Emotional Prosody Analysis: F0 Variation in CREMA-D

Introduction

The emotional state of a speaker modulates speech production, and fundamental frequency (F0) is one of the primary acoustic correlates where this modulation is reflected. The relationship between F0 and emotional expression has been extensively studied and remains an active area of research. Examining how emotion is represented in speech provides insight into human psychology and communication. Concurrently, advances in speech technology focus on accurately modeling emotional prosody in text-to-speech systems and voice assistants to increase the naturalness of synthesized voices. Changes in F0 are associated with emotional intensity across diverse languages, and understanding its variations due to emotional intensity provides a foundation for accurately modeling emotion in such systems.

However, emotion is not the only factor that affects F0, and interactions between emotion and other predictors require further examination. Baseline F0 differs systematically across demographic groups, particularly between biological sexes. These baseline differences are crucial to consider when modeling emotional variation in pitch to control for potential confounding effects. In this study, we conduct a statistical analysis to compare mean F0 values across different emotional categories and between male and female adult speakers. This allows us to examine how **emotion** and **speaker sex** jointly influence vocal pitch in a large, demographically diverse corpus of emotional speech. With a Bayesian regression approach we quantify both the main effects of these factors and their interaction, providing insight into whether emotional modulation of F0 differs between male and female speakers. According to our findings, females consistently exhibit a higher F0 compared to male speakers and also larger pitch increases from a neutral baseline when in high-arousal emotional states, indicating that emotional modulation of F0 differs between sexes.

As a secondary contribution, we present a curated dataset containing demographic information, acoustic features (mean F0, standard deviation, and interquartile range), and metadata for 7,442 emotional speech samples from the CREMA-D corpus. This dataset is publicly available and can be used for future research and pedagogical purposes in the Quantitative Methods for Linguistics and English Language Course.

Background

Numerous studies have demonstrated that F0, as a primary acoustic correlate of emotional expression, varies systematically across distinct emotions: high-arousal emotional states such as anger, happiness, and fear tend to increase mean F0 compared to low-arousal states like disgust, sadness and neutrality (Banse and Scherer 1996). At the same time, biological sex is a well-established factor that significantly influences baseline F0: adult females typically exhibit higher mean F0 than males due to anatomical differences in the size of vocal folds (Benesty, Sondhi, and Huang 2008).

While these baseline sex differences are robust and well-documented, comparatively fewer studies systematically examine the **interaction** between sex and emotional modulation of F0. Some research indicates that although both sexes show similar trends in F0 changes with emotion, the magnitude of these changes may vary. Females, for example, exhibit larger fluctuations in pitch in intense emotional speech (Scherer 2003).

Most research investigates how emotional states affect F0 across sexes, but little work has systematically examined whether emotional modulation differs between males and females. Findings remain inconclusive and inconsistent. Specifically, some studies report significant results regarding sex differences in emotional modulation (Banse and Scherer 1996), whereas others report no such effects after appropriate controls (Viscovich 2003). Moreover, existing research is often limited by small sample sizes ((Viscovich 2003), (Banse and Scherer 1996), (Scherer 2003)). As a result, there are still open questions about how biological sex shapes emotional prosody.

Data

Corpus and Speakers

For this research, CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset) was used. CREMA-D is a publicly corpus designed for emotion recognition research, consisting of 7,442 samples of audio from 91 professional actors from a range of demographic backgrounds. These actors were asked to repeat 12 sentences across a range of emotions and emotional intensities. The sentences were designed to be emotionally-neutral in their linguistic content, allowing for emotional information to be primarily conveyed through prosody, rather than semantics.

The 12 sentences are represented in (Table 1), as follows:

Table 1: CREMA-D sentence codes with sentences

Code	Sentence
DFA	Don’t forget a jacket
IEO	It’s eleven o’clock

IOM	I'm on my way to the meeting
ITH	I think I have a doctor's appointment
ITS	I think I've seen this before
IWL	I would like a new alarm clock
IWW	I wonder what this is about
MTI	Maybe tomorrow it will be cold
TAI	The airplane is almost full
TIE	That is exactly what happened
TSI	The surface is slick
WSI	We'll stop in a couple of minutes

Each sentence was repeated by each individual with the intent of conveying six different emotions (anger, disgust, fear, happiness, sadness, and neutrality) across four intensity levels (low, medium, high, and unspecified). Consequently, this dataset contains both subtle and exaggerated versions of each expression, for each emotion type.

Audio files were recorded at a 16 kHz sampling rate, in WAV format. Over 95% of clips possess 7 independent ratings from human annotators; the audio-only subsection of the dataset (used for this research) possesses an overall human recognition accuracy of 40.9%, reflecting substantial variability and ambiguity in emotional expression across the data.

The corpus includes speakers from a range of demographic backgrounds (Figure 1), making it well-suited for studying how emotional prosody interacts with speaker demographics.

