Longitudinal Phonetic Adaptation in YouTube BookTube Creators

Atrey Desai
University of Maryland, College Park, USA
adesai10@terpmail.umd.edu

1 Introduction

The widespread use of social media on the Internet has created distinctive subgroups with their own traits. In this distinct environment, content creators must adapt their presentation styles to achieve greater audience engagement. This paper investigates the existence and trends of this phenomenon in the YouTube book review and commentary community, colloquially known as "BookTube."

Prior research on short-form video platforms such as TikTok has found evidence of an "influencer voice," with distinct phonetic features such as vocal fry and exaggerated pitch variation (Adomaitis et al., 2024). Increased adherence to phonetic features was also correlated with greater video success. However, on YouTube, the BookTube community has several dissimilar features: it features long-form content with videos generally over 20 minutes, more detailed commentary, and greater variety in literary discussion.

By comparing BookTube creators' vocal patterns in videos at their channel start and the present, this study tests whether content creators adapt toward informal speech patterns in response to engagement metrics and broader platform norms.

2 Background

Prior research by Allan Bell proposed that speakers systematically adjust their linguistic style based on perceived characteristics of the audience (Bell, 1984). This form of code-switching existed before the internet, but social media allows researchers to easily collect data on how an individual may adjust their speaking patterns based on the platform they are speaking on.

Recent platform-specific work has corroborated prior findings: TikTok videos with higher rates of vocal fry, uptalk, and pitch variation resulted in higher engagement (Adomaitis et al., 2024; Murashima, 2024). This suggests the platform's young audience rewards more informal features.

In contrast, analysis of popular YouTube content creators found wider pitch ranges and fewer pauses, conveying authority and energy (Beck, 2015; Berger, 2024). This distinguishes creators from the "casual" affect of short-form video. The desire to appeal to a broader audience may also lead to phonetic changes; specifically, YouTubers reduce regional or informal markers in scripted content (Lee, 2017). Vocal fry, common in casual speech, is also often stigmatized in professional contexts, being perceived as "hesitant" (Anderson et al., 2014).

As BookTube creators establish a channel identity, exposure to the broader niche and YouTube community may lead to unconscious suppression of vocal fry and homogenization of vocal features to signal competence and credibility. Existing research has favored cross-sectional designs comparing different creators rather than tracking individuals over time. Therefore, this paper conducts a longitudinal study to determine if individual YouTube creators show measurable phonetic evolution as they gain experience. This connects to the larger theories of the emergence of platform-specific "influencer" voices through audience design processes.

3 Hypotheses

I hypothesize that YouTube BookTube creators will deliberately or indirectly adapt their phonetic features to align with the perceived norms and demographics of the platform. Specifically, they will have an increase in vocal fry percentage, increased pitch ($F0$) variation, and a higher mean pitch. Conforming to these norms will positively correlate with engagement rates (likes-to-views ratio).

4 Methods

To minimize the effect of confounding variables, samples were drawn from white women living in

Table 1: Descriptive Statistics and Significance Tests. Of note, engagement rate increased significantly while vocal fry decreased.

Measure	**Early Period**	**Late Period**	**% Change**	**Sig.**
	(Mean ± SD)	(Mean ± SD)		(Paired t-test)
Mean F0 (Hz)	216.5 ± 19.4	211.4 ± 17.1	-2.4%	$p = 0.30$ (n.s.)
F0 Range (Hz)	290.9 ± 23.0	319.3 ± 26.4	$+9.8\%$	$p = 0.030$ (*)
F0 SD (Hz)	37.2 ± 4.2	39.3 ± 4.0	$+5.8\%$	$p = 0.21$ (n.s.)
Vocal Fry (%)	15.0 ± 8.7	8.7 ± 3.9	-42.0%	$p = 0.015$ (*)
Engagement Rate	4.38 ± 1.23	7.31 ± 1.93	$+67.0\%$	$p = 0.001$ (**)

* $p < 0.05$, ** $p < 0.01$

the United States who began their YouTube channel in their early 20s and are currently in their mid-to-late 20s. These demographics are representative of the mainstream English-language BookTube community (Murray, 2025). Channels were then filtered for the following characteristics: primarily book-related content, solo-speaker format without voiceovers, a direct address style, a consistent environment with no background music, and a regular upload schedule.

