributed as follows: 21 commentary audio clips, 7 onboard audio clips, 2 post-mission press conference audio clips and 2 MCC audio clips. Each participant was assigned audio clips randomly. Loudness normalisation had been applied to all audio clips to avoid the introduction of a loudness bias. No restriction has been given in terms of test duration.

Speech intelligibility was measured using the word error rate (WER) [15] computed with the minimum-edit distance between the ground-truth sentence and the hypothesized sentence provided by each participant.

We compared subjective intelligibility scores with the performance of the Google Speech-to-Text (STT) API[2] measured with the same WER and the following non-intrusive objective metrics: ITU-T P.563 [16], speech to reverberation modulation energy ratio (SRMR) [17], and MOSNet [18]. Audio clips were downsampled to 8 kHz for ITU-T P.563 and to 16 kHz for SRMR and MOSNet to follow design guidelines of each objective metric. We computed the STT API, given that STT algorithms have been proven to be effective in predicting speech intelligibility [19, 20] and objective metrics have been used to predict STT performance [21].

Table 1: *Inferential statistical tests for assessing the correlation between subjective WER and objective metric prediction.*

Metric	Pearson coeff.	Pearson P-value	Spearman coeff.	Spearman P-value
Google STT WER	0.630	0.0001	0.679	1.93e-5
SRMR	0.081	0.658	0.112	0.538
ITU-T P563	-0.246	0.173	-0.295	0.100
MOSNet	-0.073	0.691	-0.163	0.371

Table 2: *Collected data*

Mission	Recording	Audio Clips	Speakers
Apollo 11	Onboard	266	4
Apollo 11	Commentary	220	8
Apollo 17	Commentary	51	5

We computed the Pearson and Spearman correlation coefficients which are reported in Table 1. The strength of association between Google STT WER and subjective WER is higher than other objective metrics ($Pearson = 0.630$ and $Spearman = 0.679$) and statistically significant ($P\text{-}value \leq 0.05$).

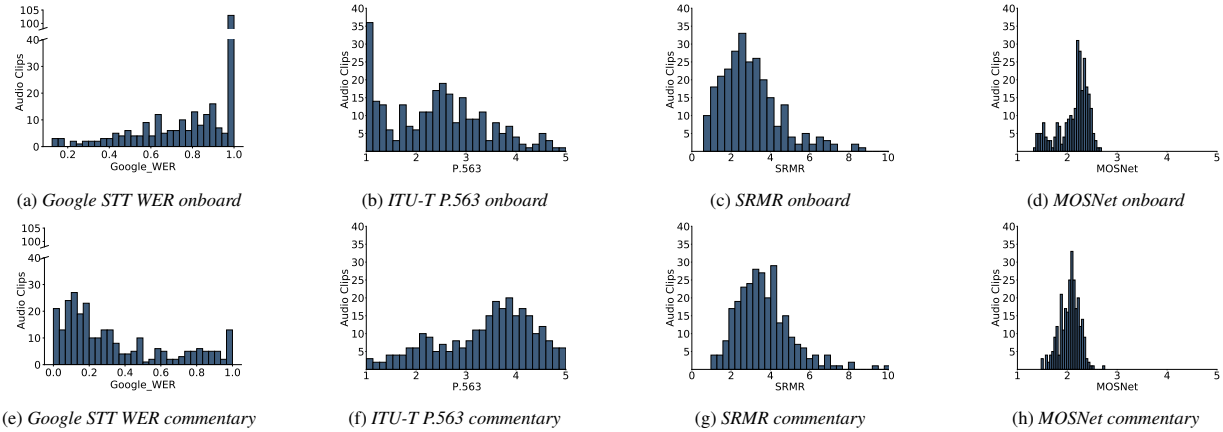
These results suggest that Google STT WER is an efficient metric to partially control the expected quality which helps in preventing repetitive data. We believe that poor correlation of objective metrics is because they were developed for audio degradations that are not found in the Apollo recordings.

