"YOUTUBE SPACE"—A PRELIMINARY INVESTIGATION OF VOWEL SPACES AND CHARISMA PERCEPTION ON YOUTUBE

Stephanie Berger[1], Margaret Zellers[1], Oliver Niebuhr[2]

[1]ISFAS, Kiel University, Germany [2]CIE, University of Southern Denmark, Denmark
sberger@isfas.uni-kiel.de, mzellers@isfas.uni-kiel.de, olni@sdu.dk

ABSTRACT

This study investigates correlations between vowel space size (measured as the mean formant ranges of the first two formants F1 and F2) and listener ratings of charisma and charisma-adjacent features (*authentic*, *enthusiastic*, *likeable*, and *persuasive*). The results show that for stimuli from a perception experiment, speakers with larger F1 ranges are perceived as more charismatic, but at the same time less authentic. The opposite is the case for a full 50-minute data set: For this data set, speakers with a smaller F1 range were perceived as more charismatic. The results differ from previous research suggesting that charisma on YouTube and other genres like business speeches may work differently.

Keywords: YouTube, charisma, perception, formants, vowel space

1. INTRODUCTION

Charisma (see [1] for an overview of the concept of charisma), i.e. the ability to draw in an audience, is incredibly important for online content creators like YouTubers. In a digital world where billions of videos are only one click away, audiences need to be kept engaged to stay and come back to a channel.

Phonetic charisma research has identified a large number of prosodic features that increase perceived charisma: a wider pitch range, faster speech rate, frequent use of emphatic accents, or a frequent use of high pitch accents, among others (e.g. [2, 3, 4, 5, 6, 7, 8, 1, 9]). News outlets that popularized the term "YouTube Voice" also include "overstressed vowels" in their assessment of the particular speaking style in YouTube vlogs [10, 11]. (Vlogs are video blogs in which a YouTuber speaks to the audience. This is semi-spontaneous speech, as it is not scripted, but planned and edited.)

The prosodic features mentioned above (as well as others, see [8] for an overview of features investigated for charismatic speech) can be connected to increased vocal effort, which in turn tends to be correlated with increased charisma in business and political speakers [1]. Increased vocal effort and increased charisma go hand in hand with clear, minimally reduced speech, and "clear speech includes 'an expansion of the vowel space"' [12, p.346]. [13] also found that speakers in psychological distress (e.g. depression) speak with significantly reduced vowel spaces compared to healthy speakers, which they correlate with decreased expressivity. That suggests the opposite—expanded vowel spaces—goes in hand with increased expressivity.

Furthermore, [14] studied differences in vowel space size and mean ranges of the first and second formants (F1 and F2) in L2 English produced by L1 German speakers. Listeners rated the speakers on five attributes: *charismatic*, *passionate*, *decided*, *trustworthy*, and *captivating*. The study found significant positive correlations between the listener ratings of all attributes and speakers' F1 and F2 ranges. This means that a larger formant range on either vowel space dimension was rated higher [14]. Large vowel spaces were also perceived as most charismatic, small vowel spaces as least charismatic by listeners. Additionally, when the vowel space was compressed in the F2 (back–front) dimension, speakers were perceived as less trustworthy and less decided, while they were perceived as less passionate and captivating if the F1 (closed–open) dimension was small [14]. Thus, in [14], the F1 range was more relevant for attributes related to emotions, while the F2 range was more closely related to attributes connected to cognition.

2. QUESTIONS AND HYPOTHESES

The current study investigates correlations between charisma ratings from a perception experiment and vowel space size, which is operationalized as the mean ranges of F1 and F2 (see Section 3.3). This study, similar to [14], includes a direct charisma rating (*charismatic*), as well as emotion-related (*likeable* and *enthusiastic*) and cognition-related attributes (*persuasive* and *authentic*). While [14] investigated business-oriented L2 English speakers, the current study investigates L1 English speaking

YouTubers who entertain their audience.

