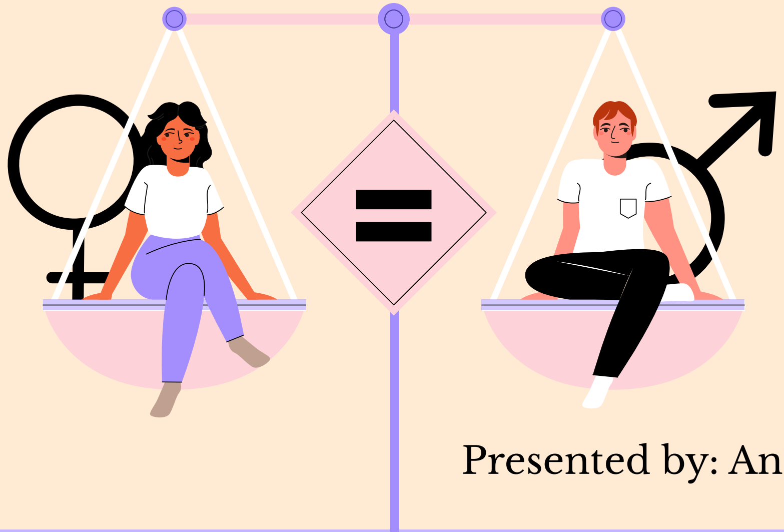
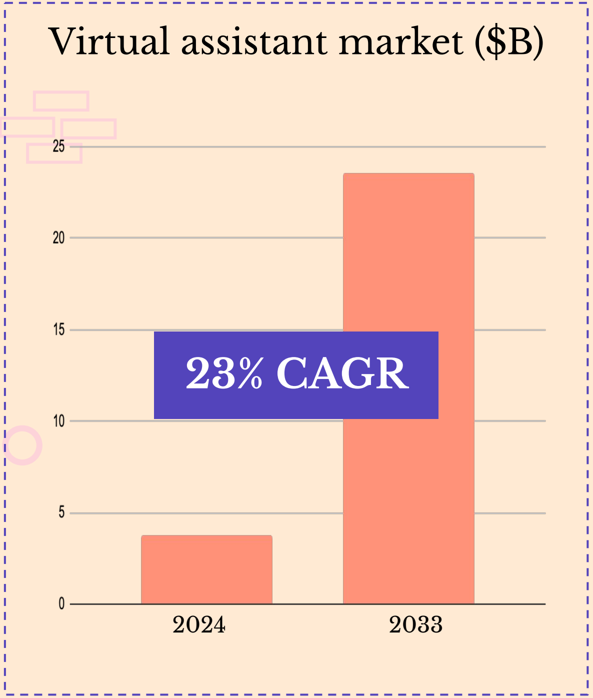
HE SAID, SHE SAID

A Gendered Twist on Virtual Assistants



Presented by: Angel Mary Oviya, Rishabh Shukla, Shubhi Phartiyal

CURRENT STATE



Current Issue: One-size-fits-all models lack personalization, leading to disengaged user experiences

Opportunity: Businesses can offer gender-sensitive, adaptive virtual assistants for improved user engagement, brand loyalty, and market positioning

Project Aim: Make corporate interactions great again!

Key Benefits
1) Creates more empathetic, relatable interactions
2) Paves the way for AI that adapts to age, personality, culture, etc.
3) Builds ethical AI that respects user individuality

REVIEWING OUR DATASET

Vocal Gender Features Dataset

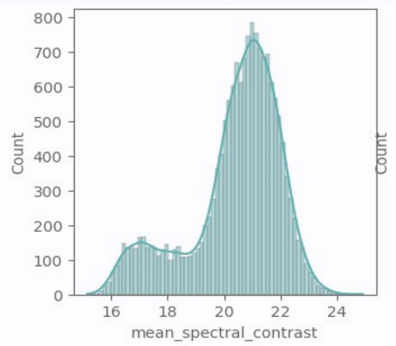
1.	**SPECTRAL**	Describes the "texture" of the sound (eg. sharp, smooth)
2.	**PITCH**	Measures how high or low the voice is
3.	**ENERGY**	Analyzes loudness and noisiness of the voice
4.	**FREQUENCY**	Captures the unique quality or tone of the voice
5.	**COMPLEXITY**	Contains unpredictability of voice pattern

