Article

How Much Does the Dynamic F0 Curve Affect the Expression of Emotion in Utterances?

Tae-Jin Yoon *

Posted Date: 1 October 2024

doi: 10.20944/preprints202409.2449.v1

Keywords: emotional speech recognition; fundamental frequency (F0); pitch contours; generalized additive mixed models (GAMMs); non-linear dynamics; speech processing

Article

How Much Does the Dynamic F0 Curve Affect the Expression of Emotion in Utterances?

Tae-Jin Yoon

Department of English Language and Literature, Sungshin Women's University, Seoul 02844, Republic of Korea; tyoon@sungshin.ac.kr; Tel.: 82-2-920-7185

Abstract: The modulation of vocal elements such as pitch, loudness, and duration plays a crucial role in conveying both linguistic information and the speaker's emotional state. While acoustic features like fundamental frequency (F0) variability have been widely studied in emotional speech analysis, challenges remain in accurately classifying emotions due to the complex and dynamic nature of vocal expressions. Traditional analytical methods often oversimplify these dynamics, potentially overlooking intricate patterns indicative of specific emotions. This study aims to enhance emotion classification in speech by directly incorporating dynamic F0 contours into the analytical framework using Generalized Additive Mixed Models (GAMMs). We utilized the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), focusing on eight distinct emotional states expressed by 24 professional actors. Sonorant segments were extracted, and F0 measurements were converted into semitones relative to a 100 Hz baseline to standardize pitch variations. By employing GAMMs, we modeled non-linear trajectories of F0 contours over time, accounting for both fixed effects (emotions) and random effects (individual speaker variability). Our analysis revealed that incorporating emotion-specific non-linear time effects and individual speaker differences significantly improved the model's explanatory power, ultimately explaining up to 66.5% of the variance in F0. The inclusion of random smooths for time within speakers captured individual temporal modulation patterns, providing a more accurate representation of emotional speech dynamics. The results demonstrate that dynamic modeling of F0 contours using GAMMs enhances the accuracy of emotion classification in speech. This approach captures the nuanced pitch patterns associated with different emotions and accounts for individual variability among speakers. The findings contribute to a deeper understanding of the vocal expression of emotions and offer valuable insights for advancing speech emotion recognition systems.

Keywords: emotional speech recognition; fundamental frequency (F0); pitch contours; generalized additive mixed models (GAMMs); non-linear dynamics; speech processing

1. Introduction

The modulation of vocal elements such as pitch, loudness, duration, and voice quality across syllables in an utterance conveys both linguistic information—such as prominence and prosodic phrasing—and non-linguistic information, notably the speaker's emotional state [1]. Emotional states are psychological conditions signaled by neurophysiological changes and associated with thoughts, feelings, and behavioral responses [2]. These states are communicated not only through facial expressions but also through vocal expressions in both linguistic and paralinguistic contexts [3].

When speaking, individuals inherently transmit their emotional status alongside linguistic content. Two prevailing theories attempt to explain how emotions influence behavior and expression: the dimensional theory and the discrete emotion theory [4]. The dimensional theory, proposed by [5] and further developed by [6], suggests that emotions can be distinguished along two primary dimensions: valence (the positivity or negativity of the emotion) and arousal (the intensity of the emotion). According to this perspective, basic emotions are defined within this two-dimensional emotional space.

In contrast, the discrete emotion theory, initially devised by [7] and extensively developed by [2], posits that there are specific, biologically and neurologically distinct basic emotions. Ekman's research supports the view that emotions are discrete, measurable, and physiologically distinct, and that certain emotions are universally recognized across cultures. These basic discrete emotions typically include happiness, sadness, anger, fear, disgust, and surprise.

Acoustic features play a critical role in emotional speech analysis and recognition [4,8]. Features such as fundamental frequency (F0) variability (pitch), voice intensity (energy), and Mel-Frequency Cepstral Coefficients (MFCCs) are commonly used in automatic emotion recognition tasks. MFCCs capture the human auditory frequency response and provide a better representation of the signal than raw frequency bands [9,10]. However, some studies have reported poor emotion classification results using MFCCs, possibly due to the embedded pitch filtering during cepstral analysis that may obscure important pitch-related emotional cues [8].

Despite the availability of rich acoustic data, challenges remain in accurately classifying emotions based on speech. This is partly due to the complexity of emotional expression and the limitations of analytical methods that often simplify dynamic vocal data, potentially overlooking intricate patterns indicative of specific emotions [11]. For instance, previous research suggests that some acoustic features are associated with general characteristics of emotion rather than specific emotional states, and that grouping emotions based on activation levels (e.g., high-activation emotions like anger and joy) can improve recognition performance [12,13]. However, distinguishing between emotions within similar activation levels remains challenging.

Intonation, particularly pitch contour patterns, has been recognized as important in manifesting emotional states [14,15]. However, previous studies often used simplistic intonation parameters, which may contribute to lower classification rates of emotional types based on acoustic features [8]. Simplifying dynamic data can reduce data size and facilitate the use of traditional statistical methods, but it may also result in the loss of potentially interesting patterns [16].

To address these challenges, more sophisticated statistical techniques are needed—particularly those capable of identifying non-linear patterns in dynamic speech data. Generalized Additive Mixed Models (GAMMs) offer such capabilities, allowing for the modeling of non-linear trajectories and interactions [17]. While several speech emotion recognition frameworks combine different feature types, the direct incorporation of dynamic F0 contours into emotion classification systems using GAMMs has not been extensively explored [8].

This study aims to classify emotional states by extracting and analyzing dynamic F0 contours using GAMMs, which have proven effective in capturing non-linear trajectory patterns in speech data [18,19]. By leveraging GAMMs, we seek to uncover the underlying patterns of vocal expression associated with different emotional states, thereby refining emotion classification in speech and advancing our understanding of the vocal expression of emotions.

Our contributions are threefold. First, we utilize the full temporal dynamics of F0 contours rather than static or simplified representations, capturing intricate pitch patterns associated with specific emotions. Second, we apply GAMMs to model non-linear relationships and individual variability in emotional speech, accommodating the complex interplay between time, emotion, and speaker characteristics. And finally, by integrating dynamic F0 contours with GAMMs, we aim to enhance the accuracy of emotion classification, particularly for emotions with similar activation levels.

The remainder of this paper is organized as follows: In Section 2, we present our methodology, including data collection, preprocessing, and the GAMM approach. Section 3 details the results of our analysis and compares them with those of previous approaches. In Section 4, we discuss the implications of our findings and potential limitations. Finally, in Section 5, we draw conclusions and indicate possible directions for future research.

2. Materials and Methods

2.1. Materials

For the statistical modeling of emotion classification, we employed the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [20], a validated multimodal dataset widely used in emotion recognition research. The RAVDESS dataset comprises 2,880 audio-visual files, including both speech and song modalities, expressed across a range of emotions.

The RAVDESS dataset was chosen for several reasons: Professional actors and controlled recording conditions ensure consistent audio quality. The dataset covers a wide range of emotions, essential for comprehensive emotion classification. Equal numbers of male and female actors facilitate gender-related analyses. Using the same sentences across emotions and actors controls for linguistic variability, allowing us to focus on acoustic features related to emotion.

For our analysis, we selected the audio-only speech files (modality code 03, vocal channel 01), focusing on the emotional expressions relevant to our study. Each actor contributed 60 recordings, resulting in a total of 1,440 audio files. This selection ensured a balanced representation of genders and emotions, providing a robust dataset for statistical modeling. These files feature 24 professional actors (12 female and 12 male), each vocalizing two semantically neutral sentences: (1) "Kids are talking by the door." (2) "Dogs are sitting by the door."

Each sentence is articulated with varying emotional expressions and intensities, except for the neutral emotion, which is presented only at a normal intensity.

2.1.1. Emotional Categories and Data Organization

The RAVDESS dataset encompasses eight distinct emotional states: (1) Neutral, (2) Calm, (3) Happy, (4) Sad, (5) Angry, (6), Fearful, (7) Disgust, (8) Surprised. Each emotion (except Neutral) is expressed at two levels of intensity: normal and strong. Every actor provides two repetitions of each sentence per emotional expression and intensity level, resulting in a comprehensive set of emotional speech data.

To streamline data access and analysis, we reorganized the audio files from their original actor-specific directories into eight consolidated folders, each corresponding to one of the emotional categories. This reclassification facilitated efficient retrieval and processing of data associated with each emotion.

Each audio file in the RAVDESS dataset is uniquely named following a structured convention that encodes metadata about the recording. The file naming convention consists of a seven-part numerical identifier in the format: Modality-VocalChannel-Emotion-EmotionalIntensity-Statement-Repetition-Actor. Table 1 summarizes the identifiers used in the RAVDESS dataset. This naming convention facilitated systematic organization and retrieval of files based on attributes such as emotion, intensity, and actor identity.

Table 1. Meta information and identifiers used in the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset.

	Meta information	Identifiers
1	Modality	01 = full-AV, 02 = video-only, 03 = audio-only
2	Vocal Channel	01 = speech, 02 = song
3	Emotion	01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised
4	Emotional Intensity	01 = normal, 02 = strong (cf. no strong intensity for the 'neutral' emotion.)
5	Statement	01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door"
6	Repetition	01 = 1st repetition, 02 = 2nd repetition
7	Actor	01 to 24 (24 actors)
8	Gender	Odd numbered actors = male, even numbered actors = female

Table 2 provides examples of file names and their corresponding metadata from the RAVDESS dataset.

Table 2. Example entries from the RAVDESS dataset.

	File name	Emotion	Intensity	Repetition	Actor	Gender	Statement
1	03-01-05-01-02-01-16	Angry	Normal	1	16	female	Dogs are sitting by the door.
2	03-01-06-01-02-02-16	Fear	Normal	2	16	Female	Dogs are sitting by the door.
3	03-01-06-02-01-02-16	Fear	Strong	2	16	Female	Kids are talking by the door
4	03-01-05-02-01-01-16	Angry	Strong	1	16	Female	Kids are talking by the door
5	03-01-07-01-01-01-16	disgust	neutral	1	16	Female	Kids are talking by the door

2.2. Methods

2.2.1. Preprocessing of F0 Contours

To enhance the accuracy of emotion classification in speech, we focused on the sonorant segments of utterances, where fundamental frequency (F0) extraction is most reliable. F0 values were converted into semitones relative to a standard 100 Hz baseline to normalize inherent pitch variances across speakers.

We utilized the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset [20], which comprises 1,440 audio files representing a range of emotions portrayed by professional actors. Forced alignment techniques were employed to synchronize the textual scripts with their corresponding audio recordings. This alignment was facilitated using Python libraries such as Parselmouth [21], librosa [22], and tgt [23], which collectively enabled the processing of audio files and extraction of vocal features like vowel duration and F0 variability—essential for nuanced emotion classification.

