Kamila Barrios

James Hatch

Ana Leon Urrita

Bias in Emotion Recognition:
How Accent Shapes AI's Perception of Speech

Introduction and Data

1.1. Motivation and Context

The increasing deployment of Artificial Intelligence (AI) in high-stakes professional settings, such as virtual interview assistance and candidate screenings, necessitates a rigorous examination of algorithmic fairness and accuracy. Many of these services incorporate emotion recognition modules that claim to objectively assess a candidate's emotional state based on vocal tone. However, the use of such technology raises profound ethical concerns regarding algorithmic bias, particularly when assessing communication styles tied to demographic characteristics like accent, ethnicity, and age. The prevalence of services like Sanas and Krisp highlights a societal preference for certain vocal norms, compelling us to question the reliability of classification systems applied to diverse voices.

This project focuses specifically on the susceptibility of emotion recognition models to biases associated with a speaker's accent. Prior research has established that models in domains like facial recognition perpetuate widespread racial biases due to non-representative training data (Scheuerman et al., 2021). We hypothesize that a similar issue exists within speech-based AI. While accent is not race, it serves as a powerful linguistic marker conveying geographic or ethnic origin, placing it in a similar sphere of potential bias. Critically, the human tendency to perceive speech through the lens of accent bias provides a strong theoretical foundation for this study. As demonstrated by Jiang et al. (2019), the presence of an unfamiliar or "outgroup" accent can significantly alter how human listeners interpret speech. In the context of AI, this bias is likely due to data imbalance—where models, trained predominantly on voices with a majority accent (e.g., General American English), struggle to accurately classify emotion in underrepresented accents, leading to systematic misclassification and potentially unfair outcomes.

1.2. Research Questions and Hypotheses

This study seeks to answer the following questions:

- How does a single, trained emotion-classifying model respond to speech exhibiting unfamiliar accents?
- Is there a bias in the classification of basic emotional valences (e.g., negative vs. positive) across different accents when analyzed by an emotion-recognition model?
- Is there a detectable bias for classifying certain emotions in relation to the speaker's age, as categorized by the model?

We hypothesize that models trained primarily on current emotional speech datasets will exhibit a significant bias, attributing a higher rate of negative emotions (such as anger and sadness) to voices with non-American accents compared to voices with a predominant American accent. We believe this discrepancy will be driven by the model's unfamiliarity with the vocal characteristics of diverse accents.

1.3. Data Description and Preparation

To address these questions, we employ three key audio datasets:

- Crema-D (Crowd-sourced Emotional Multimodal Actors Dataset): Provides 7,442 sound bites from 91 actors (aged 20–74) expressing six emotional states. This will be the primary source for training our emotion classification models.
- Speech Accent Archive: Contains 2,140 audio samples from English speakers across 177 countries and 214 native languages. This serves as the novel test set for introducing diverse accents.

Data Preparation, Variable Transformations, and Model Inclusion

The most significant initial transformation is Feature Extraction: converting all raw audio files into a sequence of Mel-Frequency Cepstral Coefficients (MFCCs). MFCCs (typically 13 coefficients per audio frame) are the industry standard for emotional speech analysis, effectively representing the spectral envelope of the sound. All files are standardized to a uniform sampling rate and trimmed of silence to ensure feature consistency. This MFCC Sequence is the sole numerical input variable for our model.

Metadata requires significant aggregation for statistical testing. Target Variable Standardization consolidates the six Crema-D emotions into three simplified Emotional Valence categories: Negative (Anger, Disgust, Fear, Sadness), Positive (Happiness), and Neutral. Separately, the raw geographic metadata from the Speech Accent Archive is tidied by mapping specific countries into a manageable set of Accent Groups (e.g., North American, South Asian, Slavic, Western European). This Accent Group variable, along with demographic variables Age and Sex, serves as the Explanatory Variables for the Kruskal-Wallis H Test used to identify post-model bias. Samples with poor audio quality or missing critical demographic data are excluded from the final analysis set.