Sample Size: ♀ = 5768 ♂ = 10380

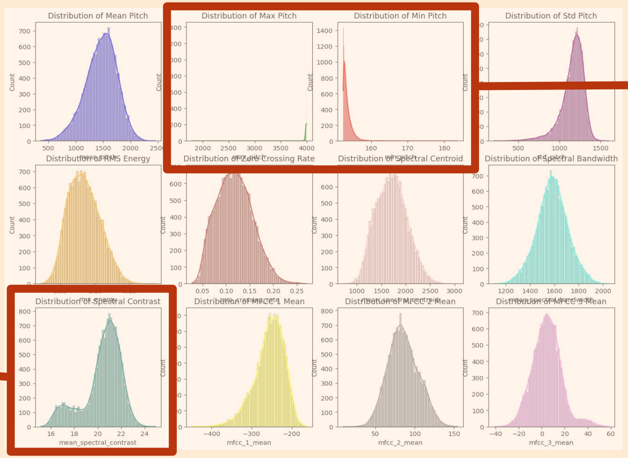
FEATURE EXPLORATION

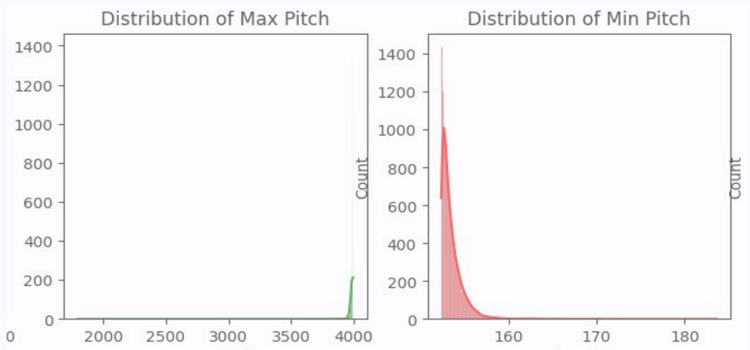
- From the 42 available features, we picked relevant ones based on descriptive statistics (mean, mode, median, range, etc) and qualitative knowledge