2.2.2. Pitch Extraction and Processing

Pitch extraction was conducted using the autocorrelation method via Praat [24], interfaced through Parselmouth. A pitch ceiling of 450 Hz was set to ensure data fidelity and to accommodate the pitch range of the speakers. The extracted F0 values were smoothed to remove non-numerical anomalies, thereby preserving the integrity of the dataset for analysis.

Our dataset comprised over 126,000 measurement points obtained from 1,618 speech trials, with an average word duration of approximately 0.78 seconds. Due to the fixed sampling rate and variations in word lengths, the number of sample points per word varied, averaging about 78 measurements per word. To address this variability and maintain consistency across the dataset, we employed the resampy package [25] to interpolate the F0 data. This interpolation adjusted the resampled files to match the maximum utterance length in the dataset while retaining the essential shape of the pitch contours. This process was crucial for ensuring that the statistical modeling would be comparable across utterances of different durations.

Our analysis concentrated on sonorant segments, particularly vowels, where F0 can be accurately extracted. We noted that adjacent obstruents and sonorants could exhibit subtle effects on F0, which warranted careful attention. By converting F0 values to semitones relative to a 100 Hz reference tone, we standardized the units for comparison and analysis across different speakers and recordings.

All RAVDESS speech files were subjected to precise time alignment using forced alignment techniques. This process allowed us to focus our analysis on both durational and segmental features, deepening our understanding of the complexities of emotional speech. Recognizing that accurate emotion classification requires a multitude of features, we ensured that our dataset was rich in relevant vocal characteristics.

Figure 1 illustrates an example of the F0 values in semitones extracted from a sample audio file ("03-01-08-01-01-01-03.wav"). The left panel shows the original F0 values, while the right panel demonstrates the resampled pitch contours using the resampy function. In the right panel, the original pitch values are represented by a solid line, and the resampled pitch values are depicted by a dotted line. This visualization highlights the effectiveness of our resampling technique in preserving the inherent shape of the pitch contours, providing a robust foundation for subsequent statistical modeling.

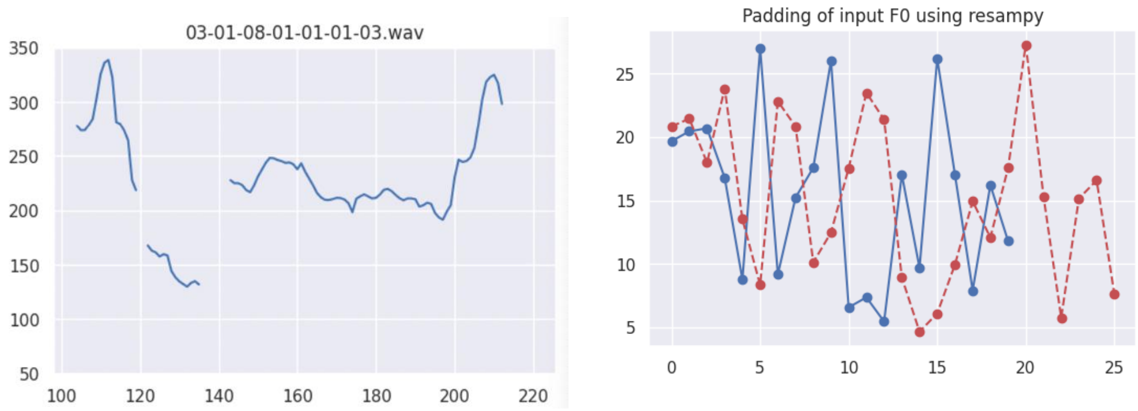

Figure 1. Left panel: Sample of F0 values in semitones extracted from the audio file "03-01-08-01-01-01-03.wav." Right panel: Resampled pitch contours using the resampy function. The original pitch values are represented by the solid line, and the resampled pitch values are represented by the dotted line.

2.3. GAMM

To capture the complex, non-linear relationships inherent in emotional speech data, we employed Generalized Additive Mixed Models (GAMMs). A Generalized Additive Mixed Model (GAMM) is an extension of the generalized linear mixed model (GLMM) that incorporates non-linear relationships in the data through smooth functions. GAMMs are adept at revealing complex patterns not apparent under linear analysis, making them particularly powerful for nuanced analyses where relationships between variables are not strictly linear or change across different levels of another variable.

In our study, we embraced the flexibility of GAMMs to interpret the intricate dynamics within our emotional speech dataset. We constructed our GAMMs using the mgcv package in R (version 1.8-36) [17], a robust environment for statistical computing and graphics. The mgcv package offers comprehensive tools for building and evaluating GAMMs, facilitating the modeling of complex data structures and relationships. It enables the modeling of complex, non-linear, and non-fixed relationships between predictors and the response variable—in our case, the various speech features influencing emotional expression. Visualization and interpretation of the models were enhanced using the itsadug package (version 2.4) in R [26], which provides utilities for plotting smooths, assessing model diagnostics, and interpreting interaction effects. This combination allowed us to effectively communicate the subtleties and strengths of our models through graphical representations.

The use of GAMMs in phonetic research is well-established, with studies demonstrating their efficacy in analyzing dynamic speech patterns: [27] applied GAMMs to investigate phonetic variation and change in Scottish English, capturing complex interactions between social factors and speech acoustics. [28] utilized GAMMs to analyze articulatory trajectories in tongue movement data, revealing non-linear patterns associated with language proficiency. [29] employed GAMMs to model the time-varying nature of electrophysiological responses in psycholinguistic experiments, showcasing the models' capacity to handle time-series data.

These precedents highlight the suitability of GAMMs for our study, which involves modeling the non-linear F0 contours associated with different emotional expressions over time. GAMMs have also been instrumental in investigating the temporal dynamics of speech and the interplay between articulatory movements and acoustic outcomes [19,30], further evidencing their utility in phonetics.

2.3.1. Model Specification

By leveraging this advanced statistical approach, our objective is to uncover the underlying patterns of vocal expression associated with different emotional states, elucidating how these patterns may vary across different speakers or linguistic contexts. Through the rigorous application and interpretation of GAMMs, we aim to contribute significantly to the growing body of knowledge in emotional speech processing and phonetic analysis, providing insights that can be applied across various fields including speech technology, clinical diagnostics, and communication studies.

The comprehensive approach we have adopted, encompassing sophisticated statistical modeling and careful visualization, is designed to offer a more nuanced understanding of the complex relationship between speech features and emotional expression. The results of our GAMM analysis promise to provide a rich, multi-dimensional perspective on the phonetic underpinnings of emotional speech, potentially informing both theoretical frameworks and practical applications in speech science.

3. Results

We developed a series of Generalized Additive Mixed Models (GAMMs) to classify the eight emotion labels in the RAVDESS dataset based on fundamental frequency (F0) contours. Our modeling approach progressively incorporated additional complexity to better capture the nuances of emotional speech.

3.1. Baseline Model

The first model (**Model 0**) was designed to estimate the average (constant) F0 across utterances among the eight emotion labels. We used the bam (Big Additive Models) function from the mgcv package in R [17] to fit the Generalized Additive Model (GAM). The bam function is specifically optimized for large datasets and complex models, making it suitable for our dataset comprising over 750,000 observations. In contrast, the alternative gam function can become prohibitively slow for complex models when fitted to datasets exceeding 10,000 data points, thus reinforcing our choice of bam for efficient computation. We implemented this model using the bam function:

model0 <- bam(F0 ~ Emotions, data=df, method="fREML") (1)

The first parameter of the function 'bam' is the formula reflecting the model specification, in this case: F0 ~ Emotions. The first variable of the formula, F0, is the dependent variable (F0 values in semitone unit across the sonorant components of utterances). The dependent variable is followed by the tilde(~), after which one or more independent variables are added. In this case, the inclusion of a single predictor, Emotions, allows the model to estimate a constant difference among its 8 levels. The parametric coefficients for Model 0 are presented in Table 3.

Table 3. Parametric coefficients of Model 0. The table includes the estimates, standard errors, *t*-values, and p-values for the intercept and each emotional category (Calm, Sad, Fear, Angry, Happy, Disgust, Surprise).

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	7.8841	0.0313	252.3	<2e-16	***
Calm	-0.6751	0.0383	-17.6	<2e-16	***
Sad	2.8221	0.0383	73.7	<2e-16	***[1]
Fear	7.3384	0.0383	191.7	<2e-16	***
Angry	6.9627	0.0383	181.9	<2e-16	***
Happy	5.7407	0.0383	150.0	<2e-16	***
Disgust	2.2176	0.0383	57.9	<2e-16	***
Surprise	6.2874	0.0383	164.3	<2e-16	***

[1] Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Table 4 provides the key summary statistics for Model 0, providing a clear overview of the model's performance and fit to the dataset. The adjusted R-squared value is 0.145, suggesting that 14.5% of the variance in the response variable is explained by the model. The restricted maximum likelihood (REML) estimate is 2.5511e+06, and the estimated scale parameter is 49.317. The model was fitted to a dataset comprising 757,440 observations.

Table 4. Summary statistics for Model 0.

R-sq.(adj)	Deviance explained	fREML	Scale est.	N
0.145	14.5%	2.5511e+06	49.317	757440

The Figure 2 illustrates the distribution of mean F0 for each emotion. It is evident in the figure that mean F0 alone is not sufficient in explaining the types of emotion.

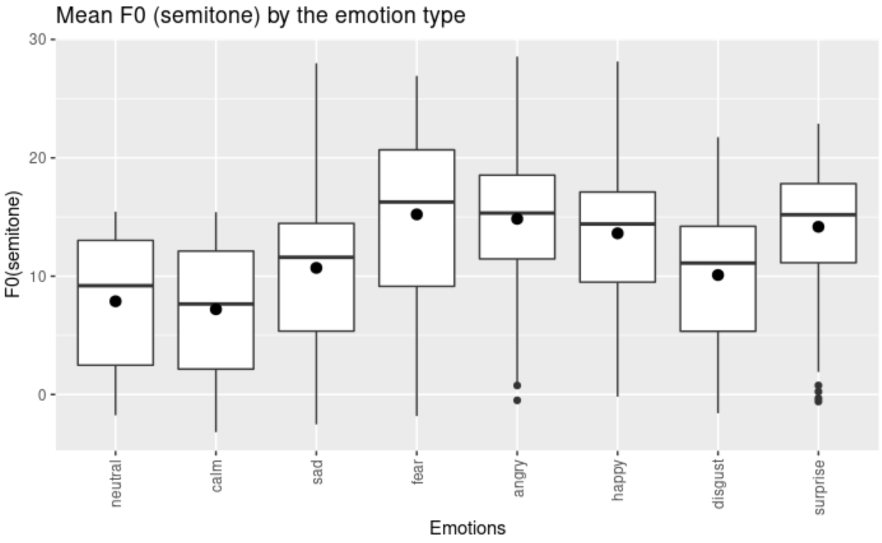

Figure 2. Boxplots of mean F0 (semitones) for each emotion category. The dot inside each boxplot denotes the mean F0 for that emotion.

