Engaging Vs. Enraging; Generating Compelling Speech

Amy Cao
caoamyc@sfu.ca
Simon Fraser University
Burnaby, BC, Canada

Ryan Zrymiak
rza80@sfu.ca
Simon Fraser University
Burnaby, Canada

Eric Wang
ywa353@sfu.ca
Simon Fraser University
Burnaby, Canada

Sina MohammadiNiyaki
sma231@sfu.ca
Simon Fraser University
Burnaby, Canada

ABSTRACT

Robotics are a vital part of affective computing. There is rapid technological growth with the increasing usage of human-robot interaction and artificial intelligence (AI). However, one critical aspect remains underdeveloped: the emotional resonance of robot voices. Current synthetic voices lack the depth necessary for genuine emotional connections between humans and robots. This study addresses this gap by proposing to generate compelling voices for robots using affective computing principles. Leveraging advanced AI algorithms, we apply different voice features to generate different styles of voice, allowing our survey participants to make comparisons. We then extract the features from the feedback, drawing meaningful conclusions. The main advantage of our approach lies in our ability to infer from real user data, fostering more natural and accurate interactions between humans and robots. By contributing an innovative solution to the challenge of compelling sounds in robotics, this research looks to contribute to the next generation of AI-generated speech in various fields including healthcare, education, and customer service.

KEYWORDS

Spectrogram, Speech, Compelling Communication, Neural Network, Generative Adversarial Network

1 INTRODUCTION

Speech is an important part of communication. Compelling speech has the power to capture the attention of its listeners, but boring speech can lose it completely. What is it that speakers use to keep their audiences engaged, and how can these features be emulated in an affective system?

Our project aims to consider several different features that contribute to effective speech, synthesize new speech samples that consider these features, and enlist participants to assess the effectiveness of these speeches.

Many researchers in the past have focused their efforts on speech synthesis. In *Tacotron: Towards End-to-End Speech Synthesis* by Wang et al.[11], the authors create a model which can produce natural-sounding phrases based on a textual input. However, we take this a step further by trying to achieve sounds that attract listeners' attention instead of just trying to mimic human-like sounds. In *The sound of emotions—Towards a unifying neural network perspective of affective sound processing*, Frühholz et al. [10] take a deep dive into how changes in sound features can elicit emotional responses from the brain. Combining affective sound processing with speech synthesis seems to be a problem that few have considered, which makes our contributions especially important.

2 APPROACH

2.1 Processing Data

The nowwithfeeling dataset of audio files is used for K-means clustering and predicting the effectiveness of speech through a Neural Network. An annotator labels the data for agreeableness on a scale from 1 to 5, and a binary value for if the speech is compelling. These two labels will be used to train the Neural Network.

Audio features from the dataset are extracted using the Python package PyAudioAnalysis[15]. These features are from Mid Term segments of the audio. As each audio file produces multiple windows of audio features, two data sets can be formed from the extracted window features; a data set where each row is a window segment used for machine learning, and a data set with each file's windows all aggregated into a mean value used for visualization.

The machine learning data set is further split into train, validate, and test data sets at 60,20,20 percent of the total data.

With 136 features, dimension reduction is performed using Principal Component Analysis (PCA) on the visualization data set to reduce complexity. Two principal components are kept as features for clustering. They hold 95 percent of the total variance in the data.

2.2 Predicting Effectiveness of Speech

A Multi-output Regression Neural Network is created to predict the level of agreeableness and presence of compelling speech. The model is created using the Keras library[9]. The model takes audio features as input, and outputs continuous numbers on level of agreeableness and compelling. The model is trained on the machine learning data sets as described in section 2.1 The Neural Network comprises dense layers and a drop-out layer. The activation function chosen was the Rectified Linear Unit (ReLU), with the final activation function as Linear.

Generated Audio is then processed with PyAudioAnalysis similarly to the training data, and is inputted to the model. The model provides another angle of evaluating the effectiveness of our generated speech.