- Then we plotted histograms for each feature and observed key details



Clear Bimodality

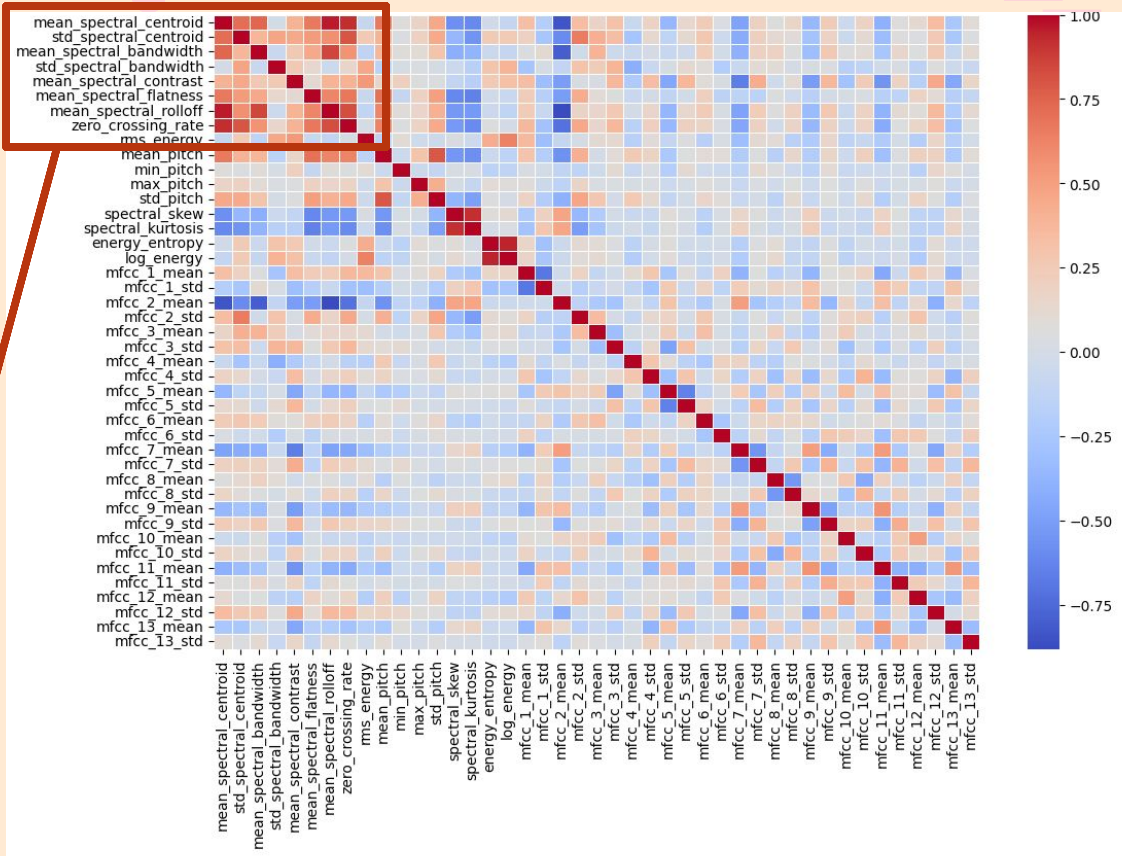




Clear Skew

CORRELATION HEATMAP

- We had 42 features which added a risk of overfitting and complexity
- Features with a correlation above 0.9 were removed to reduce redundancy
- We removed 4 features from our data analysis



FEATURE SELECTION

Feature importance ranking from Decision Tree and Random Forest, aggregated into a combined weighted average

Decision Tree	&	Random Forest	=	Combined Weighted Average
mfcc_5_mean		mfcc_5_mean		mfcc_5_mean
mfcc_1_mean		mean_spectral_contrast		mean_spectral_contrast
rms_energy		mfcc_3_std		mfcc_2_mean
mean_spectral_contrast		mfcc_2_mean		rms_energy
mfcc_2_mean		rms_energy		mfcc_3_std

Top 20 features were selected based on the combined ranking.

REDUCING FEATURE COMPLEXITY

Combined Feature Importance: Selected top 20 features using feature importance scores where higher importance scores are better gender distinguishers

Recursive Feature Elimination (RFE): Selected the best subset of 10 features, to reduce model complexity and speed up training.

Domain Knowledge Inclusion: We included 2 pitch-based features (mean_pitch & std_pitch) based on domain expertise and qualitative research

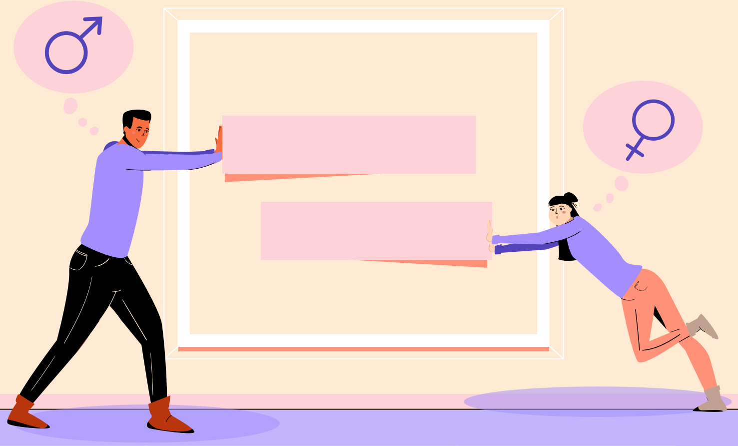
12 Selected Features				
SPECTRAL	**PITCH**	**ENERGY**	**FREQUENCY**	**COMPLEXITY**
1	2	1	8	0

MODEL TRAINING & EVALUATION

1. **Train-Test Split:** The dataset was split into 70% training and 30% testing to assess model generalization on unseen data

2. **StandardScaler:** Standardize features with larger values so they don't disproportionately affect the model, improving model stability and performance

3. **Class Imbalance:** Female voices, make up only 33% of the dataset. Using SMOTE, we addressed the imbalance by creating new synthetic samples



MODEL PERFORMANCE METRICS

General Performance Metrics

- **Accuracy** — Good starting point but misleading since classes are imbalanced
- **Macro Average** — Gives equal weight to both classes, so it ensures model performs well on both classes, regardless of their frequency
- **Weighted Average** — Accounts for class imbalance by weighting larger class *(male)* more

Model Sensitivity Metrics

- **Precision** — Measures how many of the predicted gender classes are correct
- **Recall** — Measures how many of actual class were correctly predicted
- **F1-Score** — Single metric that balances the trade-off between precision and recall