3.2. Incorporating non-linear time effects

We fit another simple model (model1) which includes the constant difference among emotions, but only a single smooth. The function s sets up a smooth over the parameter (i.e., Time). As such, model1 assumes that the pattern over time is the same for all the 8 emotions. The modified generalized additive model was specified to include a non-linear pattern over time. In this model, the helmert contrast was applied with the order of neutral, calm, sad, fear, angry, happy, disgust, and surprise (the same order that is observed in Figure 3). We implemented this model using the bam function:

$$\text{model1} \sim \text{bam}(\text{F0} \sim \text{Emotions} + \text{s}(\text{Time}),\ \text{data} = \text{df}) \qquad (2)$$

The output of model1 is given in Table 5 for coefficient estimates and Table 6 for approximate significance of smooth terms:

Table 5. Parametric coefficients of model1.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	7.8841	0.0304	259.3	<2e-16	***
Calm	-0.6751	0.0372	-18.1	<2e-16	***
Sad	2.8221	0.0372	75.9	<2e-16	***1
Fear	7.3384	0.0372	197.3	<2e-16	***

Angry	6.9627	0.0372	187.2	<2e-16	***
Happy	5.7407	0.0372	154.3	<2e-16	***
Disgust	2.2176	0.0372	59.6	<2e-16	***
Surprise	6.2874	0.0372	169.0	<2e-16	***

[1] Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

The intercept represents the estimated mean F0 for the Neutral emotion at the reference time point. The coefficients for each emotion indicate the difference in mean F0 compared to the Neutral emotion, controlling for the effect of time.

The approximate significance of the smooth term s(Time) is provided in Table 6. In the table, edf stands for the estimated degrees of freedom, which reflects the complexity of the smooth term. An edf close to 9 suggests a highly flexible function, allowing for intricate non-linear patterns in F0 over time. Ref.df for reference degree of freedom is used for hypothesis testing of the smooth term. The significant p-value (<0.001) confirms that the non-linear effect of time on F0 is highly significant.

Table 6. Approximate significance of smooth terms.

Title 1	edf	Ref.df	F	p-value
s(Time)	8.96	9	4945	<2e-16 ***

Table 6 presents the key summary statistics for Model 1. The adjusted R-squared value of 0.192 indicates that Model 1 explains 19.2% of the variance in F0, an improvement from the 14.5% explained by Model 0. This increase demonstrates the importance of modeling the temporal dynamics of F0.

Table 7. Summary statistics for Model 1.

R-sq.(adj)	Deviance explained	fREML	Scale est.	N
0.192	19.2%	2.5295e+06	46.58	757440

The adjusted R-squared value of 0.192 indicates that Model 1 explains 19.2% of the variance in F0, an improvement from the 14.5% explained by Model 0. This increase demonstrates the importance of modeling the temporal dynamics of F0.

In Figure 3, the smooth curve represents the estimated F0 contour over the normalized time course of the utterances, averaged across all emotions. The non-linear shape of the curve reflects the dynamic changes in pitch that are characteristic of spoken utterances.

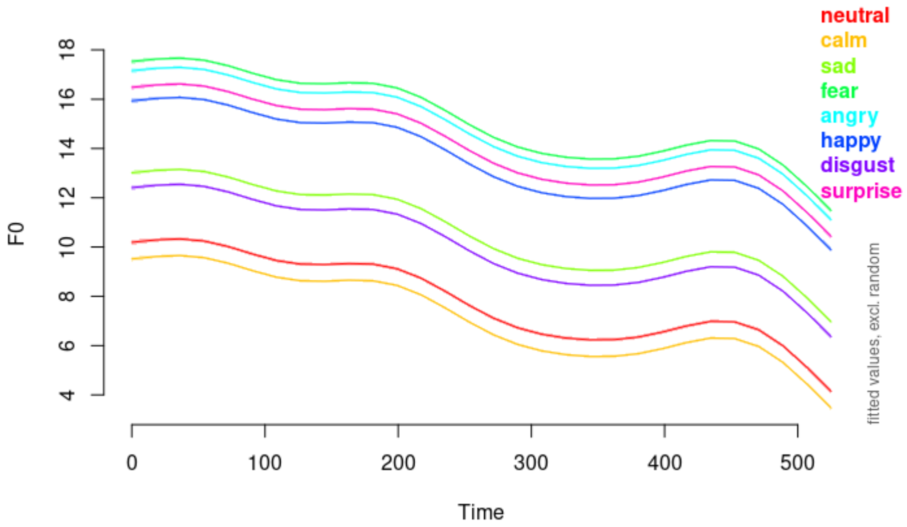

Figure 3. the modeled F0 trends over time only.

Incorporating the non-linear effect of time significantly improved the model fit, as evidenced by the increase in adjusted R-squared and deviance explained. The smooth term captures the inherent temporal dynamics of speech, which are essential for accurately modeling F0 contours.

However, Model 1 assumes that the shape of the F0 contour over time is identical across all emotions, differing only in their average levels. This assumption may not hold true, as different emotions can exhibit distinct temporal patterns in pitch modulation. For instance, emotions like surprise may have abrupt pitch changes, while sadness may show more gradual contours.

To address this limitation, we developed Model 2, which allows for emotion-specific smooth functions over time, enabling us to capture the unique F0 trajectories associated with each emotion.

3.3. Modeling Emotion-Specific Non-Linear Time Effects

While Model 1 incorporated a non-linear effect of time on F0, it assumed that the temporal pattern of F0 variation was identical across all emotions, differing only in their average levels. However, different emotions may exhibit distinct temporal patterns in pitch modulation. To capture these potential differences, we extended our modeling approach by allowing the non-linear effect of time to vary by emotion.

In Model 2, we introduced emotion-specific smooth functions over time, enabling the model to capture unique F0 trajectories associated with each emotion. We implemented this model using the bam function:

$$\text{model2} \sim \text{bam}(\text{F0} \sim \text{Emotions} + \text{s}(\text{Time}, \text{by=Emotions}), \text{data} = \text{df}) \quad (3)$$

In this model, the s(Time, by=Emotions) term allows for a separate smooth function over Time for each level of the Emotions variable. This means that each emotion can have its own unique F0 contour over time, providing a more flexible and detailed modeling of the data. The parametric coefficients from Model 2 are presented in Table 8.

Table 8. Parametric coefficients of model2.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	7.86715	0.03021	260.43	<2e-16	***
Calm	-0.67342	0.03700	-18.20	<2e-16	***
Sad	2.83017	0.03700	198.50	<2e-16	***1
Fear	7.34380	0.03700	197.3	<2e-16	***
Angry	6.97021	0.03700	188.40	<2e-16	***
Happy	5.73618	0.03700	155.04	<2e-16	***
Disgust	2.22634	0.03700	60.17	<2e-16	***
Surprise	6.32826	0.03700	171.04	<2e-16	***

[1] Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

The intercept represents the estimated mean F0 for the Neutral emotion at the reference time point. The coefficients for each emotion indicate the difference in mean F0 compared to the Neutral emotion, controlling for the emotion-specific non-linear effects of time.

The approximate significance of the emotion-specific smooth terms s_e(Time) is provided in Table 9. The edf values range from approximately 7.8 to 8.9, indicating that the smooth functions are flexible enough to capture complex non-linear patterns for each emotion. All smooth terms are highly significant ($p < 0.001$), suggesting that the emotion-specific non-linear effects of time significantly improve the model fit.

Table 9. Approximate significance of smooth terms.

Title 1	edf	Ref.df	F	p-value
s(Time):Neutral	7.844	8.654	572.0	<2e-16 ***
s(Time):Angry	8.895	8.997	764.4	<2e-16 ***

s(Time):Happy	8.882	8.996	1086.2	<2e-16 ***
s(Time):Fear	8.648	8.964	426.0	<2e-16 ***
s(Time):Sad	8.272	8.854	583.8	<2e-16 ***
s(Time):Calm	8.293	8.862	909.9	<2e-16 ***
s(Time):Disgust	8.841	8.992	779.1	<2e-16 ***
s(Time):Surprise	8.983	9.000	922.2	<2e-16 ***

The key summary statistics for Model 2 are presented in Table 10. The adjusted R-squared value of 0.202 indicates that Model 2 explains 20.2% of the variance in F0, an improvement from the 19.2% explained by Model 1. This increase demonstrates the importance of allowing for emotion-specific temporal dynamics in modeling F0 contours.

Table 10. Summary statistics for Model2.

R-sq.(adj)	Deviance explained	fREML	Scale est.	N
0.202	20.2%	2.5252e+06	46.038	757440

To assess whether Model 2 provides a significantly better fit than Model 1, we compared the two models using a likelihood ratio test (approximated via the difference in fREML scores). The comparison is summarized in Table 11. The significant reduction in the fREML score (ΔfREML = 4,294, $p < 0.001$) indicates that Model 2 provides a significantly better fit to the data than Model 1. The increase in edf reflects the additional complexity introduced by allowing separate smooth functions for each emotion.

Table 11. Comparison of the two models model1 and model2.

Model	Score	Edf	Difference	Df	P value	Sig.
Model1	2529552	10				
Model2	2525258	24	4293.991	14.0	<2e-16	***

Figure 4 illustrates the modeled F0 contours over time for each emotion, as estimated by Model 2.

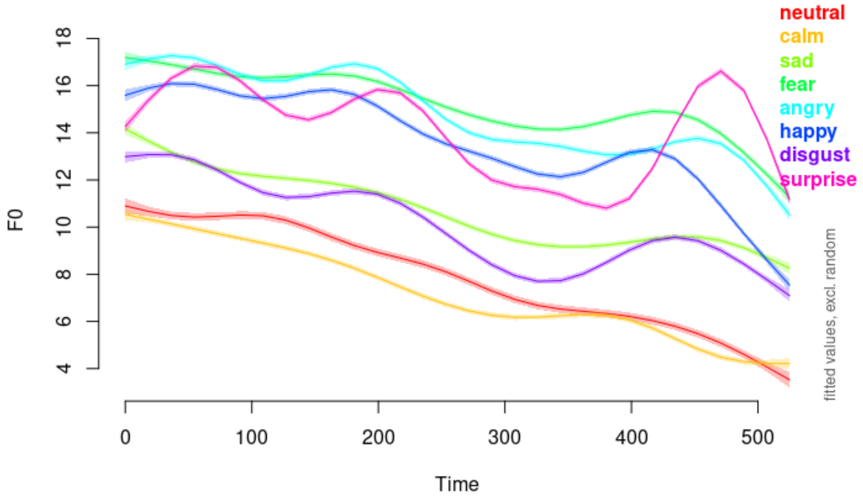

Figure 4. the modeled F0 contours over time for each emotion, estimated by Model 2.