2.3 Generating Effective Speech

2.3.1 Processing the Audio Files. We automated the extraction of crucial audio features from `.wav` files using the `librosa` [8] library. This process involved:

(1) **Feature Extraction**: Identification of key audio characteristics, such as Mel-frequency cepstral coefficients (MFCCs), zero-crossing rate, and spectral properties, for comprehensive voice analysis.
(2) **Dataset Compilation**: Aggregation of features into a structured dataset, streamlined for further analysis and model training.

2.3.2 Voice Classification Model. Our approach to voice classification encompasses:

(1) **Data Enrichment**: Manual augmentation of the dataset with qualitative traits like clarity and agreeableness to deepen our analysis.
(2) **Preprocessing**: Optimization of the dataset by refining feature representations and focusing on variables critical for classification.
(3) **Model Development**: Adoption of the Gradient Boosting Classifier[14], chosen for its robustness, facilitated by detailed preprocessing and feature scaling. The model's efficacy was evaluated using precision, recall, and F1-scores among other metrics.

2.3.3 Text-to-Speech Synthesis. To validate our findings and enrich our dataset, we integrated Text-to-Speech (TTS) technologies:

(1) **Google Cloud's TTS API**[12]: Enabled dynamic audio file generation, exploring voice diversity and the impact of voice characteristics on engagement.
(2) **Amazon Polly Integration**[6]: Expanded our project's capabilities to include a wider array of voices and languages, utilizing Polly's neural engine for natural-sounding speech synthesis.

Each phase, from audio processing and classification to text-to-speech synthesis, was pivotal in advancing our understanding of what constitutes effective speech, demonstrating the model's real-world applicability through diverse, synthesized voices. For more detailed explanations, methodologies, and analysis, refer to the extensive markdowns within our Jupyter notebook (Audio_processing_and_voice_generation.ipynb).

3 DATASET

3.1 Speech Data

We collected 162 voice data from nowwithfeeling.com, each consisting of a 15s to 25s long voice speech in .wav format. An example of the speech can be seen below. [3]

> A house has all kinds of stuff. Before you build your house, you must know what you want to do. Igloos made of snow keep you warm. Wood and stone houses with sloping roofs make the rain and snow run right off.

In addition, we downloaded about 118 hours of speech from TED-LIUM. The TED-LIUM dataset contains speech samples from 2 minutes to 45 minutes. Here is an example of part of the ChristopherdeCharms2008 speech: [4]

> You can mimic what you can see, you can program hundreds of muscles in your arms. Soon you will be able to look inside your brain and program control hundreds of brain areas you see there, I'm going to tell you about that technology.

We set the 162 nowwithfeeling data as our primary data to use in the classifier because they have consistent speech content and consistent length. We had trouble with the TED-LIUM dataset because of the size, the format, the random noise, and the inconsistency of the length of talks. Therefore, we use the TED-LIUM dataset as backup data and references.

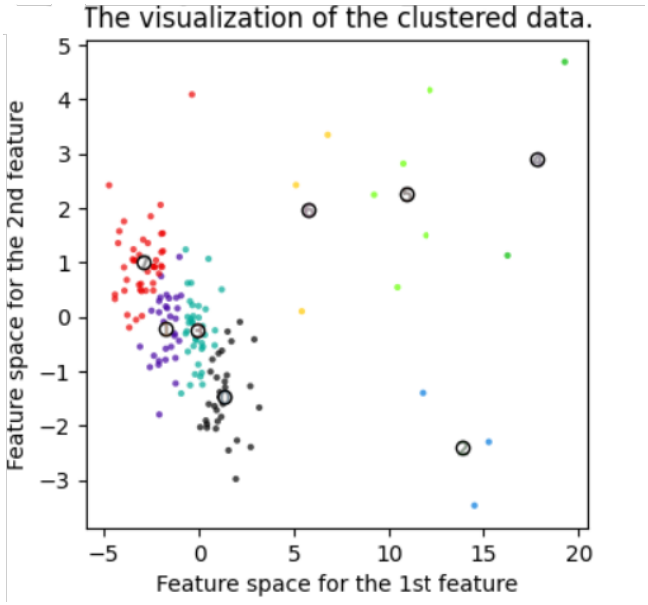

Figure 1: K-Means Clustering with K=8

3.2 Data Visualization

To gleam insights on whether or not different speeches hold similarities and thus can be grouped together, we perform K-means clustering on the visualization data set processed from the nowwithfeeling dataset. There are other clustering methods; however, K-means clustering is simple to use and is computationally inexpensive. Gaussian Mixture Model was considered but was not chosen due to hardware limitations.