Methods

Experiment 1 - Training an LSTM to create audio-based emotion classifications

To evaluate the level of accent bias in the Crema-D dataset, we decided to create a neural network that predicts the emotional valence of recordings of human speech. In order to accomplish this, we decided to rely on a series of time-based audio features that were successively fed into an LSTM. An LSTM (long short-term memory) is a type of recurrent neural network that keeps track of previous model outputs as it learns. By keeping track of its previous states and feeding those states back into the model, LSTMs are able to track patterns that are displayed over the course of successive model inputs. Since we plan to extract our audio data using spreads of numerical representations that span the course of the audio clip, an LSTM will allow us to identify patterns that occur over the entire length of each clip. This should allow us to identify successive changes in tone over time that are indicative of different emotions.

In order to generate numerical representations of our audio files, we will utilize Mel-Frequency Cepstrum Coefficient features (or MFCCs). MFCCs are time-based power representations of sound timbre. By utilizing a mel-scale transformation, a scale that is more susceptible to low-frequency changes than higher frequency changes, MFCCs are able to mathematically describe sound in ways that mirror human sound perception. Additionally, MFCC’s ability to describe sound timbre, a significant characteristic of human speech, makes it incredibly effective when attempting to train deep learning classifiers on human speech patterns.

In order to extract the MFCC data from each audio file in the Crema-D dataset, we utilized the Python-based librosa audio-analysis package. When loading librosa features, most of the variables were set to default values. Since the librosa library's default values are already tuned for human speech analysis, we ended up preserving most of the MFCC-loading variables. The only variation from this was the value of MFCC features (for variable n_mfcc), which we set at 40 MFCC in order to obtain a higher level of detail for each continuous audio clip.

As MFCC features were extracted from each audio file, each feature array was appended to a list that would eventually be used to fit the LSTM model. MFCC features in the list were then padded in order to ensure that each series of MFCCs was the same length (preventing audio clip length from disrupting the model's input size). This preprocessing sequence was utilized to extract our audio variables from both the Crema-D and Speech Accent Archive audio files.

To create validation data for our LSTM model, we also had to extract emotion labels from the Crema-D dataset. This was completed by extracting the emotion labels from the path name of each Crema-D audio file. To allow the LSTM model to interact with the emotion-based validation data, we used TensorFlow and Keras to create categorical embeddings for each emotion. These embeddings would serve as the y-variable in our model.

Our LSTM model was created using the Keras and Tensorflow APIs. The architecture of the model is as follows: the model accepts an input whose shape matches the number of MFCCs defined when loading the audio file features using the librosa library (n_mfcc). It contains 2 LSTM layers (each with 250 units), an initial dense layer (with an l2 regularizer and a relu activation) with 100 units, a dropout layer with a value of 0.3, a second dense layer with 80 units (with an l2 regularizer and a relu activation), and an output layer that outputs (a one-hot encoding that describes) one of six possible emotions embeddings.

After creating the LSTM model, we compiled it with an Adam optimizer that was set to a learning rate of 0.001. We also established categorical cross-entropy and categorical accuracy as the loss and metric values, respectively. After compiling, the model was fit using 40 epochs and a batch size of 15. During training, input examples were shuffled in order to minimize overfitting caused by repeated audio clips taken from the same subject. We also implemented early stopping callbacks in order to allow the model to stop training on its own in cases of overfitting.

Ultimately, this model was used to generate emotion predictions on each of the Speech Accent Archive accent audio files. These predictions were then grouped together to form three graphs. The first two graphs were used to display the emotion counts for the ten most represented accents in the Speech Accent Archive. Since the accent archive contained 200 accents total (with some receiving significantly fewer examples than others), we reasoned that collecting individual data for only the ten most represented countries would prevent accents with fewer recordings from skewing results. As a result, all other accents that were not represented in the ten most represented accents were grouped into a third and final graph.