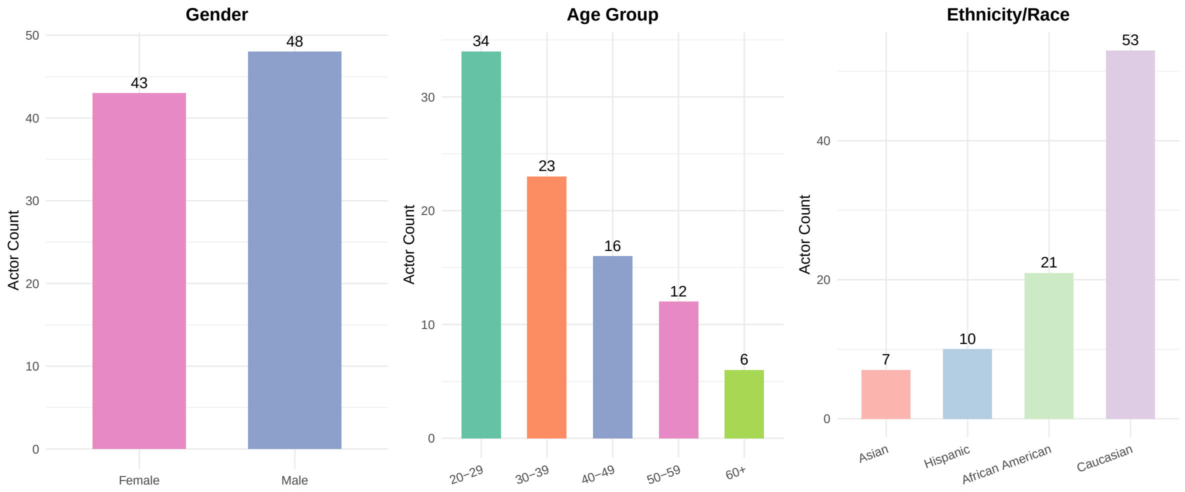

Figure 1: Demographic distribution of actors in CREMA-D corpus

Acoustic Features

F0 represents the rate of vocal fold vibration, and is a key acoustic correlate of emotional prosody. Here, F0 was extracted using CREPE (Convolutional Representation for Pitch Estimation) (Kim et al. (2018)), a neural network-based pitch tracking system known to perform well on noisy audio data. For each recording, four summary statistics were recorded (Table 2):

Table 2: Extracted F0 features for statistical analysis

Feature	Description	Emotional.Relevance
Mean F0 (Hz)	Average pitch level	High-arousal emotions raise pitch; sadness lowers it
Standard Deviation (Hz)	Pitch variability	Greater variability indicates higher arousal
Interquartile Range (Hz)	Robust measure of pitch spread	Captures expressiveness while handling outliers
Mean Confidence	Extraction quality indicator	Used for quality control

The chosen features capture the central tendency and the variability of pitch — both qualities are theoretically-motivated correlates of emotional state within speech (Juslin and Laukka 2003; Banse and Scherer 1996).

Methodology

Data Preparation and Transformation

F0 is positively skewed, as pitch cannot be negative. Furthermore, in humans, pitch is perceived non-linearly, with cochlear filter-banks operating on a logarithmic scale. Therefore log(meanF0) values were calculated and used during exploratory data analysis.

Exploratory Analysis

Before formal statistical modeling, the distribution of log-transformed F0 was examined, across emotions and demographic factors, to identify potential patterns and assess data quality.

Distribution of F0 by Emotion

Figure 2 displays the overall distribution of log F0 across the emotion categories. The distributions possess substantial overlap, reflecting substantial variability in natural pitch variability emotional expression present in CREMA-D.

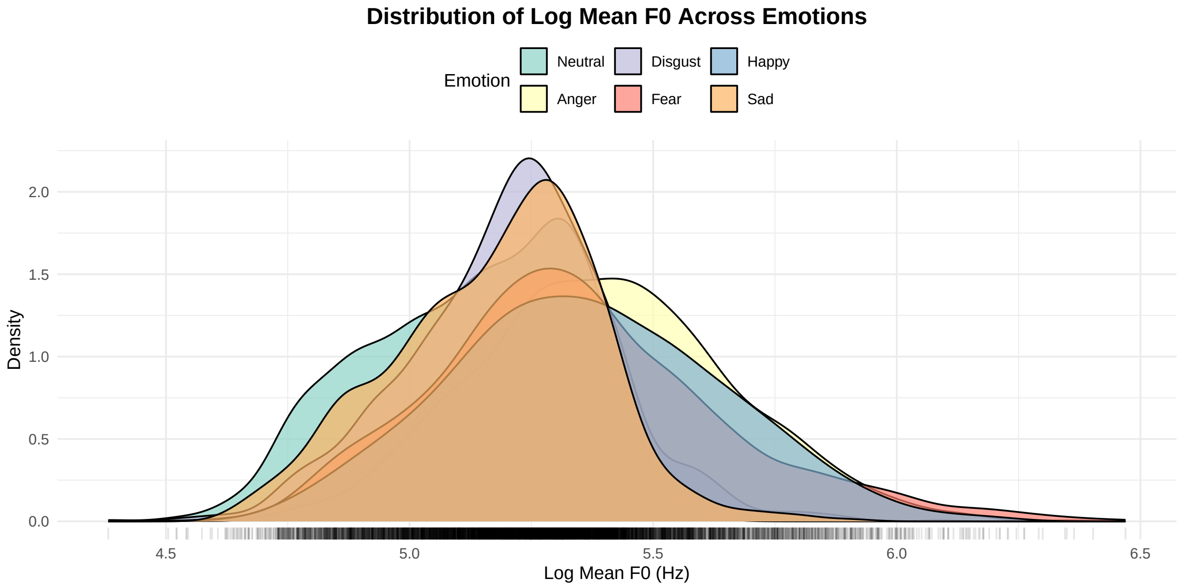

Figure 2: Density distributions of log-transformed mean F0 by emotion

Figure 3 provides a complementary view with violin plots and boxplots, showing differences in central tendency and spread across emotions.

