TOTAL VARIATION IN VOCALS OVER TIME

Elena Georgieva, Pablo Ripollés, Brian McFee

[1]Music and Audio Research Lab, New York University

ABSTRACT

Recent advancements in audio processing provide a new opportunity to study musical trends using quantitative methods. In this work, we conduct an exploratory study of 43,153 vocal tracks of popular songs spanning nearly a century, from 1924 to 2010. We use source separation to extract the vocal stem and fundamental frequency (f_0) estimation to analyze pitch tracks. In this demonstration, we calculate the total variation (TV) of each song, report trends in the metric over time and between genres, and provide listening examples on our demo website[1].

1. INTRODUCTION

Current technologies for audio processing provide new opportunities to study musical trends using quantitative methods. While researchers have analyzed music for generations, studying the evolution of music at a large scale has only been possible recently, due to the availability of large datasets [1, 2, 3]. Additionally, recent improvements in source separation technology have allowed researchers to study individual both instruments and vocals [4, 5]. However, the vocal lines of songs have been understudied, even though they are often the most salient part of a song [6, 7], and many popular songs are built around the vocal line.

In this demonstration, we examine trends in the vocal lines of 43,153 songs over 86 years (from 1924 to 2010). We use modern source separation methods to isolate vocal lines of songs from their respective accompaniments and use 30–60 second excerpts of each song. Altogether, our dataset makes up 21 days of continuous listening.

2. DATASET

We used the union of the HSP-S and HSP-L datasets ("Hit Song Prediction- Small and Large," respectively), two recent datasets intended for song popularity prediction tasks [8]. The metadata included in the datasets is based on information from AcousticBrainz[2], the Billboard Hot 100, the MSD, and last.fm[3]. Audio files 30-60 seconds in length were taken from a private mp3 sample collection of the MSD [9]. Songs with a low presence of vocals were excluded from the dataset. We compared the RMS energy of the separated vocal stem to the RMS energy of the full audio file, and removed songs with a ratio less than 0.08.

Figure 1 shows a chronological distribution of songs in demi-decade bins (i.e., 1990-1994). We observe a strong bias towards more recent songs. For each song, the dataset contains estimated labels for the singer's gender ("female" and "male")[4], estimated using the AcousticBrainz vocal gender classifier, which reports an accuracy of 87.21%. Genre tags were taken from the Tagtraum MSD annotations [10]. In our dataset, 30,797 tacks had genre labels available. Our dataset inherits biases from the MSD: songs are generally widely listened-to, the majority come from North America or Europe, and there are many more songs in later years of the dataset.

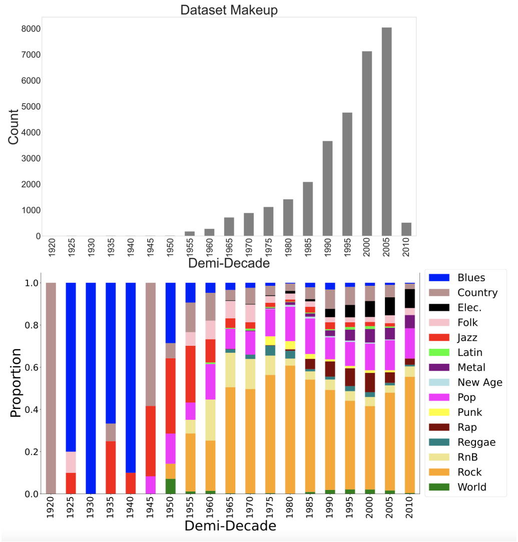

Figure 1: Top: Chronological distribution of the dataset organized in 5-year demi-decades. Bottom: Relative distribution of genres.

3. METHOD

First, we used source separation to separate the vocal line of each song from the mix. For this, we use Hybrid Transformer Demucs (HT Demucs), a hybrid temporal/spectral bi-U-Net [5].

To study pitch characteristics, we did fundamental frequency (f_0) estimation on the estimated vocal stems using PYIN [11] as implemented in Librosa v0.8.1[12]. We set the minimum frequency to 70Hz and the maximum frequency to 900 Hz, to capture likely human vocal range. We use a sampling rate of 44.1 kHz, frame length of 2048, hop length of 512; we set the number of thresholds for peak estimation to 5, and the no-trough probability to 0.99. PYIN also provides a voicing detection estimate, which we used to

[1] `https://ccrma.stanford.edu/~egeorgie/projects/totalvariation.html`

[2] https://acousticbrainz.org/

[3] https://www.last.fm/

[4] The AcousticBrainz classifier estimates gender as perceived by human listeners. We suspect that the classifier is detecting register (i.e., soprano/alto and tenor/bass) rather than gender. For clarity and consistency with prior work, we will refer to these estimates as "gender", but stress that it should be interpreted with care.

identify contiguous regions of pitched sound in the vocal stem. We converted f_0 values in hertz to cents.