In Figure 4, each panel represents the estimated F0 contour for one emotion over the normalized time course of the utterances. The distinct shapes of the contours demonstrate how different emotions exhibit unique temporal patterns in pitch modulation. For example, **Surprise** shows a rapid increase in F0 towards the end of the utterance. **Sadness** exhibits a gradual decline in F0 over time. And H**appy** displays fluctuations in F0, reflecting dynamic pitch variation.

These visualizations highlight the benefits of modeling emotion-specific non-linear time effects, capturing the nuanced acoustic signatures of each emotion. While Model 2 allows for emotion-specific non-linear time effects, it does not yet account for individual variability among speakers (actors). Speakers may differ in their baseline pitch levels and in how they express emotions acoustically. To address this, we need to incorporate random effects for speakers, as well as potential interactions between speakers and emotions. Additionally, our current model does not include random slopes for the non-linear time effects across individuals. This means that while we have accounted for the average emotion-specific F0 contours, we have not yet modeled how individual speakers might vary in their expression of these contours.

3.4. Incorporating Random Effects for Speakers

While Models 1 and 2 accounted for non-linear time effects—both general and emotion-specific—they did not consider individual variability among speakers (actors). In speech data, especially emotional speech, individual differences can significantly impact acoustic features like fundamental frequency (F0). Speakers may have different baseline pitch levels and may express emotions with varying degrees of intensity and modulation. Ignoring this variability can lead to biased estimates and reduced model accuracy.

To address this, we extended our modeling approach by incorporating random effects for speakers in **Model 3**. This allows us to account for both the random intercepts (differences in baseline F0 levels among speakers) and random slopes (differences in how speakers' F0 contours change over time and across emotions).

Model 3 introduces a random intercept for each speaker (actor) to capture individual baseline differences in F0. We implemented this model using the bam function with the following specification:

Model3 ~ bam(F0 ~ Emotions + s(Time, by=Emotions), + s(Actor, bs=”re”), data = df)

In the model, the term s(Actor, bs="re") specifies a random intercept for each actor. This models the variability in baseline F0 levels across different speakers. The term s(Time, by=Emotions) allows each emotion to have its own non-linear F0 contour over time, as in Model 2. And in Generalized Additive Mixed Models (GAMMs), random effects can be modeled using smooth terms with a special basis function (bs="re"). This approach treats random effects as smooth functions with a penalty that shrinks them towards zero unless supported by the data [17].

The parametric coefficients from Model 3 are presented in Table 12. The intercept now reflects the mean F0 for the Neutral emotion, adjusted for the random effects of speakers. The standard errors of the emotion coefficients have decreased compared to previous models, indicating more precise estimates after accounting for speaker variability.

Table 12. Parametric coefficients of model3.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	7.86671	0.97132	8.099	5.55e-16	***
Calm	-0.67314	0.02688	-25.043	<2e-16	***
Sad	2.83025	0.02688	105.295	<2e-16	***1
Fear	7.34407	0.02688	273.224	<2e-16	***
Angry	6.96962	0.02688	259.292	<2e-16	***
Happy	5.73642	0.02688	213.413	<2e-16	***
Disgust	2.22613	0.02688	82.819	<2e-16	***
Surprise	6.32624	0.02688	235.357	<2e-16	***

[1] Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1.

The significance of the smooth terms in Model 3 is provided in Table 14. The edf values for the emotion-specific smooths remain high (around 8.4 to 8.9), indicating complex non-linear F0 contours for each emotion. The term s(Actor) has an edf of approximately 23, reflecting the 24 actors in the dataset (since edf = number of levels - 1). The extremely high F-value and significant p-value indicate substantial variability among speakers.

Table 14. Approximate significance of smooth terms in Model 3.

Title 1	edf	Ref.df	F	p-value
s(Time):Neutral	8.439	8.911	1053.9	<2e-16 ***
s(Time):Angry	8.945	8.999	1448.7	<2e-16 ***
s(Time):Happy	8.938	8.999	2058.1	<2e-16 ***
s(Time):Fear	8.806	8.989	805.8	<2e-16 ***
s(Time):Sad	8.580	8.949	1095.1	<2e-16 ***
s(Time):Calm	8.610	8.956	1706.5	<2e-16 ***
s(Time):Disgust	8.915	8.998	1476.2	<2e-16 ***
s(Time):Surprise	8.991	9.000	1747.5	<2e-16 ***
s(Actor)	22.999	23.000	29457.0	<2e-16 ***

Table 15 presents the summary statistics for Model 3. The adjusted R-squared has increased dramatically to 0.579 from 0.202 in Model 2, indicating that Model 3 explains 57.9% of the variance in F0. This substantial improvement underscores the importance of accounting for speaker variability.

Table 15. Summary statistics for Model 3.

R-sq.(adj)	Deviance explained	fREML	Scale est.	N
0.579	57.9%	2.2834e+06	24.3	757440

We compared Model 3 with Model 2 as in Table **16** to assess the impact of including random effects for speakers. The large reduction in fREML score (ΔfREML = 241,865, p < 0.001) indicates that including random intercepts for actors significantly improves model fit. The increase in edf by 1 corresponds to the addition of the random intercept term.

Table 16. Comparison of the two models model2 and model3.

Model	Score	Edf	Difference	Df	P value	Sig.
Model2	2525258	24				
Model3	2283393	25	241865.61	1.000	<2e-16	***

Incorporating random effects for speakers in Model 3 significantly enhances the model's ability to capture the variability in F0 associated with different emotions and individual speakers. The substantial increase in the adjusted R-squared value demonstrates that speaker variability accounts for a large portion of the unexplained variance in previous models.

By modeling speaker-specific baseline F0 levels, we obtain more precise estimates of the fixed effects (emotions) and the smooth terms. Accounting for individual differences enhances the model's applicability to new data, as it can generalize better across different speakers. Ignoring random effects can lead to biased estimates of fixed effects and overestimation of significance levels.

While Model 3 includes random intercepts for speakers, it does not yet account for potential interactions between speakers and emotions (i.e., random slopes for emotions within speakers). Speakers may express emotions differently, leading to variability in the effect of emotions on F0 across individuals. Additionally, the model assumes that the emotion-specific F0 contours are the same for all speakers after adjusting for baseline differences. To capture individual differences in how emotions affect F0 contours over time, we need to include random slopes and possibly interaction terms between Time, Emotions, and Actors.

3.5. Incorporating Random Slopes for Emotions within Speakers

While Model 3 accounted for random intercepts for speakers, it did not consider the possibility that the effect of emotions on F0 might vary across speakers. In other words, different speakers may not only have different baseline pitch levels but may also express emotions differently in terms of pitch modulation. To capture this additional layer of variability, we extended our model to include random slopes for emotions within speakers in **Model 4**.

Model 4 includes both random intercepts and random slopes for emotions within speakers. We implemented this model using the bam function:

$$\begin{gathered}\text{model4} \sim \text{bam(F0} \sim \text{Emotions} + \text{s(Time, by=Emotions)} + \text{s(Actor, bs="re")} \\ + \text{s(Actor, Emotions, bs="re"), data=df)}\end{gathered} \tag{4}$$

The term s(Actor, Emotions, bs="re") allows the effect of each emotion on F0 to vary across speakers. This models the interaction between speakers and emotions. In the context of Generalized Additive Models (GAMs), random effects are represented using smooth terms with a "random effect" basis (bs='re'). According to [28], it is not possible to model correlations between random intercepts and random slopes in GAMs as one might in linear mixed-effects models (e.g., (1+Emotions|Actor) in lmer) [31]. Therefore, we specify random intercepts and slopes separately.

The parametric coefficients from Model 4 are presented in Table 17. The effect of **Calm** is no longer statistically significant (p = 0.28324), suggesting that after accounting for random slopes, Calm does not differ significantly from Neutral in terms of mean F0. Compared to Model 3, the standard errors of the emotion coefficients have increased, reflecting the additional variability introduced by allowing emotion effects to vary by speaker.

Table 17. Parametric coefficients of model4.

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**	
(Intercept)	7.867	1.063	7.40	1.4e-13	***
Calm	-0.673	0.627	-1.07	0.28324	
Sad	2.830	0.627	4.51	6.4e-06	***[1]
Fear	7.344	0.627	11.71	<2e-16	***
Angry	6.967	0.627	11.11	<2e-16	***
Happy	5.736	0.627	9.14	<2e-16	***
Disgust	2.226	0.627	3.55	0.00039	***
Surprise	6.326	0.627	10.08	<2e-16	***

[1] Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

The significance of the smooth terms in Model 4 is provided in Table 18. The edf values for the emotion-specific smooths remain high, indicating complex non-linear patterns for each emotion. The term s(Actor) is highly significant, confirming substantial variability in baseline F0 among speakers. The term s(Actor, Emotions) is not statistically significant (p = 0.23), suggesting that allowing the effect of emotions on F0 to vary across speakers does not significantly improve the model fit in this case.

Table 18. Approximate significance of smooth terms in Model 4.

	edf	Ref.df	F	p-value
s(Time):Neutral	8.54	8.94	1261	<2e-16 ***
s(Time):Angry	8.95	9.00	1739	<2e-16 ***
s(Time):Happy	8.95	9.00	2470	<2e-16 ***
s(Time):Fear	8.84	8.99	967	<2e-16 ***
s(Time):Sad	8.64	8.96	1313	<2e-16 ***

s(Time):Calm	8.66	8.97	2046	<2e-16 ***
s(Time):Disgust	8.93	9.00	1772	<2e-16 ***
s(Time):Surprise	8.99	9.00	2098	<2e-16 ***
s(Actor)	22.41	23.00	32268186	<2e-16 ***
s(Actor,Emotions)	161.42	184.00	120351	0.23

Table 19 presents the summary statistics for Model 4. The adjusted R-squared value has increased to 0.649 from 0.579 in Model 3, indicating that Model 4 explains 64.9% of the variance in F0.

Table 19. Summary statistics for Model 4.

R-sq.(adj)	Deviance explained	fREML	Scale est.	N
0.649	64.9%	2.2148e+06	20.245	757440

Figure 5 illustrates the modeled F0 contours over time for each emotion, as estimated by Model 3.