MODEL PERFORMANCE COMPARISON

(%)		LOGISTIC REGRESSION	SVM	RANDOM FOREST	LSTM	CNN
WEIGHTED AVERAGE		95	98	97	98	96
PRECISION	♀	91	98	97	97	94
	♂	97	98	97	98	97
RECALL	♀	94	98	95	96	95
	♂	95	99	98	98	97
F1-SCORE	♀	93	98	96	97	95
	♂	96	99	98	98	97

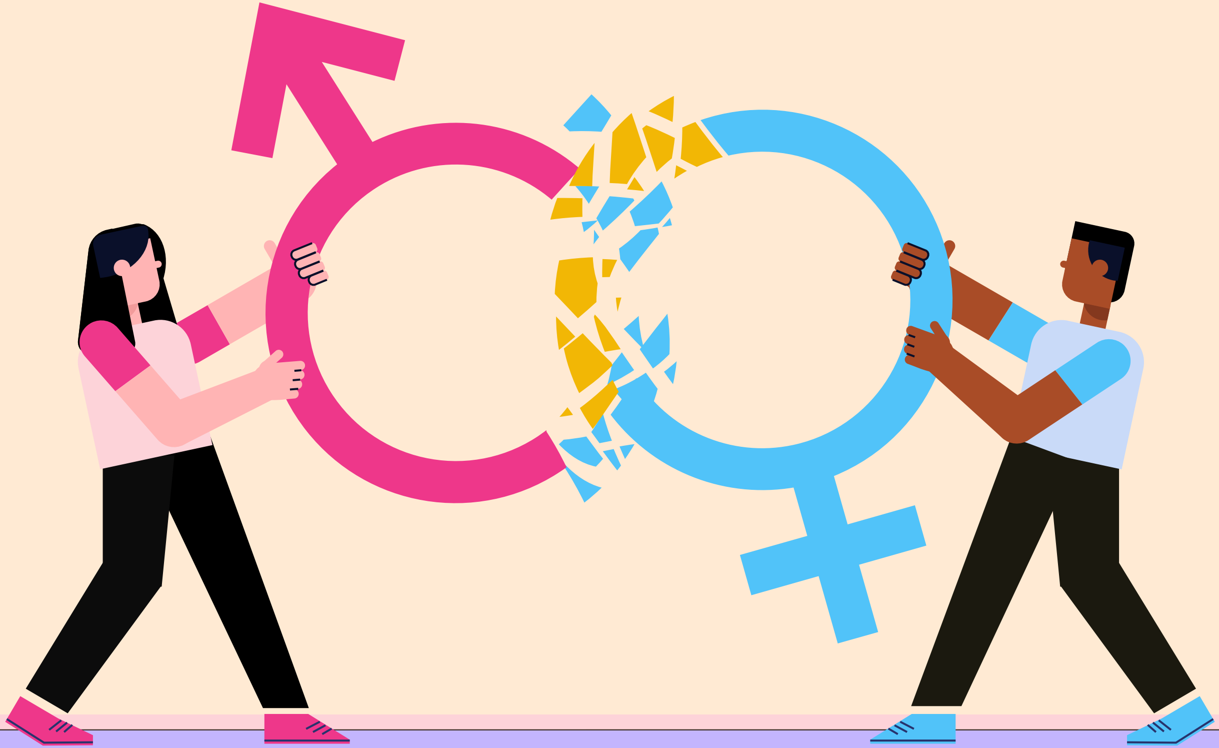
♀= Female ♂= Male

KEY FINDINGS:

- SVM is the best choice for frequency-based voice features.
- Deep learning models need more data & raw spectrograms.

TESTING ON REAL AUDIO

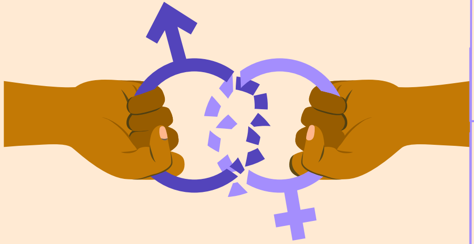
Testing Success: Recorded a few voices and fed into our trained model with 66.6% success



KEY LIMITATIONS & MITIGATIONS

LIMITATIONS		MITIGATIONS
Binary classification of gender	1	Explore spectrum-based classification or unsupervised learning for diversity
Struggles with accents, speech speeds, and background noise	2	Train on multilingual datasets, do data augmentation, and fine-tune for real-world robustness.
Skewed dataset favors Male	3	Diversify data collection and use more balancing techniques.