For each channel, I collected 10 videos, split equally from their earliest and most recent uploads. Segments were then filtered for uninterrupted solo speech and to exclude introductions, conclusions, and any extraneous audio, such as music or sound effects, that could confound phonetic analysis.

4.1 Measurements

All phonetic measurements were conducted using PRAAT (v6.4.42) (Boersma, 2001). For pitch analysis, a Pitch object (pitch floor: 75Hz, pitch ceiling: 500Hz) was created. These parameters were recommended by the PRAAT manual and manually verified using spectrograms of audio samples (Boersma). The following measurements were then extracted: mean $F0$, minimum $F0$, maximum $F0$, and $F0$ standard deviation.

For vocal fry measurements, each audio segment was opened in the PRAAT editor using the spectrogram display. Vocal fry was identified based on a low and irregular pitch track, widely spaced glottal pulses, and a creaky waveform. Pulses and pitch were also turned on to assist with identification. For quality control purposes, 10% of randomly selected segments were verified for consistency in vocal fry annotation and pitch tracking errors.

For each video, the following publicly available YouTube data was collected: view count, like count, and comment count. The engagement rate was calculated as $\left(\frac{\text{Likes}}{\text{Views}}\right) \times 100$.

The mean and standard deviation for each phonetic feature grouped by Early (the earliest 5 segments) versus Late (the recent 5 segments) were used to calculate paired-sample t-tests with a significance threshold of p < 0.05. Also, the Pearson correlation coefficients between phonetic features and engagement rates were calculated based on all 100 segments. The statistics are presented in table 1.

Download and video segmentation Python files, PRAAT scripts used to speed up the extraction of annotated TextGrids, the script for automated YouTube API querying, and downloaded YouTube segments are publicly available for replication in §8.

5 Materials

The audio corpus consists of 100 audio segments (10 channels each with 10 segments) from YouTube BookTube channels specializing in book reviews and recommendations. These creators were selected due to a consistent topic and demographics, as elaborated in §4, to control for feature-based phonetic variation that could confound temporal analysis. YouTube's publicly available engagement metrics are appropriate as an indicator of audience response. The 120-second segment length also provides enough speech sample for reliable pitch and vocal fry metrics, while remaining manageable for one person to manually annotate across 100 segments.

5.1 Tools

- PRAAT 6.4.42: Standard phonetic analysis software for acoustic measurement (Boersma, 2001)
- Google Sheets: Statistical calculations and