The value of K used is 8 as it was found to be the most optimal value using the Elbow Method[13].

The results of K Means Clustering shows separate clusters of outlier data. Perhaps more data will prove the spread out clusters to hold significance; however, in this case Interestingly, these clusters do not hold the same values of agreeableness or compelling as visually observed from graphs.

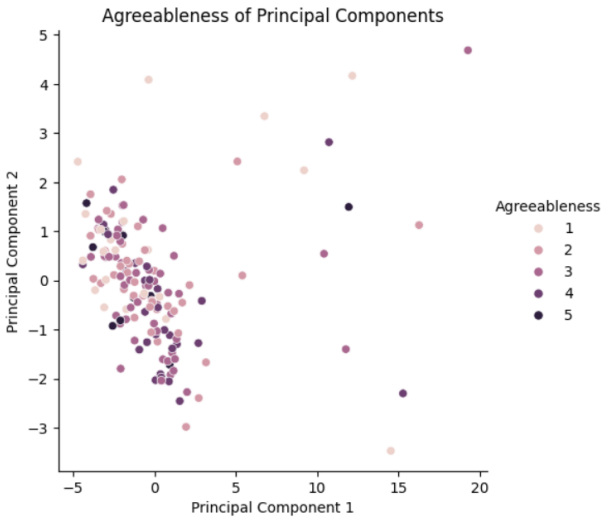

Figure 2: Agreeableness of Principal Component Data

Figure 3: Survey and Response

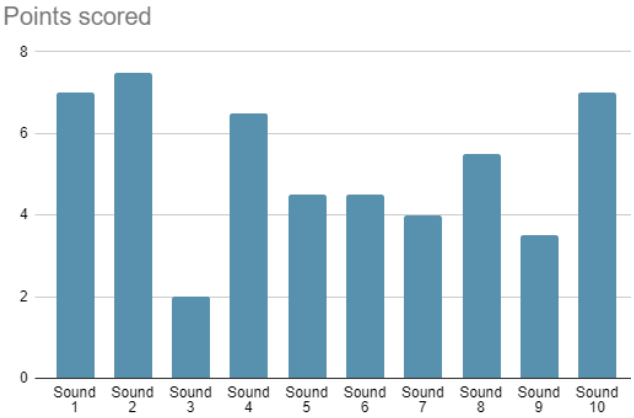

Figure 4: Internal Survey

4 EXPERIMENTS AND RESULTS

We evaluate our system by conducting surveys with human participants . Once we had processed all of our datasets and generated our own audio samples, we were then able to play those audio samples for willing participants to record their feedback. By enlisting participants to offer their thoughts on the agreeableness and engagement of the audio clips we had to share for them, we hoped that their recorded responses could allow us to validate which features of audio samples contributed to a more compelling speech.

In addition, we evaluate our audio clips from a machine learning approach. Our Neural Network is trained with nowwithfeelings data at 10 epochs and batch size of 1048. The mean predicted agreeableness is 2.76 out of 5, where 5 is the highest level of agreeableness. The mean predicted compelling value is 0.45, where 1 is the highest level of compelling.

4.1 Conducting Surveys

For our participants, we would play a generated voice clip with some enhanced verbal features or features (e.g. higher volume, faster tempo, varying pitch, etc.), and compare it with a generated voice clip where the verbal features were not adjusted. Then, for each audio sample played, we would ask our participants to rate the agreeableness and engagement of that particular audio sample on a scale of 1 to 10, with a score of 1 corresponding to an audio sample that is "Not agreeable/engaging at all" and a score of 10 corresponding to an audio sample that is "Extremely agreeable/engaging". See Figure 3 for a sample survey done by one of our participants.

We took the survey internally first to ensure the smoothness of the process, and to have the expected survey results. Here is our expected result shown in Figure 4.