This study addresses three research questions (RQ). Do the F1 and F2 ranges (as measurements of vowel space size) correlate with the ratings of charisma and charisma-adjacent attributes (RQ1)?

Are correlations between ratings and vowel spaces similar for YouTubers and the speakers from a business context from previous literature [12, 14] (RQ2)? Differences between YouTubers and other speaker groups would indicate that charisma may be encoded depending on the medium. First indications of differences between acoustic features of charismatic speech in business and on YouTube exist: e.g., pitch range is important for charisma in business [1], but it did not correlate with the video view count (considered an approximation of charisma) in [15].

In addition, we compare the formant ranges from the entirety of annotated speech material (50 minutes) to those from the stimuli used to collect the perception ratings: Do similar correlations arise between charisma ratings and vowel space (F1 and F2 ranges) differ between the full data set and the perception experiment stimuli (RQ3)?

Four hypotheses are connected to these research questions. In line with [14], larger vowel spaces (in terms of larger F1 and F2 ranges) are expected to be perceived as more charismatic (H1). Larger F1 ranges are expected to be positively correlated with higher ratings for emotion-related attributes (*likeable* and *enthusiastic*, H2). Larger F2 ranges are expected to correlate positively with higher ratings of cognition-related attributes (*persuasive* and *authentic*, H3). We also predict that the direction of correlations matches between the two data sets (H4).

3. METHODS

3.1. Speech material and measurements

The first data set (*full data set*, *FData*) consists of 50 minutes of speech material from ten English-speaking YouTubers[1] (5 male, 5 female, 5 from North America, 5 from England, age at time of video publication: 24 to 32, mean=29.1, SD=2.64). All videos are vlogs where the speaker talks to the audience about different topics like mental health, the YouTube business, or their community; only monolog passages were used for the current study.

F1 and F2 were extracted at vowel midpoints using a Praat [16] script based on [17] using Praat's default settings for its `To Formant (robust)` function with the maximum formant at 5000 Hz for male and 5500 Hz for female speakers. The segment boundaries and their SAMPA annotations were created with WebMAUS [18] and manually corrected. Four corner vowels of the vowel space (IPA: [iː], [uː], [æ] and [ɒ]) are included in the study. Formants were normalized using the `normLobanov()` function in the `phonR`-package [19] to account for the different origins and genders of the speakers and the different recording set-ups of each YouTuber.

Mean F1 and F2 of each of the four vowels were calculated per speaker. The F1 means of [iː] were subtracted from those of [æ] for the front dimension, and [uː] was subtracted from [ɒ] for the back dimension. Afterwards, the mean of these differences was calculated to obtain a mean F1 range for the center height of the vowel space. For F2, [uː] was subtracted from [iː] for the closed dimension, and [ɒ] was subtracted from [æ] for the open dimension to get the center width of the vowel space.

The same calculations were made for a second data set (a subset of *FData*: *stimuli data set*, *SData*). It comprises four interpausal units for each speaker which were stimuli used in perception experiments (see Section 3.2). The stimuli were not initially chosen to study vowel spaces, so some substitutions were necessary: For speaker ZS, the [iː] was exchanged for /ɪ/ to obtain a complete vowel space with four corner points. Speaker PL was excluded for the *SData*, as only two of four corner points were available in the experiment stimuli and no substitutions were possible.

3.2. Perception experiment

Stimuli (both manipulated in pause duration and breathing as well as unmanipulated) were presented to participants in perception experiments and were rated on 5-point Likert scales. 20 participants (male: N=8, female: N=11, prefer not to answer: N=1, age: mean=27.5, SD=5.07) rated the stimuli in terms of charisma directly. 20 other participants (male: N=8, female: N=11, non-binary: N=1, age: mean=29.95, SD=5.23) rated the stimuli on charisma-adjacent attributes (*authentic*, *likeable*, *enthusiastic*, *persuasive*). All participants were L1 English speakers, originally from the British Isles. None reported issues with their hearing. The mean ratings of the unmanipulated stimuli per speaker and scale are used for the correlations, both with the vowel spaces of *FData* and *SData*.