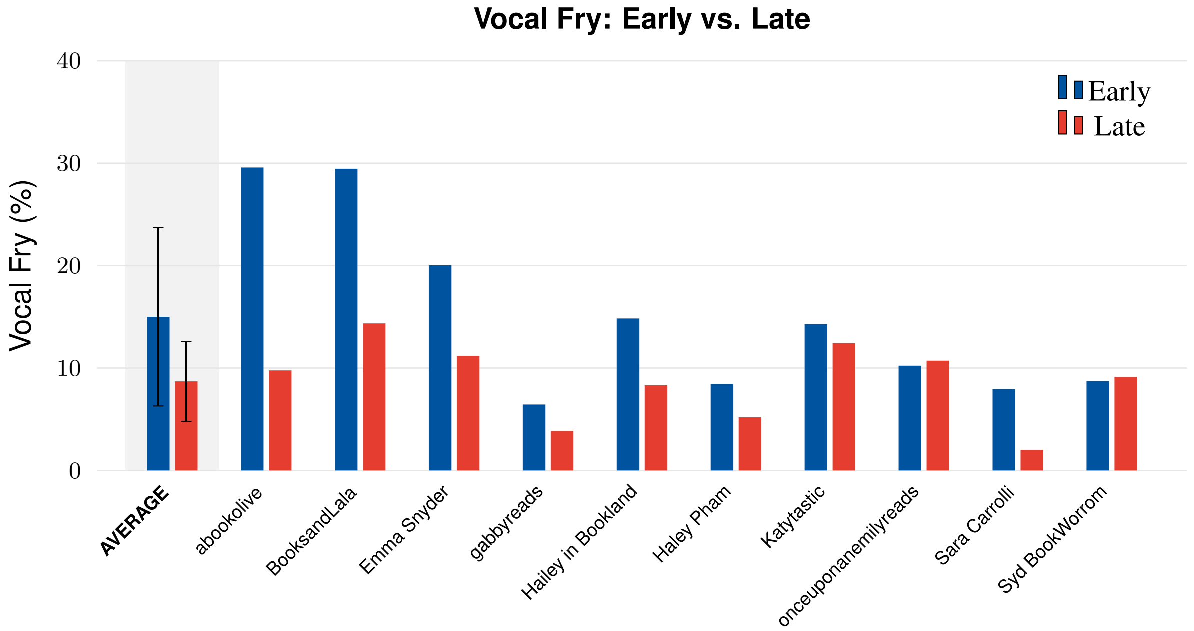

Figure 1: Comparison of Vocal Fry usage percentages between Early and Late career stages. The shaded column represents the average.

data visualization

- YouTube: Online video sharing platform

6 Results

Measurements were collected from 10 distinct YouTube channels across two time periods (Early vs. Late), yielding a total of 100 analyzed segments. Table 1 summarizes the descriptive statistics and significance tests for all acoustic and engagement measures.

6.1 Phonetic Analysis

The greatest observed change was the decrease in vocal fry between the two periods. On average, in the early period, content creators exhibited vocal fry in $15.00 \pm 8.66\%$ of analyzed speech, which dropped to $8.70 \pm 3.92\%$ in the Late period ($t(9) = 2.99, p = 0.015$). This represents a 42% reduction between the two time periods.

Pitch dynamics also showed a statistically significant change. The $F0$ Range increased by 9.8%, from a mean of 290.9 Hz to 319.3 Hz ($t(9) = -2.58, p = 0.030$). These findings are in line with earlier research that found creators adopt a more dynamic speaking style as they adapt to voice characteristics of popular YouTube creators (Berger, 2024).

However, other pitch metrics remained relatively stable. There was no statistically significant difference in Mean $F0$ (216.5 Hz vs. 211.4 Hz, $p = 0.30$) or $F0$ Standard Deviation (37.2 Hz vs. 39.3 Hz, $p = 0.21$). Double bar charts showing a breakdown in the changes on a per-channel basis are available in §8.

6.2 Engagement and Correlations

In addition, engagement rates increased dramatically between periods, rising from 4.38 ± 1.23 to 7.31 ± 1.93 ($t(9) = -4.75, p < 0.001$).

Pearson correlations between phonetic features and engagement (calculated across the 20 available data points of the aggregated data) provide some detail on this increase. There was a moderate negative correlation between mean $F0$ and engagement ($r = -0.46$) and a weak positive correlation between $F0$ range and engagement ($r = 0.22$). Generally, this suggests videos with a lower pitch but higher range received better engagement. There was no correlation between $F0$ Standard Deviation and engagement (r=0.00). Importantly, there is a weak negative correlation between vocal fry and engagement ($r = -0.23$), aligning with the finding that vocal fry decreased over time. Scatterplots showing these relationships are in the §8.

7 Conclusions

This paper determined the existence and quantified changes in vocal patterns in response to broader platform norms and channel engagement metrics by YouTube BookTube creators over time. These findings support part of the hypothesis, but also differ compared to patterns in other platforms such as TikTok.

7.1 Hypothesis Resolution

These findings suggest YouTube BookTube creators may not observe the same phonetic patterns as creators on other platforms. Contrary to the initial hypothesis, the data show vocal fry decreases over time, with a 42% drop between the early and late periods ($p = 0.015$), meaning BookTube creators tend to professionalize and inhibit their natural speech patterns, instead of becoming more informal. This is also supported by the negative correlation between vocal fry and engagement, suggesting the long-form nature of YouTube may reward more formal and academic literary content. Other aspects of the hypothesis were partially correct. While the $F0$ range increased significantly, the mean $F0$ and $F0$ standard deviation showed no significant change. There was also a 67% increase in engagement rates ($p < 0.001$) between the two periods, suggesting that creators benefit from adapting to viewer preferences, though overall, the individual correlations between phonetic features and engagement are weak and necessitate further investigation.

7.2 Limitations

The primary limitation is the limited and homogeneous sample of content creators. While white American women in their 20s are a significant portion of the current BookTube community, controlling for confounding variables limits generalization to the entire niche. Research building on this work should try to use different samples of creators (different races, male creators, etc.) to confirm similar patterns.