4. Data Collection and Exploration

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

[2]https://cloud.google.com/speech-to-text, March 2020



(a) *Google STT WER onboard* (b) *ITU-T P.563 onboard* (c) *SRMR onboard* (d) *MOSNet onboard*

(e) *Google STT WER commentary* (f) *ITU-T P.563 commentary* (g) *SRMR commentary* (h) *MOSNet commentary*

Figure 1: *Distribution of non-intrusive objective metrics: Google STT WER, ITU-T P.563, SRMR, and MOSNet for onboard (a), (b), (c), (d) and commentary (e), (f), (g), (h) respectively.*

Table 3: *Feature list*

Feature	**Group**
26 MFCCs	Spectral
10 Linear Prediction Coefficients	Spectral
Spectral Centroid	Spectral
Spectral Bandwidth	Spectral
Spectral Roll Off	Spectral
Spectral Flux	Spectral
26 Log Mel Filterbank Energies	Spectral
10 Spectral Subband Centroids	Spectral
Root Mean Square	Intensity
Loudness	Intensity
Peak Envelope	Intensity
Tonal Power Ratio	Tonalness
Spectral Flatness	Tonalness
Zero-Crossing-Rate	Tonalness
Spectral Crest Factor	Tonalness

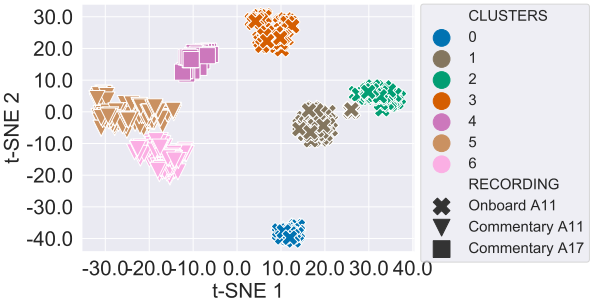


Figure 2: *HDBSCAN clustering of features shown in Table 3.*

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

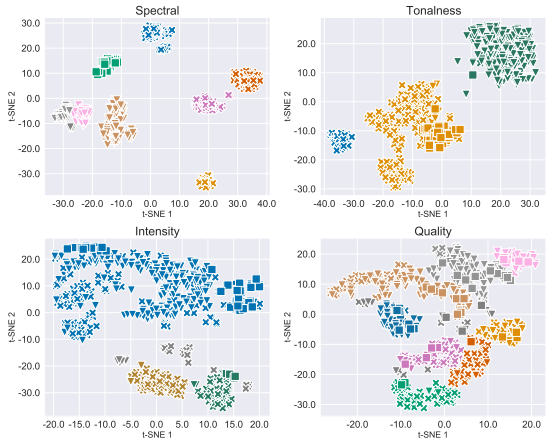


Figure 3: *HDBSCAN clustering for each feature type. Each colour represents a different cluster with noise shown in dark grey. Each marker represents: ✖ Apollo 11 onboard audio; ▼ Apollo 11 commentary audio; ■ Apollo 17 commentary audio.*

separated from the Apollo 11 commentary. We can notice that onboard audio clips are divided into 4 clear distinct groups, revealing different signal characteristics.

To gain a deeper insight into the collected data we performed a cluster analysis of different feature subgroups, using the taxonomy reported in Table 3. We cluster spectral, tonalness, intensity and quality features as shown in Fig. 3. Quality features refer to the above-mentioned objective metrics ITU-T P.563, SRMR, and MOSNet without the inclusion of Google STT WER. Spectral feature clusters are similar to the clusters obtained when considering all the features together as shown in Figure 2. This suggests that variations in spectral differences are a good indicator to identify candidate samples that will yield a range of quality ratings when stimuli will be evaluated in the large-scale crowdsourced listening test. Cluster analysis of tonalness and intensity revealed little heterogeneity with three clusters for each feature type.

To quantify how many audio clips belong to the same cluster between each feature type clustering, we calculated the adjusted Rand index [25] shown in Table 4 for each feature type pair. The low adjusted Rand index between quality metrics and each feature type means that existing objective metrics do not cluster similarly to audio characteristics. The high adjusted Rand index (0.903) between spectral clusters and all the clustered features confirms that spectral characteristics have the most variation within the Apollo data.