Interaction of Emotion and Gender

Given documented sex differences in F0 due to morphological (vocal tract) differences, this study aimed to examine how emotional modulation of pitch varies by speaker sex. Figure 4 observes that while males and females differ substantially in baseline F0, pitch modulation by emotion appear to follow broadly similar trends across sexes.

Demographic Factors

We also examined F0 variation across age groups and racial/ethnic backgrounds to assess whether these factors should be considered in statistical modeling.

While age and race show some variation in F0, our primary research question focuses on how **emotion** and **sex** jointly influence pitch production. Therefore the statistical model presented in this research centres on these two factors.

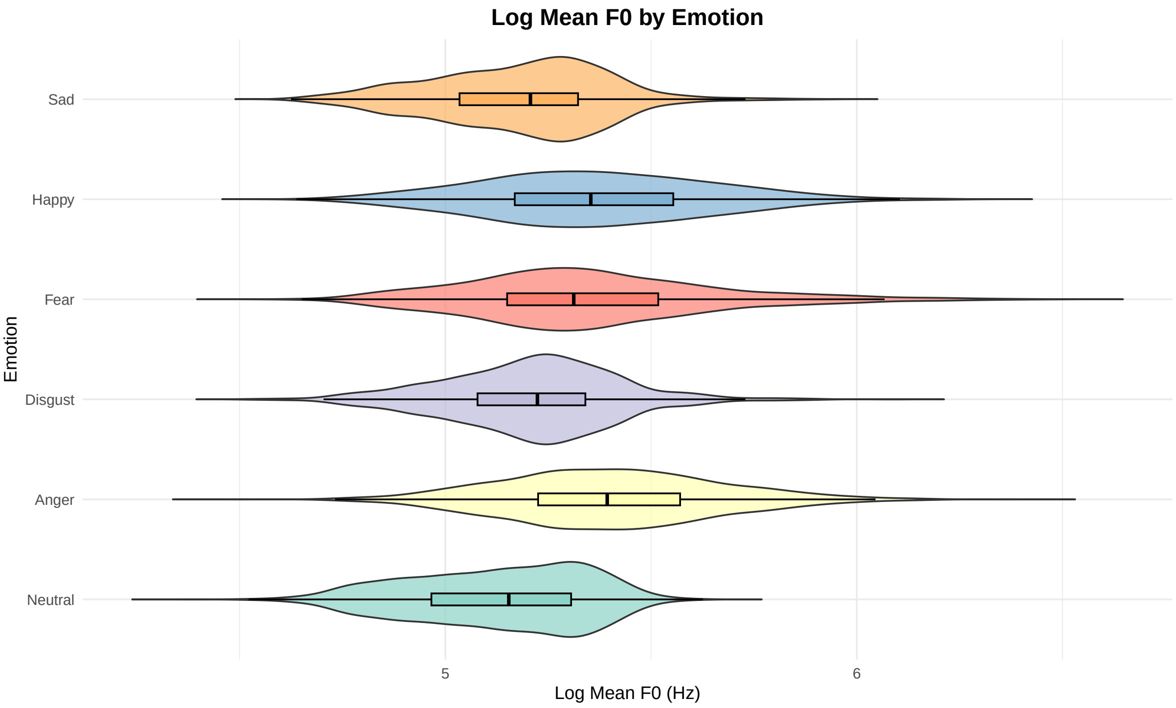

Figure 3: Distribution of log-transformed mean F0 across emotions

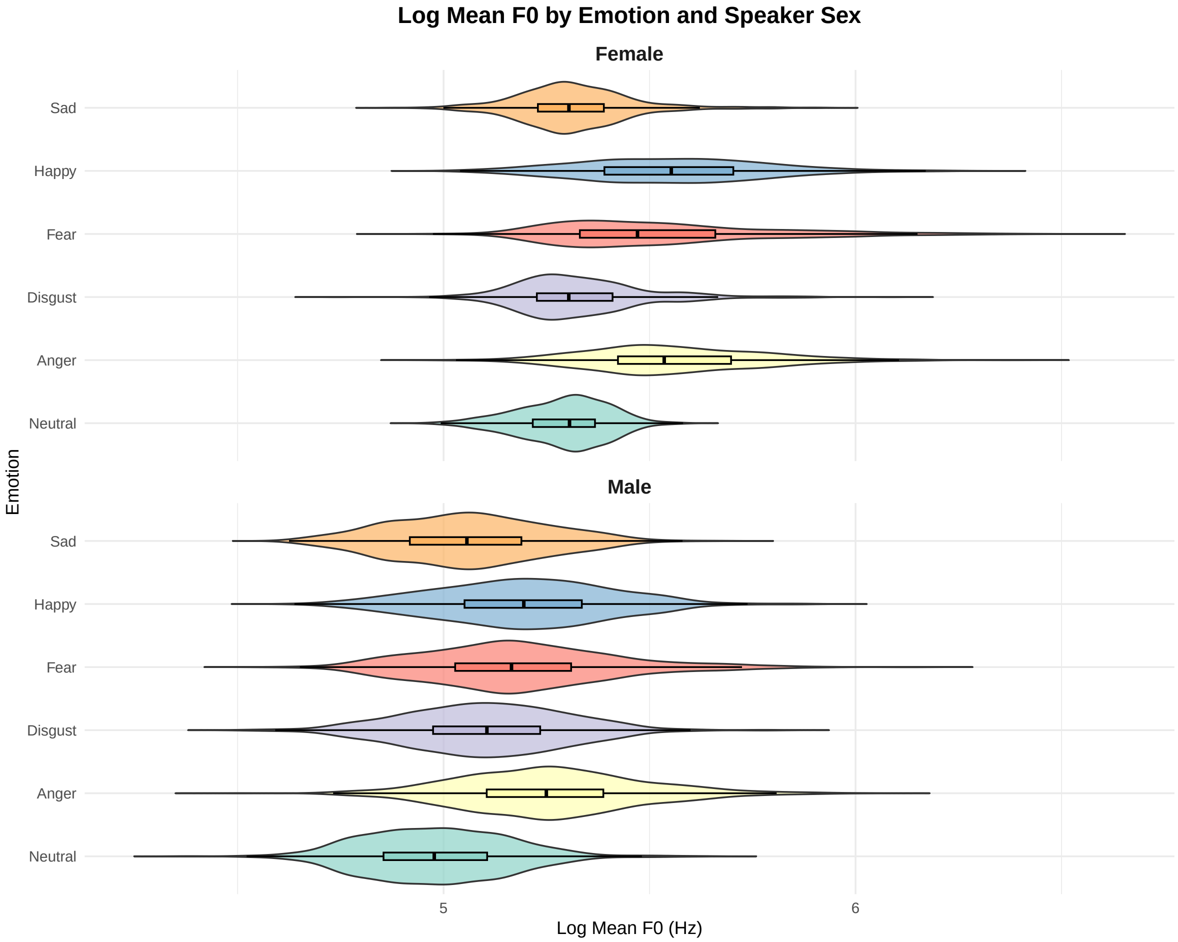

Figure 4: Distribution of log-transformed mean F0 across emotions, separated by speaker sex