Also, the current analysis is heavily reliant on engagement as a singular entity. Future work should move beyond one calculation and disaggregate engagement metrics such as likes, comments, watch time, and more. This can allow for more specialized analysis to see if different vocal features are relevant for different audience responses.

There may also be interest in qualitative data to see deliberate patterns of change in content creators over time. This can be done through conducting interviews with content creators about changes in their acoustic patterns and general trends in the BookTube community.

Finally, this study was limited to one platform because there was a lack of content creators who published across multiple platforms with different recorded videos. A comparative study of creators over platforms prioritizing different engagement lengths, such as TikTok and Instagram versus a podcast, would create firmer evidence of individuals code-switching based on platform-specific audience expectations.

References

Natalia Adomaitis, Lam Hoang, Maryam Shama, Sydney Trieu, and Kristina Zhao. 2024. The tiktok influencer voice: Do sociolinguistic features influence the success of tiktok videos?

Rindy C. Anderson, Casey A. Klofstad, William J. Mayew, and Mohan Venkatachalam. 2014. Vocal fry may undermine the success of young women in the labor market. *PLOS ONE*, 9(5):e97506.

Julie Beck. 2015. Why do so many people on youtube sound the same?

Allan Bell. 1984. Language style as audience design. *Language in society*, 13(2):145–204.

Stephanie Berger. 2024. *"Like, comment, subscribe": Perception of acoustic-prosodic features of content creators' charismatic speech on YouTube*. Ph.D. thesis, Christian Albrecht University of Kiel.

Paul Boersma. Intro 4.2. configuring the pitch contour.

Paul Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

Sarah Lee. 2017. Style-shifting in vlogging: An acoustic analysis of "youtube voice". *Lifespans and Styles*, 3(1):28–39.

Claire Murashima. 2024. Do you have "tiktok voice"? it's ok if you don't want to get rid of it. *NPR*.

Simone Murray. 2025. *The Problem of Affect: Literary Studies, BookTube, and BookTok*, 1 edition, page 99–133. Oxford University PressOxford.

8 Appendix

8.1 Public Link to Data

The data and scripts are accessible at this link.