In order to establish a balance of degradations across the data, in Figure 4 we show clusters found by HDBSCAN. The doughnut chart suggests a small imbalance in the collected data,

Table 4: *Adjusted Rand index between different feature type clusterings.*

Feature Type	All	Spectral	Tonalness	Intensity	Quality
All	1.0	0.903	0.368	0.147	0.204
Spectral	0.903	1.0	0.325	0.125	0.193
Tonalness	0.368	0.325	1.0	0.042	0.071
Intensity	0.147	0.125	0.042	1.0	0.094
Quality	0.204	0.193	0.071	0.094	1.0

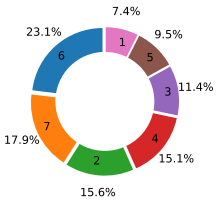


Figure 4: *Clustered degradations found by HDBSCAN.*

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 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/RC-PhD/3483 and 17/RC/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

[1] D. Schuller, "Preserving the facts for the future: principles and practices for the transfer of analog audio documents into the digital domain," *Journal of the Audio Engineering Society*, vol. 49, no. 7/8, pp. 618–621, 2001.

[2] F. Bressan and S. Canazza, "A systemic approach to the preservation of audio documents: Methodology and software tools," *Journal of Electrical and Computer Engineering*, vol. 2013, 2013.

[3] A. Ragano, E. Benetos, and A. Hines, "Adapting the quality of experience framework for audio archive evaluation," in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2019.

[4] C. F. Stallmann and A. P. Engelbrecht, "Gramophone noise detection and reconstruction using time delay artificial neural networks," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 6, pp. 893–905, 2016.

[5] M. Ciołek and M. Niedźwiecki, "Detection of impulsive disturbances in archive audio signals," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 671–675.

[6] S. J. Godsill and P. J. W. Rayner, *Digital Audio Restoration*, 1st ed. Springer-Verlag London, 1998.

[7] A. Ragano, E. Benetos, and A. Hines, "Audio impairment recognition using a correlation-based feature representation," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2020.

[8] A. Ziaei, L. Kaushik, A. Sangwan, J. H. Hansen, and D. W. Oard, "Speech activity detection for NASA Apollo space missions: Challenges and solutions," in *INTERSPEECH*, 2014.

[9] N. Harte, E. Gillen, and A. Hines, "TCD-VoIP, a research database of degraded speech for assessing quality in voip applications," in *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, 2015.

[10] J. 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 *INTERSPEECH*, 2019.

[11] S. Zielinski, F. Rumsey, and S. Bech, "On some biases encountered in modern audio quality listening tests-a review," *Journal of the Audio Engineering Society*, vol. 56, no. 6, pp. 427–451, 2008.

[12] "Apollo lunar surface journal," https://history.nasa.gov/alsj/main.html.

[13] J. H. Dabbs and O. L. Schmidt, "Apollo Experience Report - Voice Communications Techniques and Performance," *NASA TECHNICAL NOTE NASA TN D-6739*, 1972. [Online]. Available: https://www.hq.nasa.gov/alsj/tnD6739VoiceCommTechnqs.pdf

[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.

[15] A. C. Morris, V. Maier, and P. Green, "From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition," in *Eighth International Conference on Spoken Language Processing*, 2004.

[16] L. Malfait, J. Berger, and M. Kastner, "P.563 — The ITU-T standard for single-ended speech quality assessment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1924–1934, 2006.

[17] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.

[18] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, "MOSNet: Deep learning-based objective assessment for voice conversion," in *INTERSPEECH*, 2019, pp. 1541–1545.

[19] C. Spille, S. D. Ewert, B. Kollmeier, and B. T. Meyer, "Predicting speech intelligibility with deep neural networks," *Computer Speech & Language*, vol. 48, pp. 51–66, 2018.

[20] L. Fontan, I. Ferrané, J. Farinas, J. Pinquier, J. Tardieu, C. Magnen, P. Gaillard, X. Aumont, and C. Füllgrabe, "Automatic speech recognition predicts speech intelligibility and comprehension for listeners with simulated age-related hearing loss," *Journal of Speech, Language, and Hearing Research*, vol. 60, no. 9, pp. 2394–2405, 2017.

[21] L. F. Gallardo, S. Möller, and J. Beerends, "Predicting automatic speech recognition performance over communication channels from instrumental speech quality and intelligibility scores," in *INTERSPEECH*, 2017, pp. 2939–2943.

[22] ITU-T, "P.800 Methods for subjective determination of transmission quality," *Series P: Telephone Transmission Quality*, 1996.

[23] R. J. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 2013, pp. 160–172.

[24] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[25] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.