3.3. Statistical analyses

Kendall's τ rank correlation coefficients were calculated between formant ranges and mean ratings. The significance level is set at α=.05. All statistical analyses were run in RStudio [20, R version 4.1.3].

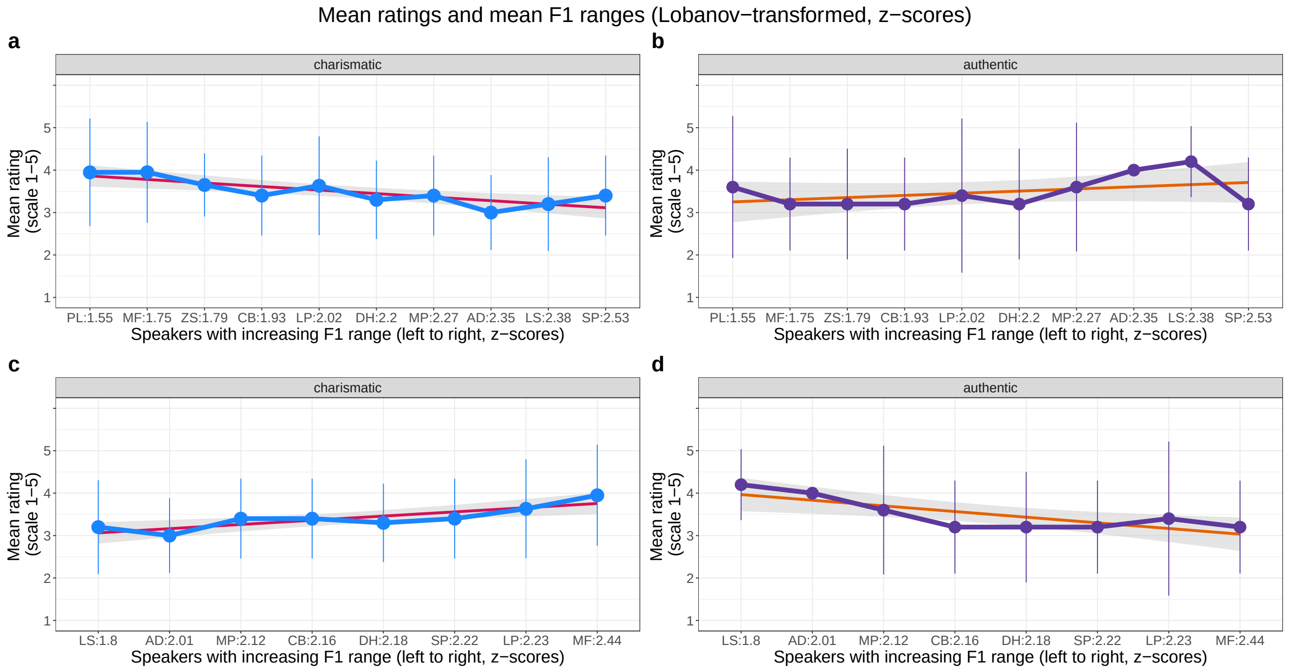

Figure 1: Data points and correlations between a) charismatic and b) authentic ratings and mean F1 range for the full data set, and c) charismatic and d) authentic ratings and mean F1 range for the stimuli. The speakers and their respective F1 ranges on the X-axis are arranged in ascending order. Whiskers represent standard deviation.

4. RESULTS

4.1. Full data set (*FData*)

There is a significant negative correlation between the *charismatic* ratings and the mean F1 range (τ = -0.64, *p* = .01). Speakers that have the largest F1 ranges in their videos overall are rated as less charismatic in the experiment (see Figure 1a). There are no other significant correlations for mean F1 range. Visually, there is also a tendency for a positive correlation for the *authentic* ratings (Figure 1b), and while this suggests the opposite direction than for the *charismatic* rating, this was not significant (*p* = .25). There are no significant correlations between the ratings on any scale and F2 range (all *p*-values >.1; see Table 1). There is a non-significant trend in that larger F2 ranges tend to be perceived as less likeable (τ=-0.4, *p*=.12).