8.2 Other Graphs

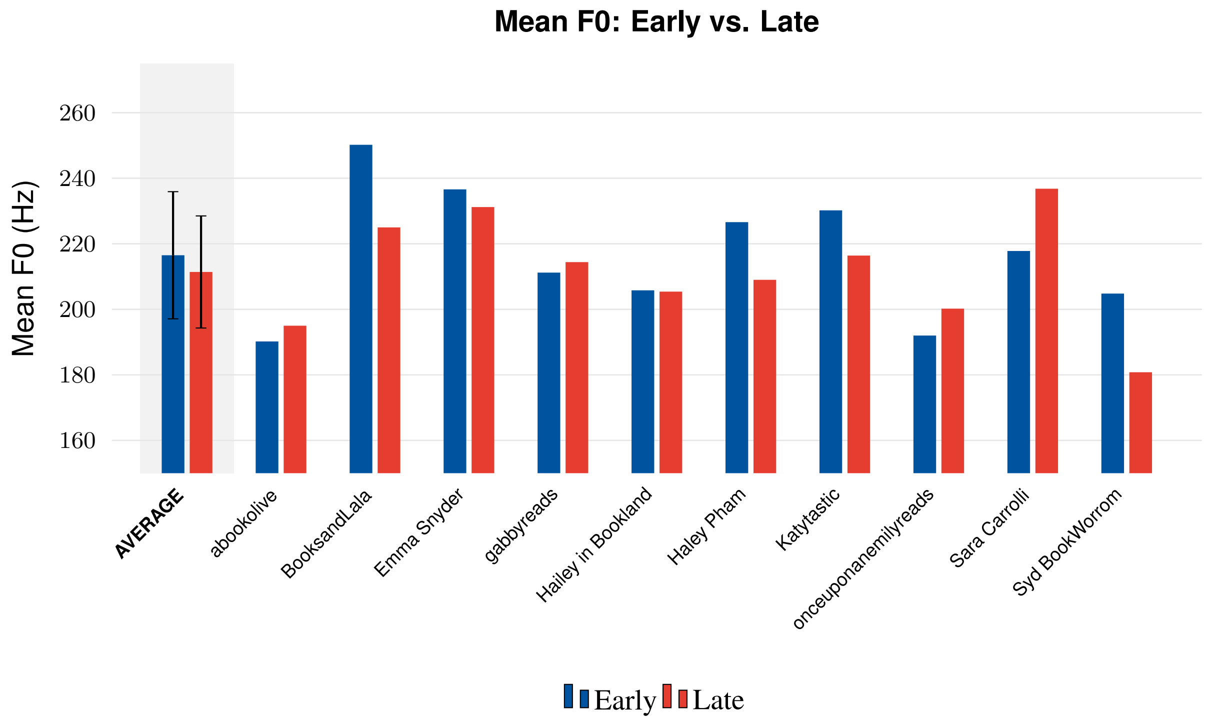

Figure 2: Comparison of Mean F0 (fundamental frequency) between Early and Late periods across creators. The difference is not statistically significant (n.s.). Error bars represent standard deviation.

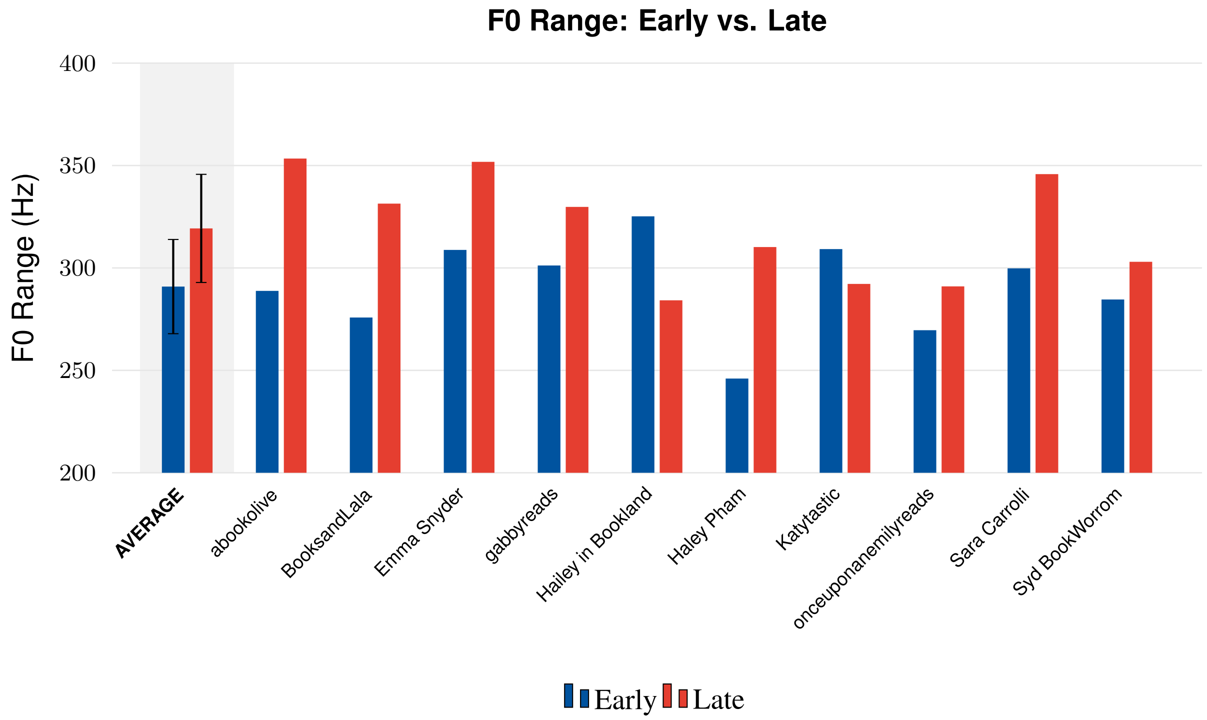

Figure 3: Comparison of F0 Range. There is a statistically significant increase ($p < 0.05$) in pitch range in the late period.