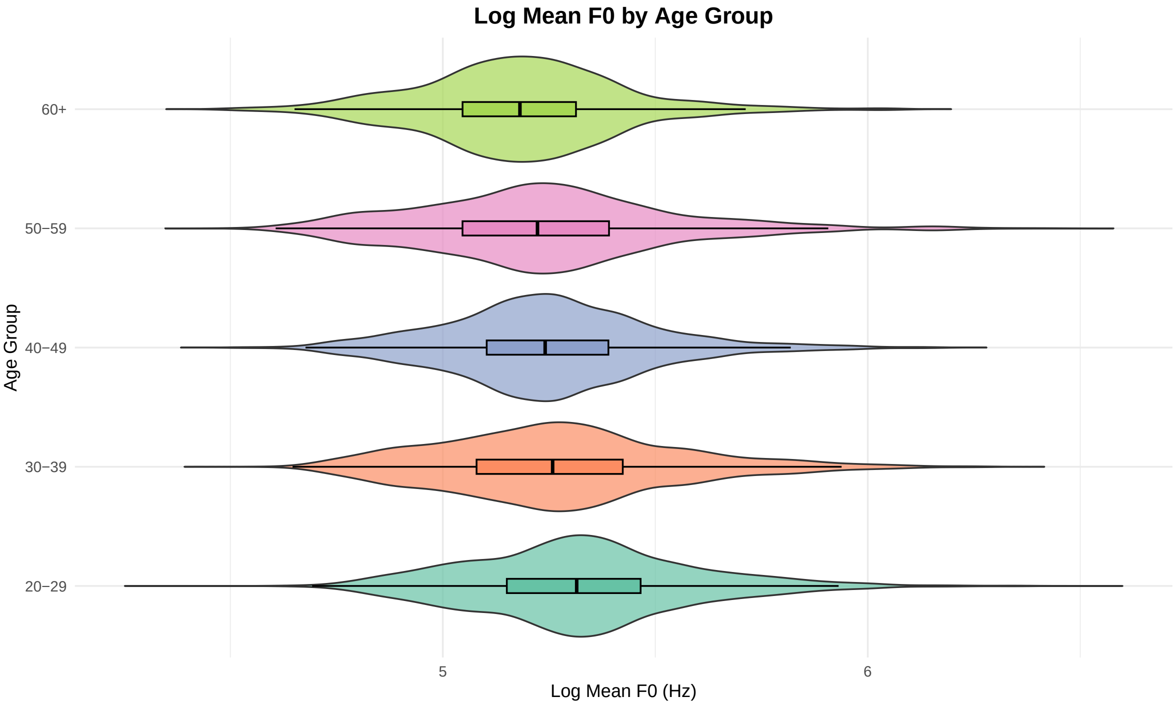

Figure 5: Distribution of log-transformed mean F0 across age groups

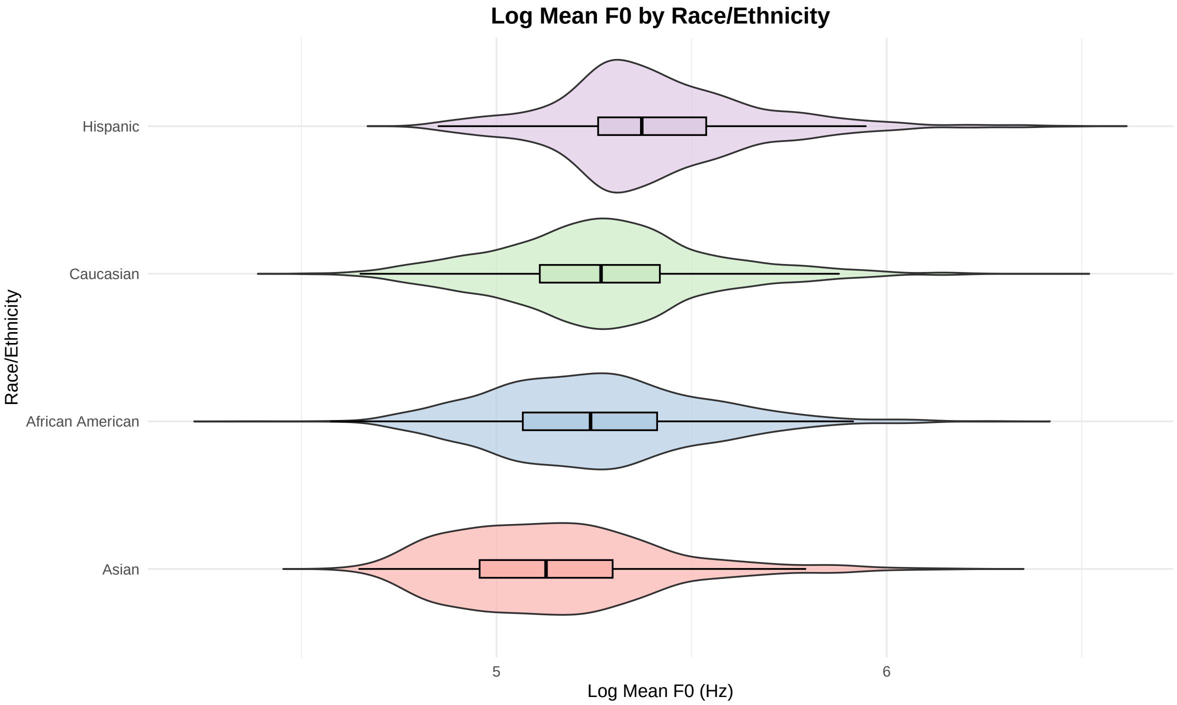

Figure 6: Distribution of log-transformed mean F0 across racial/ethnic groups

Statistical Modeling

Model Specification

For statistical analysis, a Bayesian linear regression model was implemented through the `brms` package (Bürkner 2017), which utilises Stan (Carpenter et al. 2017), a probabilistic programming language for statistical models, as a back-end to produce posterior probabilities.

The following model formula was used:

```
meanF0 ~ Emotion * Sex
```

This formula specifies mean F0 as the outcome variable, with Emotion and Sex as categorical predictors. It is equivalent to `meanF0 ~ Emotion + Sex + Emotion:Sex`, modelling the effect of both predictors and their interaction — thereby allowing for investigation of the degree of sex's impact on the effect of emotional modulation on pitch.

Emotion was coded to capture F0 differences relative to a **Neutral** reference level. Sex was coded to test the effect of emotion of F0 relative to **Female** speakers. Both predictors were coded using the default R treatment contrasts.

Distributional Assumptions

A **log-normal distribution** was assumed for the model family, meaning that F0 is modeled as following a lognormal distribution. This is equivalent to assuming that log(F0) follows a normal distribution. During model fitting, coefficient values were identified that maximized the probability of the observed data under the assumed model.

```r
# Fit the Bayesian linear regression model
f0_bm <- brm(
  meanF0 ~ Emotion * Sex,
  data    = crema_data,
  family  = lognormal,
  cores = 4,
  seed = 6725,
  file = "ch_regression_f0_bm_lognormal2"
)
```

Regression Coefficient Summary

The Bayesian regression model estimated coefficients on the log(F0) scale. Table 3 presents the posterior means and credible intervals for all model parameters. The **Intercept** represents the baseline log(F0) for female speakers producing neutral speech (the reference levels). **Emotion** coefficients indicate log-scale shifts relative to neutral. **Sex** coefficients capture the male-female baseline difference. **Interaction terms** (Emotion:Sex) indicate how emotional modulation differs between sexes.