FUTURE POTENTIAL



1

EMOTIONAL COMPREHENSION

The current model detects gender only, but voice carries emotion, tone, intent and so much more.

2

REAL TIME DEPLOYMENT

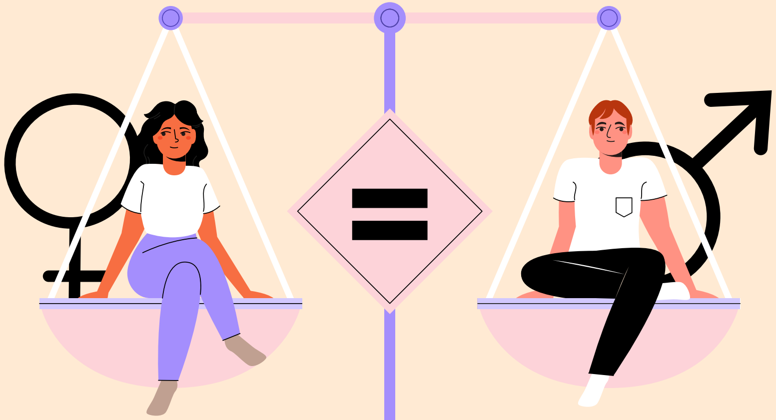
Implement this as a real-time voice assistant feature that adapts responses based on gender tone & emotional state.

3

MULTI-LANGUAGE SUPPORT

Extend training to multi-language datasets for broader applicability

APPENDIX



42 Features in Dataset

mean_spectral_centroid: The average spectral centroid, representing the "center of mass" of the spectrum, indicating brightness.
std_spectral_centroid: The standard deviation of the spectral centroid, measuring variability in brightness.
mean_spectral_bandwidth: The average width of the spectrum, reflecting how spread out the frequencies are.
std_spectral_bandwidth: The standard deviation of spectral bandwidth, indicating variability in frequency spread.
mean_spectral_contrast: The average difference between peaks and valleys in the spectrum, indicating tonal contrast.
mean_spectral_flatness: The average flatness of the spectrum, measuring the noisiness of the signal.
mean_spectral_rolloff: The average frequency below which a specified % of the spectral energy resides, indicating sharpness.
zero_crossing_rate: The rate at which the signal crosses the zero amplitude axis, representing noisiness or percussiveness.
rms_energy: The root mean square energy of the signal, reflecting its loudness.
mean_pitch: The average pitch frequency of the audio.
min_pitch: The minimum pitch frequency.
max_pitch: The maximum pitch frequency.
std_pitch: The standard deviation of pitch frequency, measuring variability in pitch.
spectral_skew: The skewness of the spectral distribution, indicating asymmetry.
spectral_kurtosis: The kurtosis of the spectral distribution, indicating the peakiness of the spectrum.
energy_entropy: The entropy of the signal energy, representing its randomness.
log_energy: The logarithmic energy of the signal, a compressed representation of energy.
mfcc_1_mean to mfcc_13_mean: The mean of the first 13 Mel Frequency Cepstral Coefficients (MFCCs), representing the timbral characteristics of the audio.
mfcc_1_std to mfcc_13_std: The standard deviation of the first 13 MFCCs, indicating variability in timbral features.

Model Selection Rationale

Logistic Regression: Efficient for binary classification with linear relationships, acting as a baseline model.