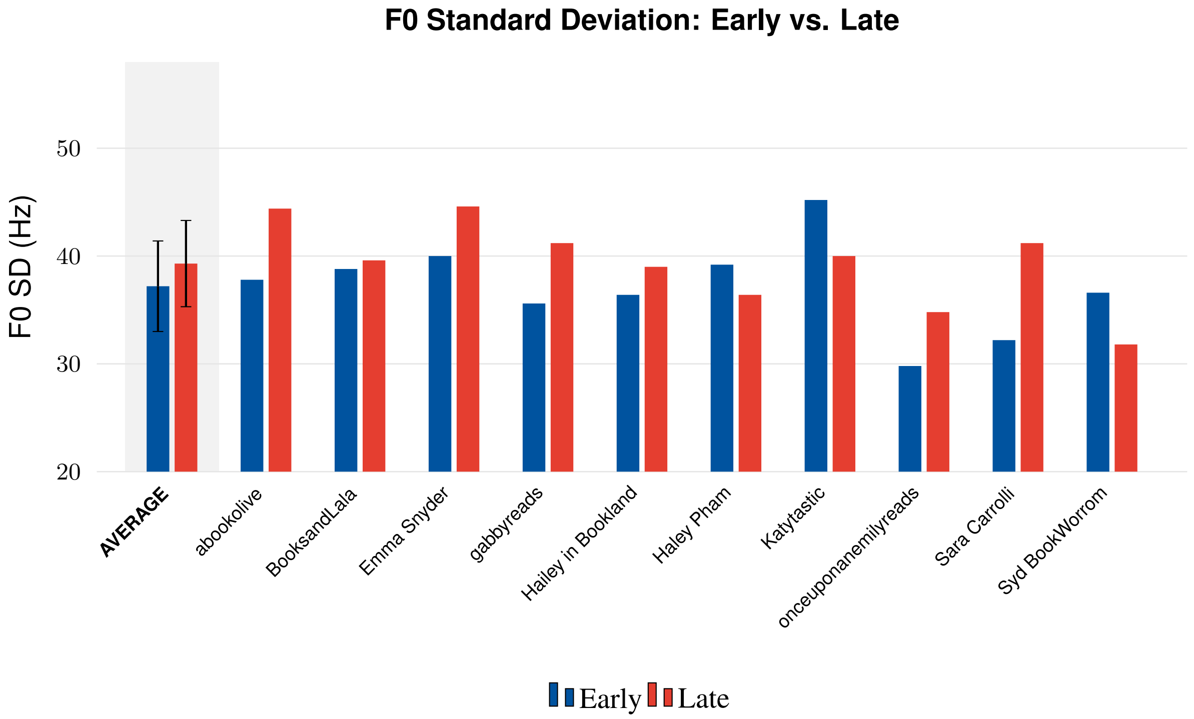

Figure 4: Comparison of F0 Standard Deviation (pitch variability). The change between periods is not statistically significant.