4.2. Stimuli (*SData*)

Two significant correlations arise for the F1 ranges. The *charismatic* rating is positively correlated with the F1 range (τ = 0.72, *p* = .02); thus speakers with a larger F1 range are perceived as significantly more charismatic than speakers with a smaller F1 range (see Figure 1c). At the same time, there is a significant negative correlation between the *authentic* rating and the F1 range (τ = -0.64, *p* = .03): the larger the F1 range, the less authentic a speaker is perceived (see Figure 1d). These two correlations are opposite, but—taking into account only the visual tendency for the *authentic* ratings in the *FData*—the correlation directions between the *FData* and the *SData* are also opposite. There are no significant correlations between the ratings on any scales and the F2 ranges of the vowel spaces of the stimuli (all *p*-values >.1; see Table 1). However, the correlation tests reveal some trends, suggesting that speakers with larger F2 ranges tend to be perceived as more charismatic (τ = 0.49, *p* = .1) and persuasive (τ = 0.46, *p* = .12), but less likeable (τ = -0.44, *p* = .13).

5. DISCUSSION

This study investigated correlations between formant ranges and charisma ratings for YouTubers. RQ1 and RQ2 are only partially supported for the *SData*, not the *FData*. There are differences between the *SData* and the *FData*, so RQ3 is confirmed for the current sample.

The correlation between *charismatic* and F1 range for the *SData* in the current study was significant and positive (as in previous literature, and partially supporting H1), while the negative correlation with

Type	Attribute	F1 range (F)	F1 range (S)	F2 range (F)	F2 range (S)
charisma	*charismatic*	τ=-0.64, p =.01*	τ=0.72, p =.02*	τ=-0.05, p =.86	τ=0.49, p =.1^{+}
emotion	*enthusiastic*	τ=-0.26, p =.31	τ=0.4, p =.18	τ=-0.17 p =.52	τ=0.08, p =.79
emotion	*likeable*	τ=-0.26, p =.32	τ=-0.15, p =.62	τ=-0.4, p =.12^{+}	τ=-0.44, p =.13^{+}
cognition	*authentic*	τ=0.31, p =.25	τ=-0.64, p =.03*	τ=0.1, p =.7	τ=-0.16, p =.6
cognition	*persuasive*	τ=0.07, p =.78	τ=0.39, p =.2	τ=0.22, p =.4	τ=0.46, p =.12^{+}

Table 1: The results of the Kendall's correlation tests for each scale attribute for the mean F1 and F2 ranges (F = *FData*, S = *SData*). Significant correlations are marked with an asterisk, non-significant trends with a plus.

the *authentic* rating does not match previous findings [14]. In the *FData*, the correlations were opposite to those in the *SData*, and also opposite to those in the literature, and not supporting H1. Unlike in [14], there were no other significant correlations for F1 ranges and attributes, and no significant correlations between attributes and F2 ranges.

Neither H2 nor H3 were supported by the data. There is no evidence that larger mean F1 ranges are positively correlated with higher ratings of emotion-related attributes, as there were no significant correlations with the attributes *likeable* and *enthusiastic*. Similarly, there is also no evidence for positive correlations between F2 range and cognition-related attributes, as there were no significant differences in the data at all, though there were tendencies for the F2 ranges that seem to align with previous research. There is a tendency for speakers with larger F2 ranges to be perceived as more charismatic and persuasive, which mirrors the results in [14], and may partially align with H3. The more emotion-related rating *likeable* has a tendency to be rated lower the larger the F2 range is which does not seem to be in line with [14]. It might also be that the *authentic* rating in this current study is more emotion-related than cognition-related, though. In [14], the term *trustworthy* was explained to participants to mean "capable of living up to [...] promises" (p.270), while the *authentic* rating (taken as similar) in the current study was explained as the speaker not putting on an act. If this is the case, there was a positive correlation between F1 range and *authentic* rating in the *SData* that may partially support H2.