Table 3: Model coefficient summary with 80% and 95% credible intervals

Parameter	Estimate	80% CI Lower	80% CI Upper	95% CI Lower	95% CI Upper
Intercept	5.291	5.280	5.302	5.273	5.308
EmotionAnger	0.269	0.254	0.284	0.246	0.293
EmotionDisgust	0.036	0.021	0.051	0.012	0.058
EmotionFear	0.233	0.218	0.249	0.210	0.256
EmotionHappy	0.261	0.246	0.277	0.237	0.285
EmotionSad	0.024	0.008	0.039	0.001	0.047
SexMale	-0.309	-0.325	-0.294	-0.332	-0.286
EmotionAnger:SexMale	0.006	-0.015	0.028	-0.025	0.038
EmotionDisgust:SexMale	0.086	0.065	0.107	0.054	0.118
EmotionFear:SexMale	-0.033	-0.054	-0.011	-0.064	-0.001
EmotionHappy:SexMale	-0.048	-0.068	-0.027	-0.079	-0.016
EmotionSad:SexMale	0.053	0.032	0.074	0.021	0.084

These coefficients are interpretable as additive effects on the log scale. For example, a coefficient of 0.10 corresponds to an approximate 10% increase in F0 when exponentiated. The predictions in Hz are obtained by exponentiating the linear combinations of these log-scale coefficients.

Results

We transform the model predictions from the log scale to Hz for substantive interpretation, examining predicted F0 values across all Emotion × Sex combinations.

Predicted F0 Values on the Hz Scale

We obtained posterior predictions of expected F0 for each Emotion × Sex combination in Hz, then computed emotion effects (differences from Neutral) and sex differences. All results below present posterior means with 80% and 95% credible intervals.

Condition Means

Table 4 presents the predicted mean F0 for each Emotion × Sex combination. Female speakers exhibited higher F0 than males across all emotions, consistent with anatomical differences in vocal fold morphology.

Table 4: Predicted mean F0 (Hz) by emotion and sex with 80% and 95% credible intervals

Sex	Emotion	Mean F0	80% CI Lower	80% CI Upper	95% CI Lower	95% CI Upper
Female	Anger	259.8	257.1	262.4	255.7	263.9
Female	Disgust	205.7	203.6	207.8	202.3	209.1
Female	Fear	250.7	248.1	253.3	246.7	254.7
Female	Happy	257.7	255.0	260.4	253.8	261.9
Female	Neutral	198.5	196.3	200.7	195.1	201.9
Female	Sad	203.3	201.2	205.4	200.1	206.6
Male	Anger	191.9	190.0	193.9	189.0	194.9
Male	Disgust	164.7	163.1	166.2	162.2	167.1
Male	Fear	178.1	176.4	179.8	175.4	180.8
Male	Happy	180.4	178.6	182.2	177.6	183.1
Male	Neutral	145.7	144.2	147.3	143.4	148.1
Male	Sad	157.3	155.8	158.9	155.0	159.7

Emotion Effects Within Sex

Table 5 shows how each emotion modulates F0 relative to Neutral speech, separately for each sex. Positive values indicate F0 increases; negative values indicate decreases. The credible intervals quantify the uncertainty in these effect estimates.

Table 5: Emotion effects on F0 (Hz)

Sex	Emotion	Mean Difference	80% CI Lower	80% CI Upper	95% CI Lower	95% CI Upper
Female	Anger	61.3	57.8	64.7	56.0	66.6
Female	Disgust	7.2	4.1	10.3	2.5	11.8
Female	Fear	52.2	48.7	55.6	46.9	57.2
Female	Happy	59.2	55.7	62.7	53.7	64.6
Female	Neutral	0.0	0.0	0.0	0.0	0.0
Female	Sad	4.8	1.7	7.8	0.2	9.5
Male	Anger	46.2	43.8	48.7	42.5	50.0
Male	Disgust	18.9	16.7	21.2	15.5	22.3
Male	Fear	32.4	30.1	34.7	28.8	35.9
Male	Happy	34.7	32.3	36.9	31.0	38.2
Male	Neutral	0.0	0.0	0.0	0.0	0.0
Male	Sad	11.6	9.4	13.7	8.2	14.9

Figure 7 visualizes these emotion effects, revealing the pattern of emotional modulation for male and female speakers. Parallel lines would indicate similar emotional modulation across sexes, while diverging patterns suggest sex-specific differences in emotional expression.