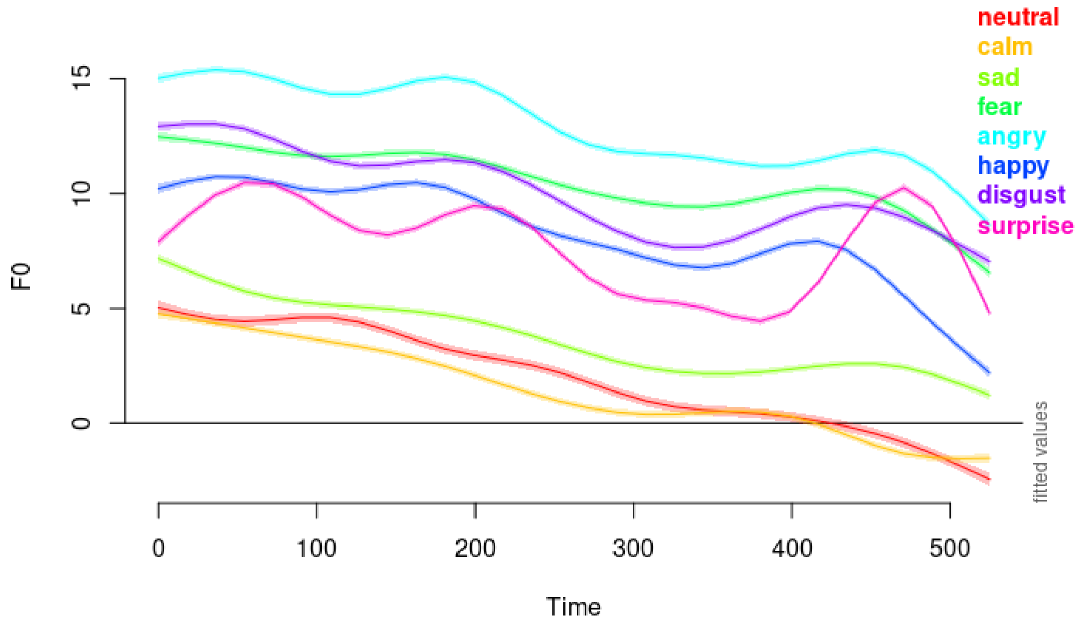

Figure 5. the modeled F0 contours over time for each emotion, estimated by Model 4.

In Table 20, we compared Model 4 with Model 3 to evaluate the impact of adding random slopes for emotions within speakers. The significant reduction in the fREML score (ΔfREML = 68,576, p < 0.001) suggests that Model 4 provides a better fit than Model 3. However, given that the random slopes for emotions within actors were not statistically significant, the improvement may be attributed primarily to other factors in the model.

Table 20. Comparison of the two models model3 and model4.

Model	Score	Edf	Difference	Df	P value	Sig.
model3	2283393	24				
model4	2214817	26	68575.483	1.000	<2e-16	***

Model 4 incorporated random slopes for emotions within speakers to account for potential variability in emotional expression across individuals. While the random slopes were not statistically significant, the model showed an improved fit over Model 3, suggesting that including these terms may still capture some variability not accounted for previously.

3.6. Incorporating Random Slopes for Emotions within Speakers

To further refine our model and capture individual differences in how speakers modulate their pitch contours over time, we extended the model to include random smooths for Time within speakers in Model 5. Model 5 adds a random smooth term for Time within speakers, allowing each speaker to have their own non-linear F0 contour over time. We implemented this model using the bam function:

```
model5 <- bam(F0 ~ Emotions + s(Time, by=Emotions) + s(Actor, bs="re")
          + s(Actor, Emotions, bs="re") + s(Time, Actor, bs="fs", m=1), data = df)
```

In the model, the term s(Time, Actor, bs="fs", m=1) allows each speaker to have their own smooth F0 contour over time. The "factor-smooth" basis (bs="fs") models interactions between a continuous variable (Time) and a factor (Actor). Setting m=1 penalizes the first derivative (the speed) of the smooth, resulting in less wiggly estimates compared to the default second derivative penalty (which penalizes acceleration).

The parametric coefficients from Model 5 is identical to those from Model 4. The significance of the smooth terms in Model 5, which differs from that of Model 4, is provided in Table 21. In Model 5, the term s(Time, Actor) is statistically significant (p = 0.0038), indicating that allowing each speaker to have their own F0 contour over time significantly improves the model fit. The edf for s(Actor) is lower than in previous models, suggesting that some of the variability previously captured by random intercepts is now being modeled by the random smooths over time.

Table 21. Approximate significance of smooth terms in Model 5.

	edf	Ref.df	F	p-value
s(Time):Neutral	8.72	8.79	68.5	<2e-16 ***
s(Time):Angry	8.80	8.84	111	<2e-16 ***
s(Time):Happy	8.90	8.92	151	<2e-16 ***
s(Time):Fear	8.83	8.86	107	<2e-16 ***
s(Time):Sad	8.76	8.80	85.8	<2e-16 ***
s(Time):Calm	8.75	8.80	62.5	<2e-16 ***
s(Time):Disgust	8.84	8.87	94	<2e-16 ***
s(Time):Surprise	8.94	8.95	271	<2e-16 ***
s(Actor)	11.20	23.00	0.97	<2e-16 ***
s(Actor, Emotions)	161.42	184.00	863	<2e-16 ***
S(Time, Actor)	197.42	214.0	1.84e+07	0.0038**

Table 22 presents the summary statistics for Model 5. The adjusted R-squared value has increased to 0.665 from 0.649 in Model 4, indicating that Model 5 explains 66.5% of the variance in F0.

Table 22. Summary statistics for Model 5.

R-sq.(adj)	Deviance explained	fREML	Scale est.	N
0.664	66.5%	2.1982e+06	19.347	757440

Figure 6 illustrates the emotion-specific F0 contours over time from Model 5, adjusted for speaker-specific temporal patterns.

doi:10.20944/preprints202409.2449.v1

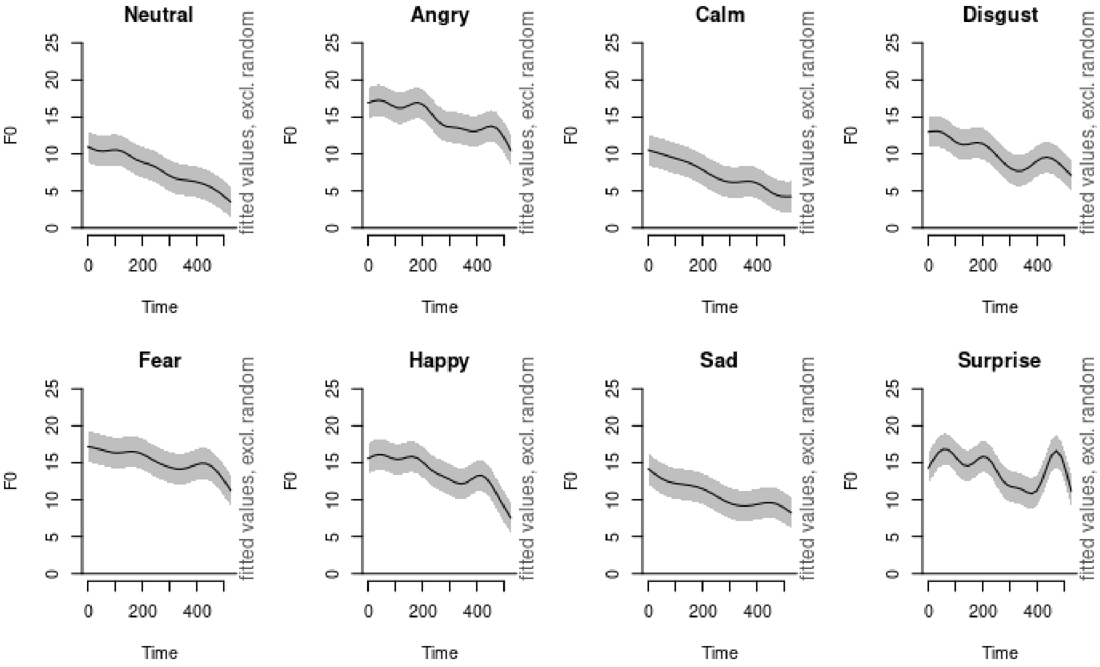

Figure 6. F0 contours of each emotional types modeled with model4.

Model 5 represents the most comprehensive model in our approach to emotion classification using dynamic F0 values, incorporating random intercepts, random slopes for emotions, and random smooths for Time within speakers. Including random smooths for Time within speakers significantly improved the model fit, as evidenced by the increase in adjusted R-squared and the significance of the s(Time, Actor) term. This model accounts for individual differences in how speakers modulate their pitch over time, providing a more accurate representation of the data. The emotion coefficients remained largely consistent with previous models, but the standard errors decreased slightly, indicating more precise estimates. As such, model 5 provides the best fit to the data, explaining 66.5% of the variance in F0.

3.7. Adding Additional Factors

While our primary focus is on Model 5 due to its emphasis on F0 dynamics, we also explored the inclusion of additional factors—such as utterance duration, statement type, gender, and intensity—to provide a more comprehensive understanding of their potential impact. These factors are provided here for reference. Table 23 summarizes the deviance explained by adding these factors to the model.

Table 23. Summary of deviance explained by adding additional factors (duration, statement type, gender, and intensity) to model5.

Additional Factors	Deviance explained
Utterance Duration by Emotions	69.9%
Statement + Utt. Duration by Emotions	71.4%
Gender + Utt. Duration by Emotions	71.4%
Intensity	73.1%
Intensity + Utt. Duration by Emotions	76.3%
Intensity +Statement + Utt. Duration by Emotions	76.3%
Intensity +Statement + Gender + Utt. Duration by Emotions	76.3%

4. Discussion

Our findings shed light on the intricate role of dynamic F0 contours in speech emotion recognition and highlight both consistencies and discrepancies with previous research. [8] provided a foundational summary of the acoustic characteristics associated with various emotions. They reported that anger typically exhibits the highest energy and pitch level; disgust is characterized by a low mean pitch, low intensity, and slower speech rate compared to neutral speech; fear correlates with a high pitch level and increased intensity; and sadness is associated with low mean intensity and pitch. Additionally, they emphasized that pitch contour trends are valuable in distinguishing emotions, particularly noting that fear resembles sadness with an almost downward slope in the pitch contour, which helps separate it from joy.

However, statistics such as mean and variance of pitch, while informative, are rudimentary and may not capture the complex temporal dynamics of emotional speech. Our study advances this understanding by directly modeling dynamic F0 contours using Generalized Additive Mixed Models (GAMMs) for the classification of eight basic emotions in the RAVDESS corpus.

One of the key advantages of using GAMMs in our study is the **direct modeling of dynamic F0 contours**. Unlike traditional methods that might rely on summary statistics, GAMMs allow us to capture the non-linear, temporal patterns in pitch that are critical for differentiating emotions. This approach not only improves classification performance but also enhances interpretability, enabling us to understand which aspects of the pitch contour contribute to recognizing specific emotions.