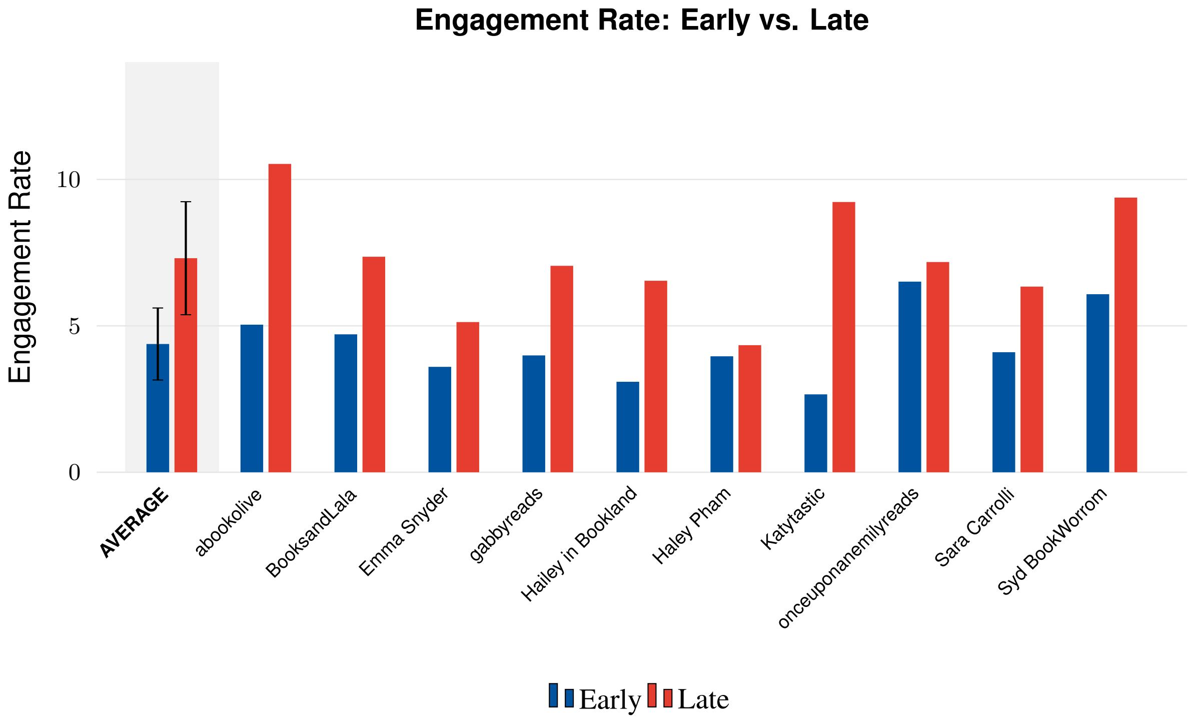

Figure 5: Comparison of average Engagement Rate. There is a highly significant increase ($p < 0.01$) in engagement in the late period.

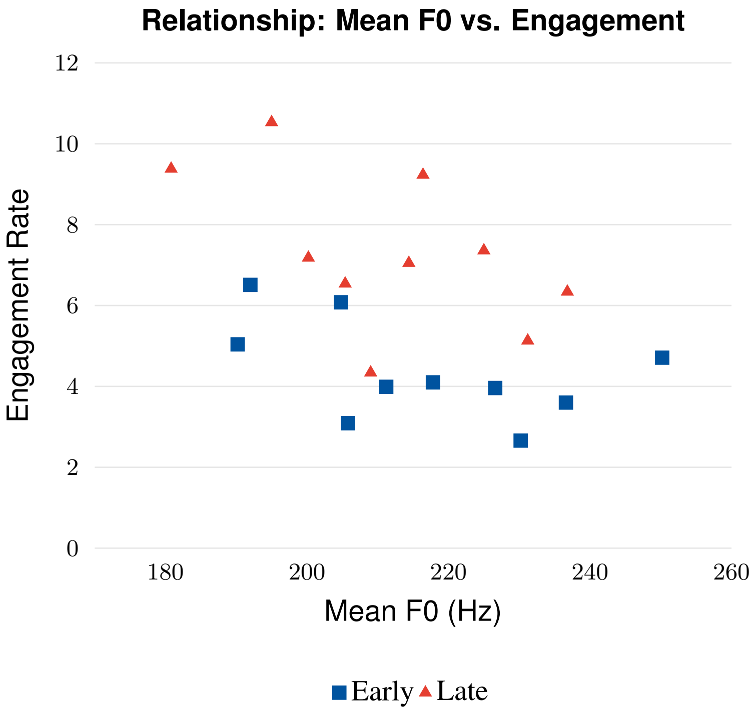

Figure 6: Scatter plot showing the relationship between Mean F0 and Engagement Rate across both periods.

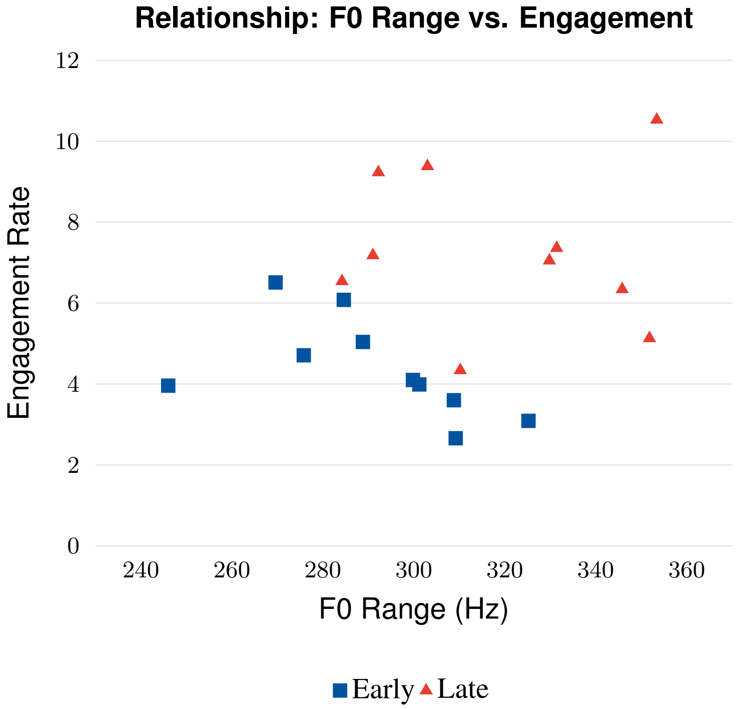

Figure 7: Scatter plot showing the relationship between F0 Range and Engagement Rate.