We further predicted (H4) that the direction of the correlations would match between the two data sets. This was not the case. Where the correlation with the *charismatic* rating was negative in the full data, it was positive in the *SData*. The (visual, not significant) correlation with the *authentic* rating in the *FData* was positive, while the (significant) correlation was negative in the *SData*. This suggests that for charisma and authenticity, formants may work well as predictors for perception of the stimuli the ratings were made on, but that there is more going on in larger data sets that cannot be extrapolated by using only the perception of a small subset. In particular, the overall shape and area of the vowel spaces differs and should be investigated in future studies.

The difference in the direction of correlation for *charismatic* and *authentic* in the *SData* might be explained by a balancing act on YouTube: Speakers with larger F1 ranges (i.e. exhibiting increased vocal effort) were perceived as more charismatic, and at the same time less authentic. YouTubers tend to try to appear approachable, friendly and genuine [21], while still having to keep an audience engaged and entertained. In the context of YouTube, speech is then likely perceived as less authentic when it is produced with more effort. This is similar to results from [22] which found that moderately reduced German speech was perceived as most sincere, compared to unreduced or reduced speech.

Charisma may be a different dimension that, on YouTube, is less tied to authenticity, as the goal is entertainment. It is also likely that YouTubers intentionally or subconsciously adjust their speech to be more colloquial, like a friendly conversation. This may coincide with what Labov calls the "principle of attention" in that speaking styles "can be ordered along a single dimension, measured by the amount of attention paid to speech. [...] Casual and intimate styles can be stationed at one end of this continuum, and frozen, ritualistic styles at the other" [23, p.112]. Vlogs may be positioned somewhere towards the casual end of the continuum, but still not at the extreme, as they are planned and edited (see also [24]).

In addition to investigating the vowel space area and shape instead of F1/F2 ranges, future studies should look at a larger sample of YouTubers. Preliminary studies have also shown that the jaw movement is connected to formants and a more consistent predictor for speaker charisma [25]. While physical measurements from YouTube videos are impossible, future studies may also include distance measurements of mouth and jaw facial landmarks calculated from videos with programs like OpenFace [26].

6. REFERENCES

[1] O. Niebuhr, A. Brem, J. Michalsky, and J. Neitsch, “What makes business speakers sound charismatic? A contrastive acoustic-melodic analysis of Steve Jobs and Mark Zuckerberg,” *Cadernos de Linguistica e Teoria da Literatura*, vol. 1, no. 1, pp. 1–40, 2020.

[2] F. Biadsy, A. Rosenberg, R. Carlson, J. Hirschberg, and E. Strangert, “A cross-cultural comparison of American, Palestinian, and Swedish perception of charismatic speech,” in *Speech Prosody*, vol. 37, 2008.

[3] A. Rosenberg and J. Hirschberg, “Charisma perception from text and speech,” *Speech Communication*, vol. 51, no. 7, pp. 640–655, 2009.

[4] R. Signorello, F. D’Errico, I. Poggi, D. Demolin, and P. Mairano, “Charisma perception in political speech: A case study,” in *International Conference on Speech and Corpora (GSCP 2012)*, 2012, pp. 343–348.

[5] F. D’Errico, R. Signorello, D. Demolin, and I. Poggi, “The perception of charisma from voice: A cross-cultural study,” in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013, pp. 552–557.

[6] S. Berger, O. Niebuhr, and B. Peters, “Winning Over an Audience–A Perception-based Analysis of Prosodic Features of Charismatic Speech,” in *Proc. 43rd Annual Conference of the German Acoustical Society, Kiel, Germany*, 2017, pp. 1454–1457.