4.1. Comparative Analysis with Previous Studies

Several prior studies have attempted emotion classification using the RAVDESS dataset with varying degrees of success. For instance, [32] utilized Support Vector Machines (SVM) with features selected via Continuous Wavelet Transform (CWT) and achieved an accuracy of **60.1%**. [33] focused on a subset of four emotions and obtained an accuracy of **57.14%** using group multi-task feature selection.

Deep learning approaches have reported higher accuracies. [34] implemented a Deep Neural Network (DNN) using spectrograms and achieved **65.9%** accuracy. [35] fine-tuned a pre-existing DNN and, using VGG-16, reported an accuracy of **71%**. Similarly, [36] employed a one-dimensional deep CNN with a combination of acoustic features like MFCCs, mel-scaled spectrograms, chromograms, spectral contrast features, and tonnetz (e.g. tone network) representations, reaching an accuracy of **71.61%**.

While these deep learning models outperform human accuracy rates of **67%** reported by [20], they often function as black boxes, offering limited insight into which features drive the classification. The use of multiple acoustic features in these models combines various sound characteristics, potentially leading to improved performance. However, this complexity makes it challenging to discern the contribution of individual features to the overall classification, which can hinder the interpretability and explainability of the results.

4.2. Current Approaches to Contributing to Interpretability

In contrast, our approach achieved an accuracy of **68.2%**, which is higher than human performance (67%) and comparable to some deep learning models (71% in [35] and 71.61% in [36]). More importantly, by focusing on dynamic F0 contours and employing GAMMs, we provide a transparent model that elucidates which acoustic features are critical for emotion recognition. This interpretability is crucial for applications where understanding the basis of the classification decision is as important as the decision itself, such as in clinical settings or human-computer interaction design.

For example, our model highlights why certain emotions are more challenging to distinguish. The similarity in pitch contours between calm and neutral speech suggests that listeners and models alike may struggle to differentiate these emotions based solely on pitch information. This insight can guide future research to incorporate additional features, such as spectral properties or articulatory cues, to improve classification accuracy for these emotions.

4.3. Discrepancies with Previous Literature

In our analysis, as illustrated in Figures 5 and 6, we observed that **anger**, **disgust**, and **fear** all exhibit downward pitch contours at elevated pitch levels. Emotions such as **sadness, calm, and neutral** are characterized by downward pitch contours at subdued pitch levels. Notably, **sadness occurs at a slightly higher pitch level than calm and neutral**, which might reflect a difference in emotional arousal. The **similarity between the pitch contours of calm and neutral** suggests that these two emotions may be challenging to distinguish for listeners, potentially leading to misclassification. This difficulty underscores the importance of nuanced acoustic analysis in emotion recognition systems.

Among these emotions, **anger** starts at a higher pitch level than both **disgust** and **fear**, aligning partially with [8] findings regarding anger's high pitch. While disgust also shows a downward contour similar to anger, it displays more fluctuation in the pitch contour compared to fear.

To further examine the nuances between **anger** and **disgust**, we present the modeled F0 contours for these emotions in Figure 7. As depicted in the left panel of Figure 7, both emotions demonstrate very similar shapes in their dynamic F0 movements, indicating a shared downward trajectory in pitch over time. However, the **anger** emotion (red line) consistently exhibits higher pitch levels compared to **disgust** (blue line) throughout the utterance.

The right panel of Figure 7 illustrates the estimated difference in F0 between the angry and disgust conditions, along with the confidence interval of the difference. This visual comparison highlights that while the dynamic patterns of pitch movement are similar for both emotions, the overall **pitch level** serves as a distinguishing feature. The elevated pitch levels associated with anger support the notion of increased arousal and intensity in this emotion, which is reflected acoustically.

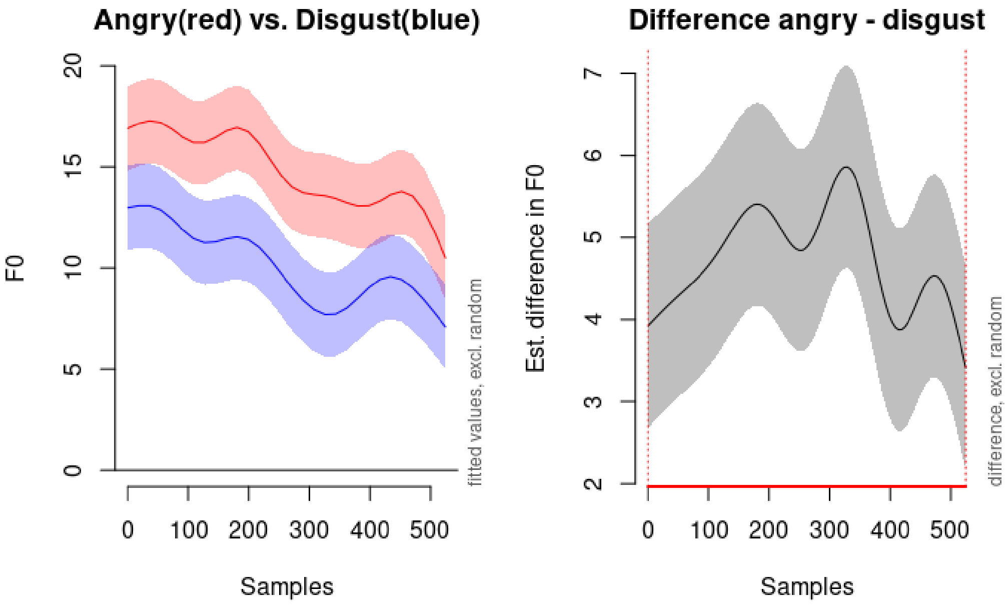

Figure 7. Comparison of pitch contours between Angry (red) and Disgust (blue) speech samples. The left panel displays the fitted pitch values in semitones over 500 samples, with shaded areas representing the confidence intervals excluding random effects. The right panel shows the estimated difference in F0 (fundamental frequency) between the Angry and Disgust conditions, with the shaded area indicating the confidence interval of the difference.

Interestingly, **disgust in our data does not conform to the low mean pitch reported previously**. In [8], **disgust** is expressed with a low mean pitch, a low intensity level, and a slower speech rate than the neutral state does. However, it is not the case in our data. Instead, it displays elevated pitch levels with more fluctuation in the pitch contour compared to fear. This discrepancy suggests that disgust may be expressed differently in the RAVDESS dataset, potentially due to cultural, linguistic, or methodological differences. The expression and perception of emotions can vary across cultures and languages. What is considered a typical acoustic manifestation of an emotion in one culture may differ in another. Or it can be the case that the actors in the RAVDESS dataset may have employed a different vocal strategy to convey disgust, perhaps emphasizing certain prosodic features to make the emotion more discernible. Methodological difference may have led to the differences in the finding; our use of dynamic modeling with GAMMs allows for capturing temporal variations in pitch contours that static measures like mean pitch may overlook.

The acoustic properties of **clam** and **sadness** emotions are characterized by downward pitch contours at subdued pitch levels, reflecting their low arousal states. This downward trajectory in pitch aligns with previous research indicating that lower pitch and reduced variability are common in less active or more negative emotional states. However, there are notable differences between the two emotions. Sadness consistently exhibits a slightly higher pitch level than calm throughout the utterance. This elevated pitch in sadness may convey a sense of emotional weight or poignancy, distinguishing it from the more neutral or relaxed state of calm. While both emotions are low in arousal, sadness typically carries a negative valence, whereas calm is generally neutral or slightly positive. This difference in emotional valence might be subtly reflected in the acoustic properties, such as the slight elevation in pitch for sadness.

As illustrated in Figure 8, the left panel displays the fitted pitch contours for calm (red line) and sadness (blue line) over 500 samples. Both contours demonstrate a similar downward trend, but the pitch level for sadness remains consistently higher than that of calm. The shaded areas represent the confidence intervals excluding random effects, showing the reliability of these observations. The right panel of Figure 8 presents the estimated difference in F0 between the calm and sad conditions, with the shaded area indicating the confidence interval of the difference. The fact that the confidence interval does not cross zero throughout the entire time window (0 to 525 samples) suggests that the difference in pitch level between calm and sadness is statistically significant across the utterance. These findings highlight the similarities between **calm** and **sadness** in terms of pitch contour shape—both exhibit a downward slope indicative of low arousal. However, the differences in overall pitch level provide an acoustic cue for distinguishing between the two emotions. The higher pitch in sadness may reflect a slight increase in emotional intensity or a different emotional valence compared to calm.

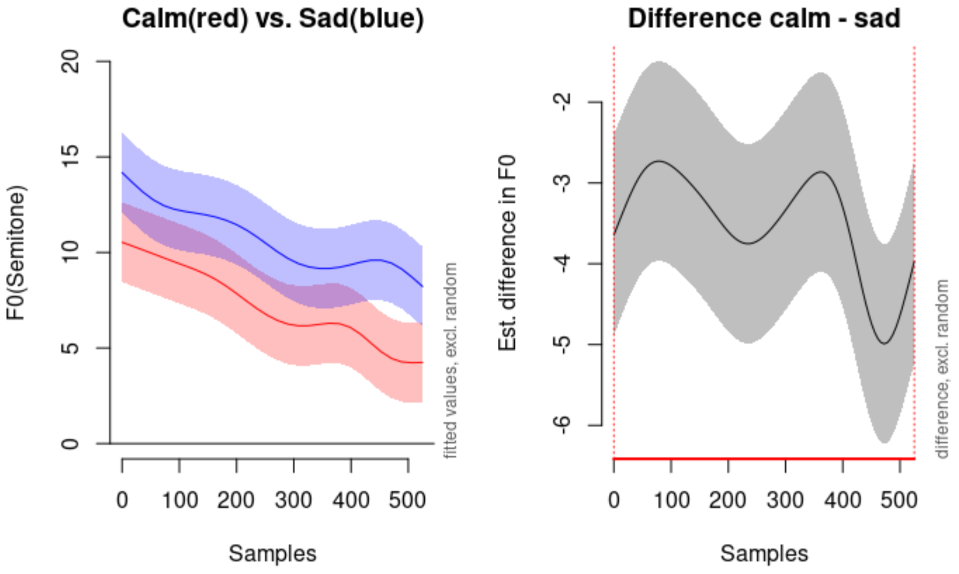

Figure 8. Comparison of pitch contours between Calm (red) and Sad (blue) speech samples. The left panel displays the fitted pitch values in semitones over 500 samples, with shaded areas representing the confidence intervals excluding random effects. The right panel shows the estimated difference in F0 (fundamental frequency) between the Calm and Sad conditions, with the shaded area indicating the confidence interval of the difference.