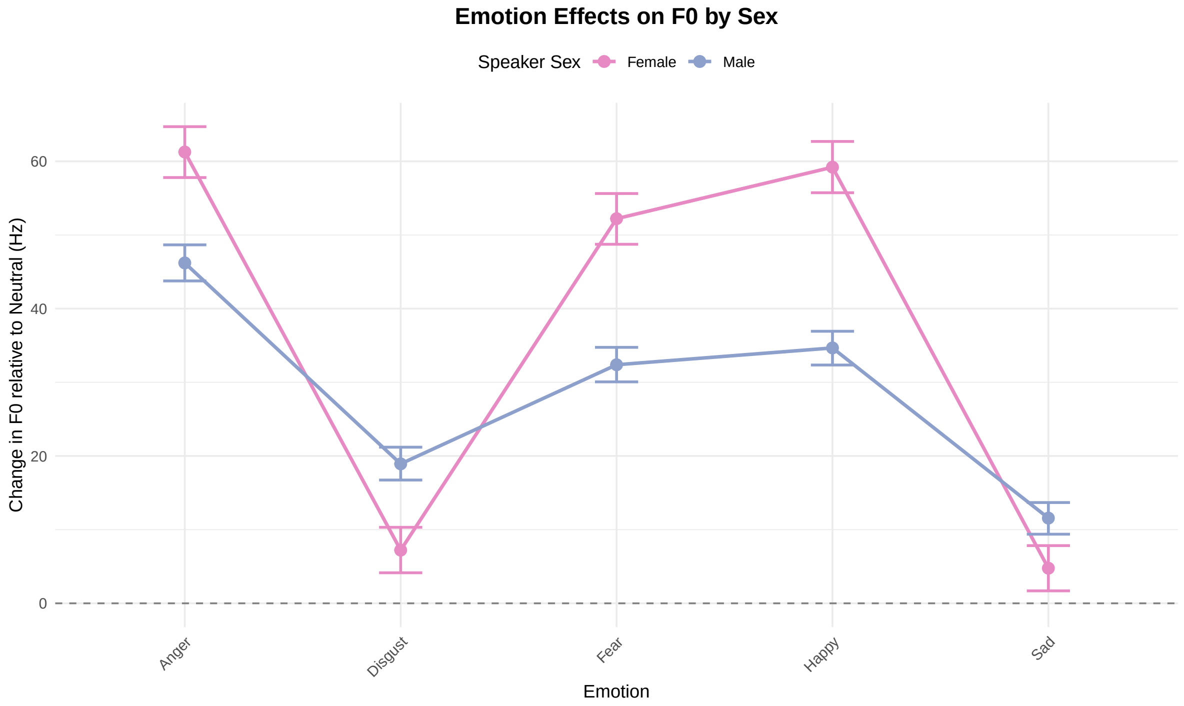

Figure 7: Interaction between emotion and sex: F0 change relative to Neutral. Error bars represent 80% credible intervals.

Sex Differences Across Emotions

Table 6 presents the male-female F0 difference for each emotion. Negative values indicate females have higher F0 than males. These results show whether the baseline sex difference varies across emotional contexts.

Table 6: Sex differences in F0 (Hz; Male − Female) by emotion, with 80% and 95% credible intervals

Emotion	Mean Difference	80% CI Lower	80% CI Upper	95% CI Lower	95% CI Upper
Anger	-67.8	-71.1	-64.6	-72.9	-62.8
Disgust	-41.1	-43.8	-38.4	-45.2	-37.0
Fear	-72.6	-75.7	-69.5	-77.4	-67.9
Happy	-77.3	-80.5	-74.2	-82.1	-72.4
Neutral	-52.8	-55.6	-50.1	-56.8	-48.8
Sad	-46.0	-48.6	-43.4	-50.1	-42.0

Posterior Distributions of Mean F0

Figure 8 displays the posterior distributions of mean F0 for each emotion-sex combination. The distributions represent the uncertainty about the true mean F0 for each condition, with wider distributions indicating greater uncertainty. The figure reveals both the substantial baseline difference between male and female speakers, and patterns of emotional modulation within each sex.