To account for differences in accent representation, we modified our three, count-based graphs to display relative frequencies (in terms of percentages of the accent's total representation) for each accent. This allowed us to more easily compare the LSTM's classifications of emotions between accents with vastly different sample sizes. In cases where differences appear to be smaller, a Kruskal-Wallis test might be utilized for additional statistical clarity.

(Intended) Experiment 2 - Transfer learning on a highly accurate emotion classifier

After training an emotion-classification model using the Crema-D dataset, we noticed that the categorical accuracy of our testing data tended to settle around 60%. Additional changes to the architecture of our model did not seem to improve this value. While 60% accuracy is still greater than the 17% accuracy we would expect from a model that was classifying audio as one of six emotions entirely by chance, we were skeptical about the ability of our model to classify audio files in a manner that allowed us to make accurate conclusions about the Crema-D dataset. This was an issue because one of the only simple ways of improving the model's testing data accuracy—introducing the model to new datasets—could not be performed on account of it preventing us from making generalizations about solely the Crema-D dataset.

To this end, we have proposed a secondary experiment that involves the use of transfer-training to tailor the weights of existing emotion classifiers to Crema-D specifically. By starting with a highly accurate, pre-trained model, we can circumvent the limitations of our single dataset model while allowing us to ensure that our dataset remains tailored to Crema-D. To ensure this, we will run two sets of classifications: the first will use the base, high-accuracy model to generate emotion-based predictions about the Speech Accent Archive data without receiving exposure to Crema-D. The second will utilize transfer learning to tailor the network weights at the top levels of the emotion classifier to the Crema-D dataset. Ultimately, base predictions in this second model and differences between the two prediction series will indicate biases within the Crema-D dataset.

Results

1. Demographic Imbalance in the CREMA-D Dataset

We first examined whether demographic variables (age, sex, and race) were evenly distributed across the six emotion categories in the CREMA-D dataset. These analyses were conducted using a test split.

1.1 Age Distribution

Across all emotions, the average speaker age falls within a range of approximately 35-38 years, indicating that the CREMA-D dataset is dominated by middle-aged speakers.

	Age
Emotion	
ANG	35.785156
DIS	35.961240
FEA	35.396887
HAP	37.472574
NEU	36.346320
SAD	35.688000

There is no significant age variation between the emotion classes. This suggests that age related analyses might not produce useful results, being that most emotions are represented by speakers of nearly identical ages. This raises concerns about whether a model trained on this dataset is capable of generalizing to speakers of other age groups.

1.2 Gender Distribution

Sex	Female	Male
Emotion		
ANG	122	134
DIS	117	141
FEA	130	127
HAP	117	120
NEU	110	121
SAD	110	140

The representation of Gender is not equal, and the imbalance varies by emotion. For example, sadness (SAD) and disgust (DIS) feature significantly more male speakers, while fear (FEA) is slightly female-dominant. These uneven patterns may lead the model to implicitly learn correlations between certain emotions and specific genders, potentially embedding gender bias.

1.3 Race Distribution

Race	African American	Asian	Caucasian	Unknown
Emotion				
ANG	59	24	169	4
DIS	75	13	165	5

FEA	50	18	186	3
HAP	55	20	161	1
NEU	52	23	154	2
SAD	65	20	160	5

CREMA-D is heavily dominated by Caucasian speakers, and the rest of the racial groups are vastly underrepresented. For each emotion category, Caucasian speakers make up approximately 65-75%, African American speakers account for about 20-25%, and Asian speakers only represent 5-10%. This imbalance makes it likely that the model will become better at gauging emotion for some races and generalizing poorly for underrepresented groups, leading to algorithmic racial bias.

2. LSTM Emotion Classification Performance

To train the LSTM model, we extracted MFCC features from CREMA-D WAV files. Training accuracy improved steadily, reaching about 74.7%, and validation accuracy peaked at around 57%.