4.2 Interpreting Survey Results

We took all 19 survey results, added a total score to each voice sample and divided the total by 19 to find an average. Here is the bar graph for our final survey results: Most samples scored near the average, which might indicate our generated sounds are neither compelling nor uncompelling. There are two clear outliers, sound number 2 and sound number 6. Sound number 2 being the highest falls into our expectations during internal testing. Sound number 6 unexpectedly scored the lowest. Sound number 3 was expected to score the lowest, but the final result displays it to be slightly below average. See Figure 5.

4.3 Statistical Testing

The most significant difference between Sound 2 and Sound 6 is their speaking rate. Therefore, we propose a statistical test on whether the speaking rate affects our research. From our survey, the top three speaking rate sound samples have scores of 7.16, 5.58, and 5.63. with a mean equal to 6.12 and a standard deviation of 0.90.[1] The other seven samples have a mean equal to 5.01. Null hypothesis H0: The speaking rate has no effect on the score of the speech. Alternative hypothesis Ha: The speaking rate has an effect on the score of the speech. p-value = 0.05. Assume H0 is true, the sampling distribution is 0.9/sqrt(3) = 0.52. The z statistic = (5.01-6.12)/0.52

	Score
Sound 1	5.421052632
Sound 2	7.157894737
Sound 3	4.947368421
Sound 4	4.947368421
Sound 5	5.263157895
Sound 6	3.421052632
Sound 7	5.578947368
Sound 8	6.526315789
Sound 9	5.631578947
Sound 10	4.578947368

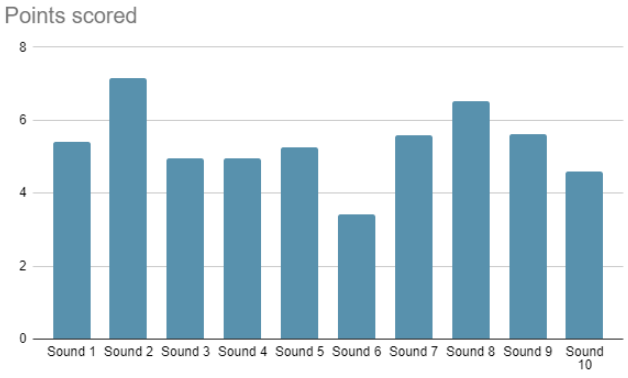

Figure 5: Final Survey Graph

= -2.13.[5] This means an average score of 6.12 is a 2.13 standard deviation away from the mean. According to the z-score table, if we assume H0 is correct, the probability of getting a sample of our result is 0.0332. It is below the p-value of 0.05. Therefore, we reject the null hypothesis H0: The speaking rate has no effect on the score of the speech. This statistical testing shows there is a very strong indicator that the speaking rate has an effect on the score of our surveys. [2]

4.3.1 Comparison of Objective Measures and Subjective Perceptions. Our study aimed to objectively identify voices with compelling attributes by comparing generated audio features against a set of compelling averages, using Euclidean distances for measurement. Yet, our observations indicate a clear divergence between these objective metrics and the subjective listener feedback presented in the survey.

These results imply that quantitative audio statistics, though valuable, fall short of encompassing the factors that shape human perceptions of a voice's charisma and appeal. Indeed, voices that were a close match to our compelling standard based on Euclidean distance did not necessarily captivate listeners.

Highlighting the intricacies of vocal evaluation, our research advises against relying solely on objective data. It points to the vital role of subjective perception in voice quality assessments, advocating for future studies to merge quantitative analyses with qualitative listener responses.

5 DISCUSSION

5.1 The Problems

After recording all of our participants' responses, we observed that there was a noticeable variation in the individual ratings of participants, but on average, 80% of the voices that were rated had an average score greater than 4/10 and less than 7/10. This result made our next stage of interpreting results more difficult as we were expecting extreme scores. Since most samples are close to the overall mean, it is hard to justify excluding them in our next round of sound generation. This finding can represent different things. One explanation suggests our generated voices are neither very compelling nor very dull. The other explanation is people simply interpret compelling voices differently, which we think is more plausible.

5.2 Possible Solutions

One solution to our survey problem is to generate sounds with more emphasized features. For example, if we are to still use speaking rate, pitch, and accent as our parameters, some samples should have two parameters with low and equal values, but one parameter has various values. We believe doing this allows us to interpret a specific feature more directly.