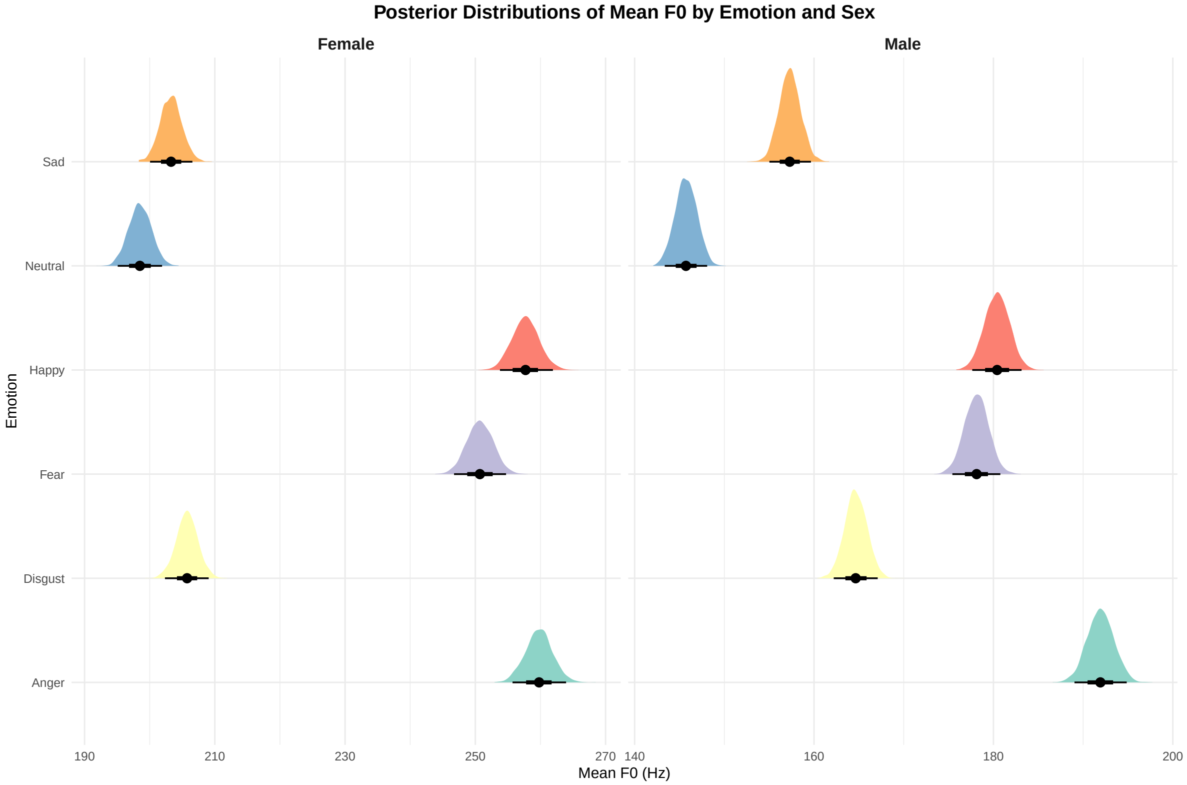

Figure 8: Posterior distributions of mean F0 by emotion and speaker sex

Discussion

All emotions caused mean F0 to increase across speakers, but the magnitude of this effect varied between emotions. The pattern observed is consistent with previously-mentioned arousal-based theories of emotional prosody, where studies found that speech produced in high-arousal emotions (e.g. anger, fear, and happiness) produce more substantial F0 increases relative to neutral speech than low-arousal emotions (e.g. disgust and sadness). Reflecting this, large pitch increases were observed for high-arousal emotions: anger caused F0 to rise by ~61 Hz in females (95% CI [56, 67]), and ~46 Hz in males (95% CI [43, 50]); fear by ~52 Hz in females

(95% CI [47, 57]), and ~32 Hz in males (95% CI [29, 36]); and happiness by ~59 Hz in females (95% CI [54, 65]), and ~35 Hz in males (95% CI [31, 38]). Smaller changes were found for low-arousal emotions: disgust caused F0 to rise by ~7 Hz in females (95% CI [3, 12]), and ~19 Hz in males (95% CI [16, 22]); and sadness by 5 Hz in females (95% CI [0, 10]), and 12 Hz in males (95% CI [8, 15]).

Males were found to possess substantially lower average F0 values than females when speaking neutrally, with a baseline difference of ~53 Hz (95% CI [49, 56]), due to anatomical differences between males and females. However, the magnitude of this gap varied across emotions, reflecting sex-dependent differences in pitch variation across emotional states. For high-arousal emotions, females displayed larger pitch increases than males: for anger, female F0 rose by ~15 Hz more than for males; for fear female F0 rose by ~20 Hz more than in males; and for happiness, ~25 Hz more than males. Inspecting the interaction term confirms the observed differences for fear (80% CI [-0.054, -0.012]; 95% CI [-0.065, -0.002]) and happiness (80% CI [-0.068, -0.027]; 95% CI [-0.08, -0.016]). For anger, no reliable difference was found at either an 80% confidence interval ([-0.014, 0.027]), or at 95% interval ([-0.026, 0.04]).

For low-arousal emotions, this trend reversed, with males displaying larger pitch increases. For disgust, male pitch rose by ~12 Hz more than females; and for sadness, pitch rose by ~7 Hz more than for females. Interaction terms confirm sex-dependent differences in pitch modulation for disgust (80% CI [0.065, 0.107]; 95% CI [0.054, 0.119]) and sadness (80% CI [0.032, 0.073]; 95% CI [0.021, 0.085]).