In our study, the pitch contours of **happiness** and **surprise** display distinct characteristics that aid in differentiating these emotions acoustically. The pitch contour of **happiness** begins at a mid-level pitch and exhibits a **steeper downward slope** compared to other emotions. This pattern aligns with previous research suggesting that happiness involves increased pitch variability and dynamic intonation patterns, reflecting a cheerful and expressive vocal demeanor. In contrast, the pitch contour of **surprise** is notably distinctive. As illustrated in Figure 9, surprise shows significant fluctuation throughout the utterance and features an **elevated pitch towards the end**, surpassing even other high-activation emotions like anger and fear. This elevation in pitch at the conclusion of the utterance may mirror the suddenness and heightened intensity typically associated with surprise. The dynamic rise in pitch could be indicative of an exclamatory expression, which is characteristic of how surprise is often vocally manifested.

The left panel of Figure 9 demonstrates the fitted pitch contours for both emotions. While both happiness and surprise start at similar pitch levels, their trajectories diverge significantly over time. Happiness maintains a relatively steady pitch before descending sharply, whereas surprise exhibits considerable fluctuation and culminates in a pronounced pitch increase at the end. The right panel of Figure 9 highlights the estimated difference in F0 between the two emotions. The areas where the confidence interval does not cross zero indicate time windows where the difference in pitch is statistically significant. These differences suggest that listeners may rely on the distinctive pitch patterns, particularly the end-of-utterance elevation in surprise, to differentiate it from happiness.

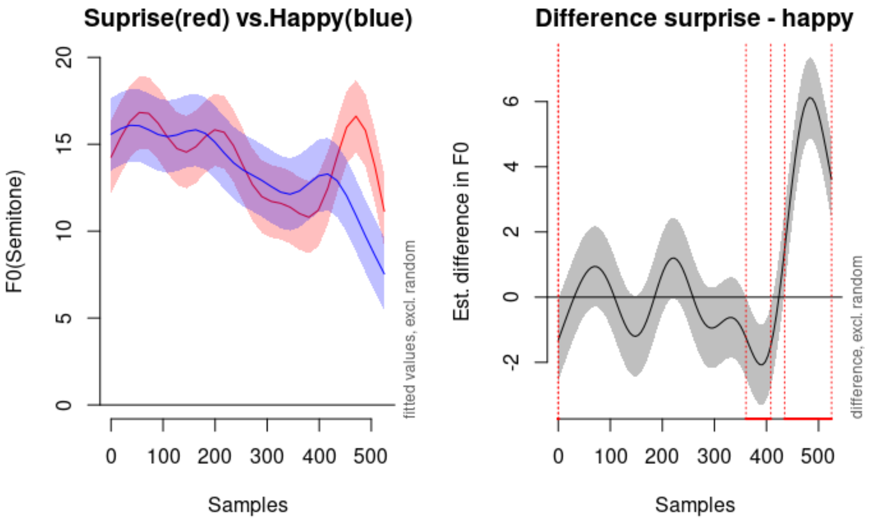

Figure 9. Comparison of pitch contours between Surprise (red) and Happy (blue) speech samples. The left panel displays the fitted pitch values in semitones over 500 samples, with shaded areas representing the confidence intervals excluding random effects. The right panel shows the estimated difference in F0 (fundamental frequency) between the Surprise and Happy conditions, with the shaded area indicating the confidence interval of the difference.

Our results emphasize the importance of considering both the **shape** and **level** of pitch contours in emotion recognition. The similarities in contour shape indicate that relying solely on the dynamic movement of pitch may not suffice for accurate classification between certain emotions. However, incorporating pitch level differences enhances the discriminative power of the model, enabling better differentiation between closely related emotional expressions like anger and disgust. The unique pitch elevation in surprise, contrasted with the steeper downward slope in happiness, provides acoustic cues that are critical for distinguishing between these two positive high-arousal emotions. This differentiation is crucial, as happiness and surprise can be easily confused due to their shared characteristics of high activation and positive valence.

Our use of Generalized Additive Mixed Models (GAMMs) allows us to capture nuanced acoustic differences among various emotional states. By modeling the dynamic F0 contours, we enhance the interpretability of the emotion classification system, providing insights into how specific acoustic features correspond to different emotions. This approach is particularly valuable for low-arousal emotions like **calm** and **sadness**, where acoustic differences are more subtle and require sophisticated modeling techniques to detect.

One of the challenges we encountered is the differentiation between **neutral** and **calm** emotions. According to our modeling, the neutral emotion is not statistically significantly different from calm emotion. As shown in **Figure 10**, which compares the predicted F0 contours for neutral and calm, the pitch contours are remarkably similar. The right panel of Figure 10 indicates with vertical red dots that the time window of significant difference is found merely between 116.67 to 121.97 milliseconds and 222.73 to 270.45 milliseconds—a very brief duration overall. The following observation from [20] provides context for our findings. The inclusion of both neutral and calm emotions in the RAVDESS dataset acknowledges the challenges performers and listeners face in distinguishing these states. The minimal acoustic differences captured by our GAMM analysis reflect this perceptual similarity, as noted in [2] (note that in the quotation author style citation style is replaced with numeric style citation):

"[T]he RAVDESS includes two baseline emotions, neutral and calm. Many studies incorporate a neutral or 'no emotion' control condition. However, neutral expressions have produced mixed perceptual results [1], at times conveying a negative emotional valence. Researchers have suggested that this may be due to uncertainty on the part of the performer as to how neutral should be conveyed [3]. To compensate for this, a calm baseline condition has been included, which is perceptually like neutral but may be perceived as having a mild positive valence. To our knowledge, the calm expression is not contained in any other set of dynamic conversational expressions." [20]

Figure 10 illustrates the modeled F0 contours and differences between neutral and calm emotions. The left panel displays the fitted pitch values in semitones over 500 samples, with shaded areas representing the confidence intervals excluding random effects. The contours for neutral (red line) and calm (blue line) are nearly overlapping, indicating highly similar pitch patterns. The right panel shows the estimated difference in F0 between the neutral and calm conditions, with the shaded area indicating the confidence interval of the difference. The vertical red dots mark the brief time windows where the difference is statistically significant.

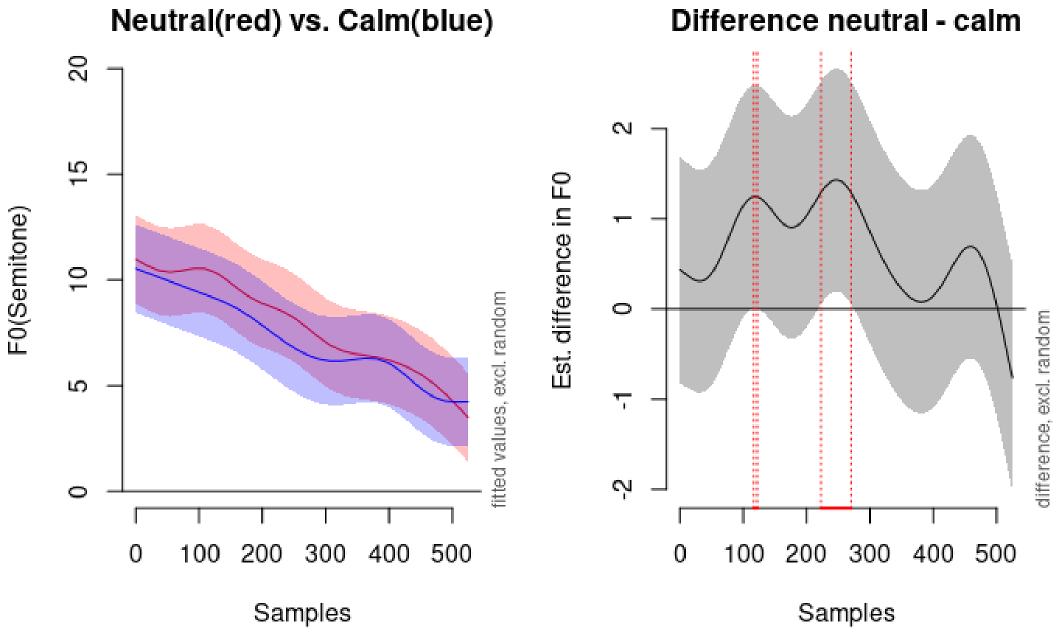

Figure 10. Comparison of pitch contours between Neutral (red) and Calm (blue) speech samples. The left panel displays the fitted pitch values in semitones over 500 normalized time points, with shaded areas representing the 95% confidence intervals (excluding random effects). The right panel shows the estimated difference in F0 (fundamental frequency) between the Neutral and Calm conditions, with the shaded area indicating the 95% confidence interval of the difference. Vertical red lines denote time windows where the difference is statistically significant.

Additionally, Figure 10 also includes comparisons of the modeled F0 contours for **angry** and **happy** emotions (top panel), demonstrating how our method effectively captures and illustrates distinctions between emotions with higher arousal levels. The clear differences in pitch contours and levels between angry and happy support the effectiveness of GAMMs in modeling and interpreting emotional speech across the arousal spectrum.

In conclusion, our use of GAMMs facilitates a deeper understanding of the acoustic properties of emotional speech. By capturing both prominent and subtle differences among emotions, especially in cases where traditional statistical methods might overlook minor yet significant variations, our approach enhances both the interpretability and the accuracy of emotion classification systems.

5. Conclusions

In this study, we employed a Generalized Additive Mixed Model (GAMM) to analyze the role of dynamic F0 contours in the classification of eight basic emotions using the RAVDESS corpus. The F0 values, extracted over the sonorant portions of speech at ten equidistant points, were concatenated across two predetermined sentences. By focusing on these controlled utterances, we aimed to isolate the effect of pitch dynamics on emotion recognition.

Our findings confirm previous observations about the informative role of pitch in expressing emotions. The dynamic modeling of F0 contours provided insights into the specific acoustic patterns associated with different emotional states. For instance, we observed that emotions such as anger, disgust, and fear exhibit downward pitch contours at elevated pitch levels, while happiness displays a steeper downward slope starting from a mid-level pitch. These nuanced differences highlight the significance of dynamic pitch features in distinguishing between emotions.

One of the key advantages of our approach is its interpretability. While deep learning-based automatic speech recognition systems have achieved higher accuracy in classifying emotion types—surpassing both human raters and our pitch contour-based modeling—they often function as "black boxes," offering limited insight into the features driving their performance. In contrast, our method allows for a more transparent understanding of why certain emotions are harder to distinguish than others. By directly modeling dynamic F0 contours, we can explain, for example, why emotions like calm and neutral are challenging to differentiate due to their similar pitch patterns.