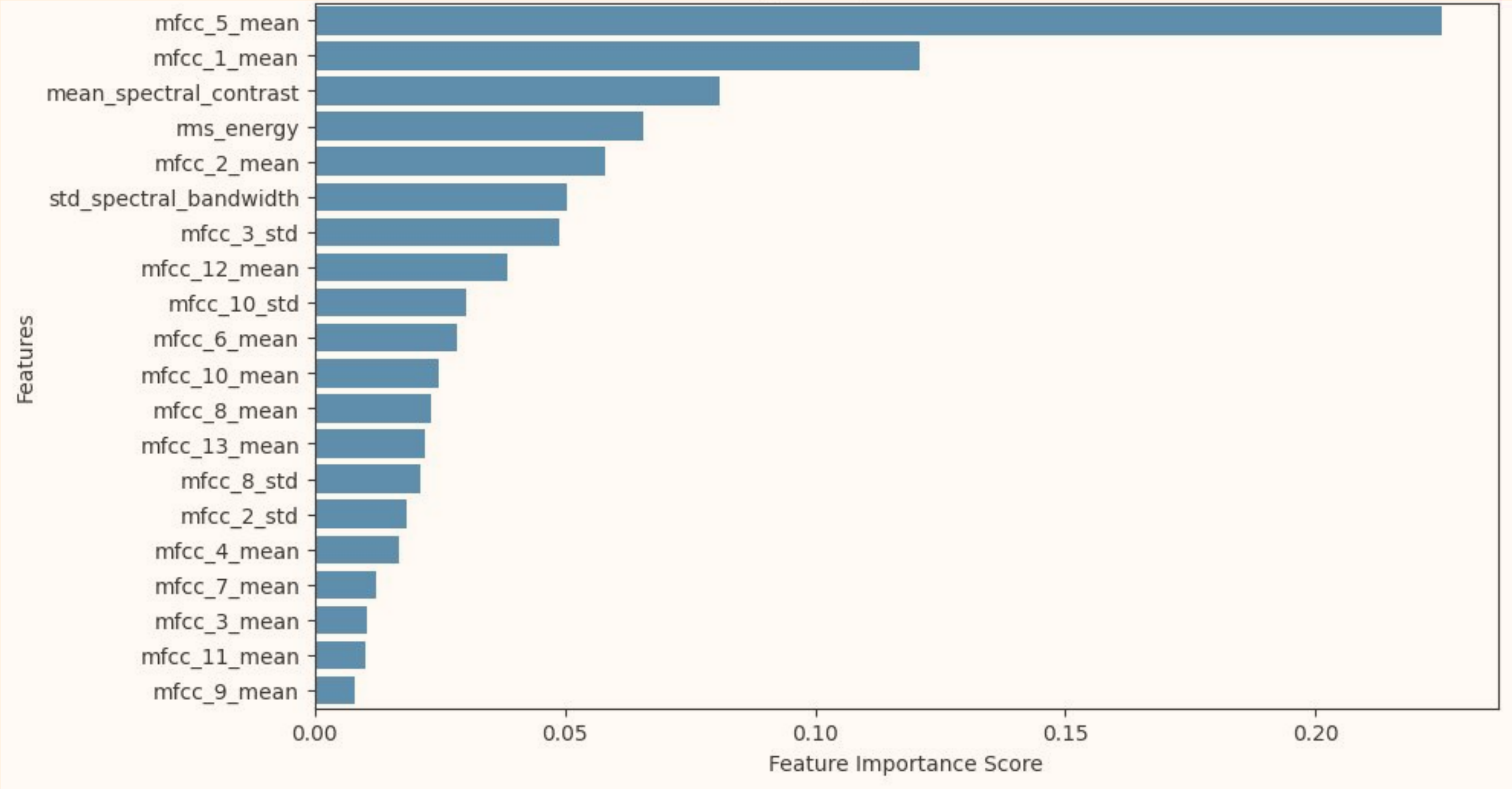
SVM: Maps data into a higher-dimensions to find non-linear decision boundaries, enhancing performance for complex gender patterns.

Random Forest: Captures non-linear patterns, handles noise, and doesn't require feature scaling.

LSTM: Captures long-term dependencies in sequential datas. It effectively handles context and order dependencies but can be computationally expensive.

CNN: Excels at feature extraction from structured data - recognizing patterns in voice characteristics without needing sequential memory.

Top 20 Feature Selection



Most Critical Features Selected

Spectral Features:
mean_spectral_contrast: Males emphasize lower frequencies, females have more high-frequency energy

Pitch Features:
mean_pitch: Differentiates vocal range between genders (lower for males, higher for females).
std_pitch: Measures pitch variation (more fluctuation in female voices).

Energy Features:
rms_energy: Represents loudness (males typically have higher energy due to low-frequency components).

MFCCs (Mel-Frequency Cepstral Coefficients):
mfcc_3_std: Variability in mid-frequency components (differentiates speech patterns).
mfcc_10_std: Variation in high-order spectral features (helps differentiate vocal tone).
mfcc_6_mean: Mid-range spectral properties linked to formants (vocal tract length variations).
mfcc_10_mean: High-frequency details (stronger in female voices).
mfcc_13_mean: High-frequency characteristics (useful for distinguishing timbre).
mfcc_2_std: Variability in low-frequency components (more prominent in male voices).
mfcc_4_mean: Mid-frequency distribution (different resonance patterns).
mfcc_9_mean: Mid-to-high frequency characteristics (helps differentiate vocal texture).