[7] O. Jokisch, V. Iaroshenko, M. Maruschke, and H. Ding, “Influence of age, gender and sample duration on the charisma assessment of German speakers,” *Proc. 29. Konferenz Elektronische Sprachsignalverarbeitung (ESSV2018)*, 2018.

[8] S. Berger, O. Niebuhr, and A. Brem, “Of voices and votes: Phonetic charisma and the myth of Nixon’s radio victory in his first 1960 TV debate with Kennedy,” in *An den Rändern der Sprache*, ser. Kieler Forschungen zur Sprachwissenschaft, M. Elmentaler and O. Niebuhr, Eds. Peter Lang, 2020, pp. 109–145.

[9] S. Berger and M. Zellers, “Pitch accent position, peak height, and prominence level relative to accented vowel onset on YouTube,” in *1st International Conference on Tone and Intonation (TAI)*, 2021, pp. 137–141.

[10] J. Beck, “The Linguistics of ‘YouTube Voice’,” *The Atlantic, December 7, 2015*, 2015.

[11] S. Hagi, “The rise of the ‘YouTube Voice’ and why vloggers want it to stop,” *Vice.com, March 28, 2017*, 2017.

[12] O. Niebuhr and S. Gonzalez, “Do sound segments contribute to sounding charismatic? Evidence from a case study of Steve Jobs’ and Mark Zuckerberg’s vowel spaces,” *International Journal of Acoustics and Vibration*, vol. 24, no. 2, pp. 343–355, 2019.

[13] S. Scherer, L.-P. Morency, J. Gratch, and J. Pestian, “Reduced vowel space is a robust indicator of psychological distress: A cross-corpus analysis,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4789–4793.

[14] O. Niebuhr, “Space fighters on stage—How the F1 and F2 vowel-space dimensions contribute to perceived speaker charisma,” *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung*, pp. 265–277, 2020.

[15] S. Berger, O. Niebuhr, and M. Zellers, “A Preliminary Study of Charismatic Speech on YouTube: Correlating Prosodic Variation with Counts of Subscribers, Views and Likes,” in *INTERSPEECH*, 2019, pp. 1761–1765.

[16] P. Boersma and D. Weenink, “Praat: doing phonetics by computer,” 2018. [Online]. Available: http://www.praat.org/

[17] Y. Xu and H. Gao, “FormantPro as a tool for speech analysis and segmentation,” *Revista de Estudos da Linguagem*, vol. 26, no. 4, pp. 1435–1454, 2018.

[18] T. Kisler, U. Reichel, and F. Schiel, “Multilingual processing of speech via web services,” *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.

[19] D. R. McCloy, “phonR: Tools for phoneticians and phonologists,” *R package version*, vol. 1, 2016.

[20] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[21] R. Kyncl and M. Peyvan, *Streampunks: How YouTube and the new creators are transforming our lives*. Random House, 2017.

[22] O. Niebuhr, “Clear Speech-Mere Speech? How Segmental and Prosodic Speech Reduction Shape the Impression That Speakers Create on Listeners,” in *INTERSPEECH*, 2017, pp. 894–898.

[23] W. Labov, “Some principles of linguistic methodology,” *Language in Society*, vol. 1, no. 1, pp. 97–120, 1972.

[24] S. Lee, “Style-shifting in vlogging: An acoustic analysis of “YouTube voice”,” *Lifespans & Styles: Undergraduate Working Papers on Intraspeaker Variation*, vol. 3, no. 1, pp. 28–39, 2017.

[25] O. Niebuhr and A. Gutnyk, “Pronunciation engineering: Investigating the link between jaw-movement patterns and perceived speaker charisma using the MARRYS cap,” in *2021 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*. IEEE, 2021, pp. 1–6.

[26] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, “Openface 2.0: Facial behavior analysis toolkit,” in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 59–66.

[1] The nine videos used in this investigation are collected in a playlist on YouTube: https://bit.ly/3N1TnfC