However, our study also has limitations that warrant discussion. The analysis was conducted using predetermined sentences, which constrains the generalizability of our findings to a broader range of speech contexts. The use of controlled utterances, while beneficial for isolating specific acoustic features, may not capture the variability inherent in natural, spontaneous speech. This limitation suggests that caution should be exercised when extrapolating our results to more diverse linguistic environments.

Despite this constraint, focusing on fixed sentences provided a controlled setting to delve deeply into the role of dynamic F0 in emotion classification. It allowed us to attribute differences in emotion recognition specifically to pitch contours, minimizing the influence of lexical or syntactic variations. This approach contributes valuable insights into how dynamic pitch features function as acoustic correlates of emotional expression.

It is also important to acknowledge that pitch contour alone does not bear the entire burden of conveying expressive meaning in speech. Emotions are complex and multifaceted, often communicated through a combination of prosodic features such as intensity, duration, speech rate, and spectral qualities. While our study was solely restricted to the pitch contour variable—due to its significant role in arousing sensations in listeners and its previously underexplored modeling—it is clear that incorporating additional acoustic features could enhance emotion recognition systems.

Future research should consider expanding the scope of analysis to include other prosodic and spectral features. Integrating variables such as intensity, speech rate, and formant frequencies could provide a more comprehensive understanding of emotional speech. Additionally, applying dynamic modeling techniques like GAMMs to spontaneous speech samples or a wider variety of sentences could improve the generalizability of the findings and contribute to the development of more robust emotion recognition models. Additionally, including dynamic and multimodal properties could be an interesting area for future research. For example, [37] achieved an 80.08% accuracy in classifying eight emotions on the RAVDESS dataset using a multimodal emotion recognition system that combines speech and a facial recognition model.

In conclusion, our study demonstrates that dynamic F0 contours play a crucial role in emotion classification and that modeling these contours using GAMMs offers both interpretability and valuable insights into the acoustic properties of emotional speech. While the use of predetermined sentences presents limitations in terms of generalizability, it also serves as a strength by providing a controlled environment to explore the impact of pitch dynamics. Our approach underscores the importance of transparent and interpretable models in speech emotion recognition, paving the way for future studies to build upon these findings and develop more sophisticated, multimodal emotion recognition systems.

Funding: This work was supported by Sungshin Women's University Research Grant 2023 and by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A5A2A01061716).

Data Availability Statement: The RAVDESS database used in this paper is available under request from https://zenodo.org/record/1188976#.YTscC_wzY5k, accessed on 10 October 2023.

Acknowledgments: In this section, you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Scherer K.R; Banse R,; Wallbott H.G.; Goldbeck T. Vocal cues in emotion encoding and decoding. *Motivation and Emotion* **1991,** *15(2)*, pp. 123–48. https://doi.org/10.1007/BF00995674
2. Ekman, P. Are there basic emotions? *Psychological Review* **1992**, *99(3)*, pp. 550-553. https://doi.org/10.1037/0033-295X.99.3.550
3. Juslin P.N.; Laukka P. Communication of emotions in vocal expression and music performance: different channels, same code? *Psychol Bull* **2003**, *129(5)*, pp. 770–814. https://doi.org/10.1037/0033-2909.129.5.770
4. Yoon, S.; Son, G.; Kwon, S. Fear emotion classification in speech by acoustic and behavioral cues. *Multimed Tools Appl* **2019,** *78*, pp. 2345–2366. https://doi.org/10.1007/s11042-018-6329-2
5. Plutchik, R. A general psychoevolutionary theory of emotion, In *Theories of Emotion; Plutchik, R., Kellerman, H., Eds.; Academic Press, New York*, **1980**; pp. 3–33.
6. Russell, J.; Bachorowski, J.; Fernández-Dols, J. Facial and vocal expressions of emotion. *Annual review of psychology* **2003**, *54(1)*, pp. 329–349.
7. Tomkins, S. *Affect imagery consciousness: Volume I: The positive affects*; Springer publishing company, New York, 1962,
8. Ververidis, D.; Kotropoulos, C. Emotional Speech Recognition: Resources, Features, and Methods. *Speech Communication* **2006**, *48*, pp. 1162-1181. http://dx.doi.org/10.1016/j.specom.2006.04.003
9. Davis S.; Mermelstein P. Evaluation of acoustic parameters for monosyllabic word identification. *The Journal of the Acoustical Society of America* **1978**, *64(S1)*, S180-S181. https://doi.org/10.1121/1.2004059
10. Abdulmohsin, H. A new proposed statistical feature extraction method in speech emotion recognition. *Computers Electrical Engineering* **2021**, *93*, pp. 107172. https://doi.org/10.1016/j.compeleceng.2021.107172
11. Rodero, E. Intonation and Emotion: Influence of Pitch Levels and Contour Type on Creating Emotions. *Journal of Voice* **2011**, *25(1)*, pp. e25-e34. https://doi.org/10.1016/j.jvoice.2010.02.002.
12. Whissel, C. The dictionary of affect in language. In *Emotion: Theory, Research and Experience*; Plutchik, R. Kellerman, H. Eds.; Academic Press, New York, 1989; *Vol. 4*, pp.113-131.
13. Juslin P.N.; Laukka P. Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. *Emotion* **2001**, *1(4)*, pp. 381–412. https://doi.org/10.1037/1528-3542.1.4.381
14. Bänziger, T.; Scherer, K.R. The role of intonation in emotional expressions. *Speech Communication* 2005, *46(3-4)*, pp.252-267. https://doi.org/10.1016/j.specom.2005.02.016

15. Hirose, K.; Sato, K.; Asano, Y.; Minematsu, N. Synthesis of F0 contours using generation process model parameters predicted from unlabeled corpora: application to emotional speech synthesis, *Speech Communication* 2005, *46(3-4)*, pp. 385-404. https://doi.org/10.1016/j.specom.2005.03.014
16. Morrison, G.S. L1-Spanish Speakers' Acquisition of the English /i/—/I/ Contrast: duration-based perception is not the initial developmental stage. *Language and Speech* **2008,** *51(4)*, pp. 285-315. https://doi.org/10.1177/0023830908099067
17. Wood, S.N. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, New York, **2017**. https://doi.org/10.1201/9781315370279
18. van Rij, J. Overview GAMM analysis of time series data. Available online: http://www.sfs.uni-tuebingen.de/~jvanrij/Tutorial/GAMM.html. (Accessed on 23 Oct. 2023).
19. Chuang, Y.; Fon, J.; Papakyritsis, I.; Baayen, H. Analyzing phonetic data with generalized additive mixed models. In *Manual of clinical phonetics*; Ball, M. Ed.; Routledge: London, **2021,** pp.108-138. https://doi.org/10.4324/9780429320903
20. Livingstone, S.R.; Russo, F. A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS ONE* **2018,** *13(5)*, e0196391. https://doi.org/10.1371/journal.pone.0196391
21. Jadoul, Y.; Thompson, B.; de Boer, B. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics* **2018**, *71*, 1-15. https://doi.org/10.1016/j.wocn.2018.07.001
22. McFee B.; Raffel C.; Liang D.; Ellis, D.; McVicar, M.; Battenbert, E.; Nieto, O. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, Austin, Texas, 6-12 July, **2015**.
23. Buschmeier H.; Wlodarczak M. TextGridTools: A TextGrid processing and analysis toolkit for Python. In *Tagungsband der 24. Konferenz zur elektronischen sprachsignalverarbeitung (ESSV 2013)*, Bielefeld, German, 26-28 Mar, **2013**.
24. Boersma, P.; Weenink, D. Praat: doing phonetics by computer [Computer program Version 6.1.38]. Available online: http://www.praat.org/ (accessed on 2 January 2021)
25. McFee B. resampy: efficient sample rate conversion in Python. *Journal of Open Source Software* **2016**, *1(8)*, pp. 125. https://doi.org/10.21105/joss.00125
26. van Rij, J.; Wieling, M.; Baayen, R. H.; van Rijn, H. itsadug: Interpreting Time Series and Autocorrelated Data using GAMMs. **2022**; R package version 2.4.1
27. Stuart-Smith, J.; Lennon, R.; Macdonald, R.; Robertson, D.; Sóskuthy, M.; José, B.; Evers, L. A dynamic acoustic view of 528 real-time change in word-final liquids in spontaneous Glaswegian. In Proceedings of the 18th International Congress of Phonetic Sciences, Glasgow, U.K., 10-14 August **2015**.
28. Wieling, M. Analyzing dynamic phonetic data using generalized additive mixed modeling: a tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics* **2018**, *70*, pp. 86-116.
https://doi.org/10.1016/j.wocn.2018.03.002
29. Baayen H.; Vasishth S.; Kliegl R. et al. The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language* **2017**, *94*, pp. 206-234. https://doi.org/10.1016/j.jml.2016.11.006
30. Sóskuthy M. Evaluating generalised additive mixed modelling strategies for dynamic speech analysis. *Journal of Phonetics* **2021**, *84*, pp. 101017. https://doi.org/10.1016/j.wocn.2020.101017
31. Winter, B.; Wieling, M. How to analyze linguistic change using mixed models, Growth Curve Analysis and Generalized Additive Modeling. *Journal of Language Evolution* **2016**, *1(1)*, pp. 7–18. https://doi.org/10.1093/jole/lzv003
32. Shegokar, P.; Sircar, P. Continuous wavelet transform based speech emotion recognition. In 2016 10 International Conference on Signal Processing and Communication Systems (ICSCS), IEEE, Surfers Paradise, Australia, 19-21 Dec. **2016**, p. 1-8. https://doi.org/10.1109/ICSPCS.2016.7843306
33. Zhang, B.; Provost, E.M.; Essi, G. Cross-corpus acoustic emotion recognition from singing and speaking a multi-task learning approach. In 2016 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Shanghai, China, 20-25 Mar, 2016, pp. 5805-5809. https://doi.org/10.1109/ICASSP.2016.7472790
34. Zeng, Y.; Mao, H.; Peng, D.; Yi, Z. Spectrogram based multi-task audio classification. *Multimed. Tools. Appl*. **2019**, 78, 3705-3722. https://doi.org/10.1007/s11042-017-5539-3

doi:10.20944/preprints202409.2449.v1

35. Popva, A.S.; Rassadin, AG.; Ponomarenko, A.A. Emotion recognition in sound. In International Conference on Neuroinformatics. Moscow, Russia, 2-6 October 2017, pp. 117-124.
36. Issa, D. M.; Yazici, A. Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control* **2020,** 59, 101894. https://doi.org/10.1016/j.bspc.2020.101894
37. Luna-Jiménez, C.; Griol, D.; Callejas, Z.; Kleinlein, R.; M. Montero, J.; Fernández-Martínez, F. Multimodal emotion recognition on RAVDESS dataset using transfer learning. *Sensors* **2021**, *21(22)*, pp. 7665. https://doi.org/10.3390/s21227665

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.