The other solution is to conduct a more complex survey which involves asking participants' background, preferences, and interpretation of compelling sound. For example, the new survey asks "Which country are you from? Do you prefer teachers who speak fast? Do you think robotic voices are compelling?" We prefer this solution over the other but acknowledge it will require significantly more commitment from our participants and us. We consider this to be an extension of our future project, and we are excited to see the effect of adding personal factors as new parameters.

6 CONCLUSION

In conclusion, our project set out an important lesson to address the individual interpretation of compelling sound in human-robot interaction. Through the usage of affective computing principles and AI algorithms, we successfully generated compelling voices, conducted surveys, and improved the quality of the voices as a result. We encountered many complexities inherent in sound generation. During the voice generation stage, although we completed data classifying, utilizing machine learning such as CNNs and RNNs remains challenging. We decided to utilize the classified results and apply features to existing sound generation models. The planned second round of sound generation utilizing GANs was also modified due to the unexpected survey results, as explained in our discussion. Utilizing GANs was proven to be a challenging task as we were only able to generate white noise. We modified round two sound generations by conducting a statistical test and applied features using Amazon Polly. Moving forward, we would love to incorporate more advanced machine learning models, and more parameters into our studies, such as an individual's cultural background. Nevertheless, We believe by leveraging real user data and applying innovative ideas, our efforts serve as a significant contribution to the affective computing field.

REFERENCES

[1] [n. d.]. Calculating Standard Deviation Step by Step. https://www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/variance-standard-deviation-population/a/calculating-standard-deviation-step-by-step

[2] [n. d.]. Hypothesis testing and p-values | Inferential statistics | Probability and Statistics | Khan Academy. https://www.youtube.com/watch?v=-FtlH4svqx4

[3] [n. d.]. Now With Feeling. https://nowwithfeeling.com/

[4] [n. d.]. TED-LIUM Corpus. https://www.tensorflow.org/datasets/catalog/tedlium

[5] [n. d.]. Z-Table. https://www.z-table.com/

[6] Amazon Web Services. [n. d.]. *Amazon Polly.* https://aws.amazon.com/polly/

[7] Yannick Estève Anthony Rousseau, Paul Deléglise. 2012. TED-LIUM: an Automatic Speech Recognition dedicated corpus. http://www.lrec-conf.org/proceedings/lrec2012/pdf/698_Paper.pdf

[8] Brian McFee and Colin Raffel and Dawen Liang and Daniel PW Ellis and Matt McVicar and Eric Battenberg and Oriol Nieto and Tej Chajed and Zhiyuan Luo and Carrie Grimes and others. 2022. *LibROSA Documentation.* librosa development team. https://librosa.org/doc/latest/index.html

[9] Francois Chollet et al. 2015. *Keras.* https://github.com/fchollet/keras

[10] Fruhholz et al. 2016. The sound of emotions—Towards a unifying neural network perspective of affective sound processing. https://www.sciencedirect.com/science/article/pii/S0149763416300082?casa_token=6eHpMMbLrBcAAAAA:Kk1KyYGVaIAmM_FPtFjzl6-XCCxket2_c6jwqpIRgxMZkeQtgfd73S3ZdiwkGoG667IZxnwy_A

[11] Wang et al. 2017. Tacotron: Towards End-toEnd Speech Synthesis. https://arxiv.org/abs/1703.10135

[12] Google Cloud. [n. d.]. *Text-to-Speech AI.* https://cloud.google.com/text-to-speech

[13] Dhendra Marutho, Sunarna Hendra Handaka, Ekaprana Wijaya, and Muljono. 2018. The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News. In *2018 International Seminar on Application for Technology of Information and Communication.* 533–538. https://doi.org/10.1109/ISEMANTIC.2018.8549751

[14] scikit-learn contributors. [n. d.]. *sklearn.ensemble.GradientBoostingClassifier.* https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html

[15] Tyiannak. [n. d.]. GitHub - tyiannak/pyAudioAnalysis: Python Audio Analysis Library: Feature Extraction, Classification, Segmentation and Applications. https://github.com/tyiannak/pyAudioAnalysis