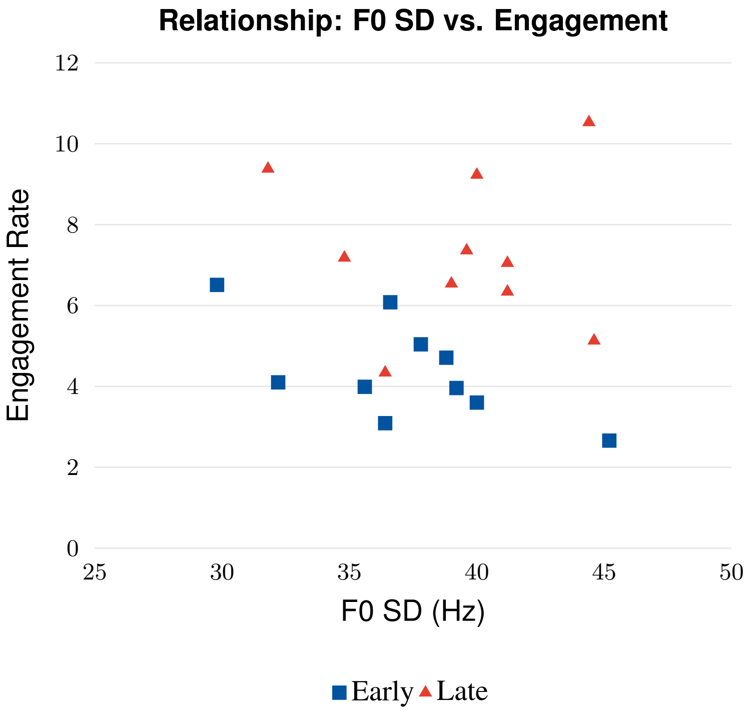

Figure 8: Scatter plot showing the relationship between F0 Standard Deviation and Engagement Rate.

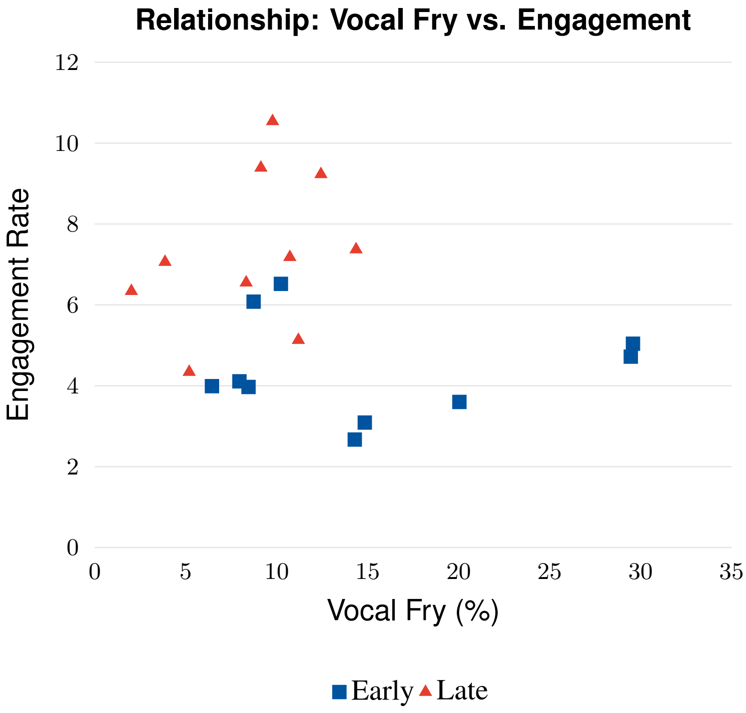

Figure 9: Scatter plot showing the relationship between Vocal Fry and Engagement Rate.