Due to the study's constraints, including limited control over confounding factors affecting F0 (described in the following section), the observed associations should not be interpreted as evidence of causality. To draw definitive conclusions, further investigation is required.

Limitations and future work

The dataset used consists of posed emotional speech produced by actors trained to exaggerate and therefore the results cannot be generalized to spontaneous, natural speech.

Additionally, it is important to mention that examining the intensity of each emotion lies beyond the scope of this study, so the results are only a comparison between emotional states and not between intensities of each emotion. Because intensity can modulate F0 independently of emotion category, examining graded emotional levels represents an important direction for future research.

Furthermore, this study is restricted to mean F0 as means of emotional expression, although there are more acoustic parameters that interact with F0 to convey emotion in speech, such as timbre, voice quality and formant frequencies. Incorporating these features would allow for a more complete characterization of emotional prosody.

Finally, only biological sex was considered as a demographic factor. Other sources of speaker variability, including age, social background and hormonal state may modulate emotional expression. Future work should systematically examine how these factors interact to shape F0 patterns in emotional speech.

References

Banse, Rainer, and Klaus R Scherer. 1996. "Acoustic Profiles in Vocal Emotion Expression." *Journal of Personality and Social Psychology* 70 (3): 614–36.

Benesty, Jacob, M. Mohan Sondhi, and Yiteng Arden Huang, eds. 2008. *Springer Handbook of Speech Processing*. Springer Handbooks. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-49127-9.

Bürkner, Paul-Christian. 2017. "Brms: An r Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software* 80 (1): 1–28. https://doi.org/10.18637/jss.v080.i01.

Carpenter, Bob, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76.

Juslin, Patrik N, and Petri Laukka. 2003. "Communication of Emotions in Vocal Expression and Music Performance: Different Channels, Same Code?" *Psychological Bulletin* 129 (5): 770–814.

Kim, Jong Wook, Justin Salamon, Peter Li, and Juan Pablo Bello. 2018. "CREPE: A Convolutional Representation for Pitch Estimation." In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 161–65. IEEE.

Scherer, Klaus R. 2003. "Vocal Communication of Emotion: A Review of Research Paradigms." *Speech Communication* 40 (1-2): 227–56.

Viscovich, Nada. 2003. "Acoustic Analysis of Vocal Emotion in Speech." *[Journal Details Needed]*.