A APPENDIX

3.1 Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? "Nowwithfeeling is a collaborative site for psychologists and researchers to collect affective information around the globe through fun activities. Emotion can be expressed in different ways depending on cultural, context, and individual differences. Collecting affective information will help us understand more about how people express and interpret emotions!" [3] TED-LIUM is an English-language TED talk dataset. It is to provide a resource to TensorFlow catalog [4]

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? Nowwithfeeling is created by the ROSIE Lab led by Dr. Angelica Lim, Assistant Professor of Professional Practice in the School of Computing Science at Simon Fraser University. [3] TED-LIUM dataset is created by Anthony Rousseau, Paul Deléglise, Yannick Estève, from Laboratoire Informatique de l'Université du Maine (LIUM), University of Le Mans, France. [7]

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number. Unable to find the information.

3.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description. They represent people and their voices. TED-LIUM also contains TED talks interacting with other people.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? All instances.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? nowwithfeeling consists of .wav sound files. TED-LIUM consists of .sph sound files.

Is there a label or target associated with each instance? If so, please provide a description. No.

Are relationships between individual instances made explicit? No.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them. TED-LIUM provided default training and testing split. However, they do not apply to us.

Are there any errors, sources of noise, or redundancies in the dataset? We found duplicate data inside the interestingdata folder in nowwithfeeling. We did not use them.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? They are self-contained.

Does the dataset contain data that might be considered confidential? No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? The voices of both datasets, and the content of the TED talks dataset might make people uncomfortable.

Does the dataset identify any subpopulations (e.g., by age, gender)? No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? The TED-LIUM dataset has individual names written on the file names.

Does the dataset contain data that might be considered sensitive in any way? The TED-LIUM dataset might contain TED talks that are considered sensitive to people.

3.3 Collection Process

How was the data associated with each instance acquired? Data was reported through survey responses.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? A software program was used to record participants' responses.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? The dataset was not sampled from a larger set.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)? Family and friends were surveyed for data collection. They were not compensated for their responses.

Over what timeframe was the data collected? The data was collected during the week of April 7, 2024 to April 14, 2024.

Were any ethical review processes conducted (e.g., by an institutional review board)? No ethical review processes were conducted.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? Data was collected from surveying individuals directly.

Were the individuals in question notified about the data collection? Yes.

Did the individuals in question consent to the collection and use of their data? They were informed that participating in the survey meant that they were given us consent to collect and use their responses in our dataset.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? No, the dataset will be erased following the conclusion of our research.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? No.

Any other comments? No.

3.4 Preprocessing/cleaning/labeling

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? No.

3.5 Uses

Has the dataset been used for any tasks already? Yes, we used nowwithfeeling for classification.

Is there a repository that links to any or all papers or systems that use the dataset? No.

What (other) tasks could the dataset be used for? Other research in voice-related affective computing. For example, teaching voice research.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? It is possible. Voice datasets contain different ways of speaking. There can be people speaking from different cultural backgrounds. Some research might conclude a stereotyping result such as voices from ... background is less compelling than voices from ... background. For example, to prevent harms such as stereotyping, researchers must not label them based on the guessed culture of the sound origin.

Are there tasks for which the dataset should not be used? We standby that the dataset should not be used on any tasks that are considered unethical, such as making cultural stereotypes, system manipulation/nudging/deception, systems with their own emotions, algorithmic bias, etc...

Any other comments? No.

A.1 Contributions

Amy Cao:

- Processed nowwithfeeling and generated datasets
- Represented data in various visual forms
- Created a Neural Network for predictions on generated audio

Sina MohammadiNiyaki:

- Processed nowwithfeeling and manually labelled the dataset
- Implemented sound generation using Google Cloud TTS and Amazon Polly
- Helped write the report

Eric Wang:

- Initial sound generation models using RNNs and Encoder/Decoder. Unsuccessful.
- Data collection
- Poster
- Conducting surveys and processing results
- Writing the report

Ryan Zrymiak:

- Developed initial (unsuccessful) sound generation model
- Developed survey response recorder
- Surveyed participants for voice compellingness feedback
- Wrote report Introduction and Experiment & Results sections