Epoch	Train Accuracy	Val Accuracy	Train Loss	Val Loss
1	0.697	0.550	0.8827	1.2986
2	0.720	0.575	0.8474	1.2611
3	0.717	0.548	0.8463	1.3893

4	0.722	0.568	0.8473	1.2627
5	0.733	0.555	0.7974	1.3411
6	0.747	0.567	0.7659	1.3457

The model output reveals a gap between training and validation accuracy, which suggests the existence of overfitting. This leads us to believe that although the model might learn the training distribution, it still struggles to generalize. If we recall the demographic imbalance mentioned earlier, we can see how it might be attributed to the reliance on demographic details rather than purely emotional cues, especially if certain racial groups dominate specific emotional categories.

3. Accent-Based Evaluation (In Progress)

   We aim for our project to assess bias in emotion classification based on different accents, but the transfer learning pipeline we are contemplating is still under development, and no accent-based results are ready at this time.

Discussion

The overarching goal of this project was to investigate whether an emotion-recognition model trained on a widely used emotional speech dataset displayed systematic biases when exposed to voices with varying accents. Although the accent-based evaluation is still in progress, the results we have obtained thus far have already revealed several important insights about model behavior, dataset characteristics, and broader implications for fairness and transparency in speech-based AI.

Our analyses of the CREMA-D dataset reveal significant demographic imbalance across race, gender, and age. The data is made up primarily of Caucasian speakers, with most being in their mid-30s. Gender is unevenly represented as well, with specific emotions portrayed by one gender more than the other. This suggests that the model is trained on demographically limited data, and likely causes it to rely on correlations between emotion and demographics, and not necessarily on acoustic patterns relevant to emotion.

The accuracy results of the LSTM reinforce this concern. While the training accuracy reached roughly 74.7%, the validation accuracy didn't notably improve past about 57%. These findings suggest overfitting to the training population. Then, when the model operates based on these demographic correlations as opposed to emotion itself, the results become unreliable. The model's instability can then be identified even before accent variation is introduced, as per our first research question. Since the training data is demographically narrow, the model connects its understanding of emotion to the vocal patterns of the majority group. This strongly implies that when we eventually input speech from actors with unfamiliar accents, the model is likely to misinterpret the speech due to its limited exposure to vocal diversity.

The second research question asks whether bias exists in the classification of basic emotional valence across accents. While our current model has not yet been applied to accented speech, the validation accuracy patterns already hint at a potential valence-specific bias. We've established that the model might associate certain emotion categories with specific demographics, so we can imagine a scenario where a group of unfamiliar accents may be perceived as "negative-sounding", while others may be categorized as neutral or positive. This will be confirmed or contradicted once we evaluate the model on different accents.

Our third research question details whether emotion classification differs based on the actor’s age. While the CREMA-D dataset contains speakers from ages ranging between 20 and 74, the distribution is so tightly clustered around the mid-30s that meaningful age comparison becomes nearly impossible. This severely limits our ability to detect any age-related bias. As a result, the model likely generalizes poorly with younger and older speakers, this time due to a lack of exposure rather than a learned bias.

To an extent, the discrepancies we have identified both within the Crema-D dataset and our resulting findings indicate a need for well-documented, transparent data. Despite advertising itself as a diverse dataset with a demographically diverse set of speakers, our data reveals biases in the dataset towards certain demographics. Ultimately, it is the culmination of these slight biases that requires data scientists to use multiple datasets in order to attain a large set of diverse data to train models. The tendency to create multiple smaller datasets and piece them together into one diverse one indicates a preference for model work over data work, as highlighted by Scheuerman et al. (2021). Overall, however, this preference is dangerous as data is just as important to model outcomes as the architecture models train on and the scientists who curate and implement the data to solve real-world problems. As a result, we echo Scheuerman et al.’s claim and argue that datasets deserve more transparency and attention.

References

Jiang, X., Gossack-Keenan, K., & Pell, M. D. (2019). To believe or not to believe? How voice and accent information in speech alter listener impressions of trust. *Quarterly Journal of Experimental Psychology*, *73*(1). https://journals.sagepub.com/doi/full/10.1177/1747021819865833

Scheuerman, M. K., Hanna, A., & Denton, R. (2021). Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1-37. https://dl.acm.org/doi/10.1145/3476058