Using this information we chose to focus on one measure: total variation (TV)[13] after dropping unvoiced frames. TV summarizes the rate of pitch change and is defined in Equation 1:

$$\mathrm{TV}(x) = \frac{1}{N} \sum_{i=1}^{N-1} |x_{i+1} - x_i| \tag{1}$$

for a given f_0 contour $x = (x_1, \ldots, x_N)$. TV is calculated independently for each voiced region within a song and then aggregated to a single total.

4. EXPERIMENT AND RESULTS

We used R (4.2.2) and RStudio (2022.12.0+353) to implement linear regression with the *lm* function. Post-hoc tests were implemented using the *emmeans* package with Tukey correction for multiple comparisons.

We first ran a linear regression to examine the relationship between TV and the year of track release (TV∼year). Results showed a significant positive relationship (t(43152) = 7.604, p < 0.001; see black line in Figure 2).

Next, we ran a second model that included the interaction between year and gender (TV∼year*gender). Here, we found a main effect of year (β = 0.011, t=2.142, and p< 0.05), and a main effect of gender (β = -87.53, t=-6.105 p < 0.0001). There was also a significant interaction between year and gender (β = 0.045, t=6.210 p < 0.0001). TV is increasing, but the regression slope is steeper for male singers (see blue and orange lines in Figure 2).

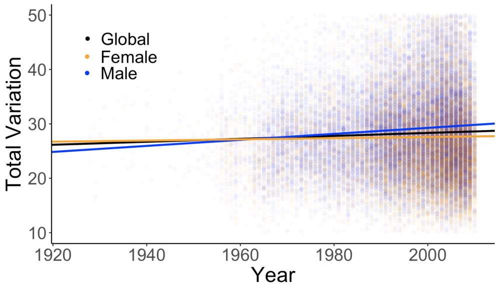

Figure 2: TV as a function of year. Each dot represents a song. The black line represents the regression slope for the model TV∼year. The orange and blue lines represent the regression slopes for the model TV∼year*gender.

We then calculated a linear regression between TV and musical genre (TV∼genre) for the eight genres having more than 950 songs represented in the dataset. We found a significant main effect of genre (F(7, 27602) = 625.81, p < 0.001). Post-hoc tests showed that all genres were significantly different than one another (all p values <0.05) except for folk and pop (t=-2.650, p=0.138) and country and rock (t=-0.245, p = 1.00; see Figure 3).

Finally, we calculated independent linear regressions between TV and year for the eight genres. For each genre, we included data starting in the year in which each of the genres became prevalent. The only significant main effect of year on TV was that rap music showed a significant negative relationship between year and TV (β=-0.169, t=-5.217, p<0.0001; see Figure 4).

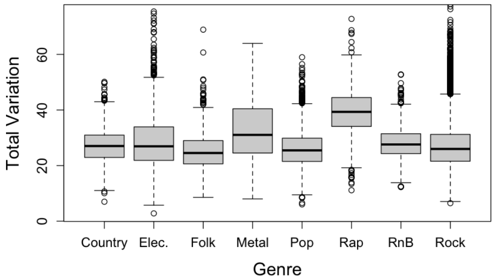

Figure 3: TV in each of the eight genres across the whole dataset. Means are shown with interquartile range, 95% confidence interval error bars, and outliers.

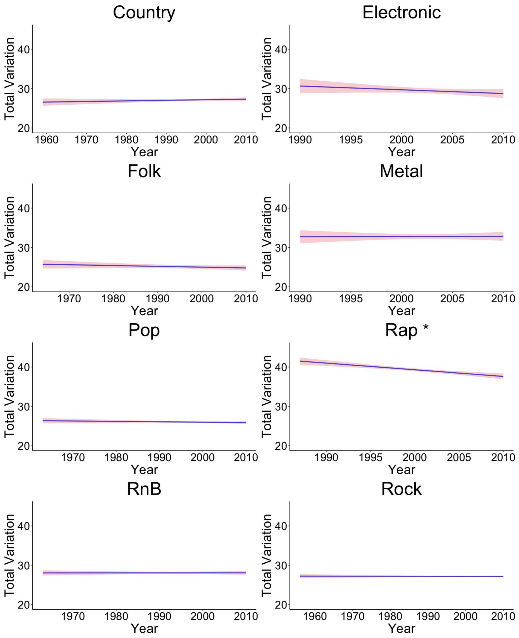

Figure 4: Relationship between TV and year for each musical genre. A significant main effect, denoted by "*", was found for rap music.

5. DISCUSSION

The results showed a significant positive relationship between TV and year. We observed, notably, that the rap genre had a higher TV than the other genres (Figure 3), showing that rap songs feature more pitch variation than other musical genres, on average. This could be because rap vocals tend to have less sustained pitches than the vocals of other genres. This can be observed in the provided audio examples with high TV.

There has been previous work on vocal pitch content in rap music, affirming that pitch variance in rap music is a complex and significant feature of the genre [14, 15]. Thus, the increasing TV over time in Figure 2 can likely be explained by the increasing prevalence of rap in the dataset over the years.

In future work, we will study mean pitch and pitch class entropy of these songs over time and across musical genres.

6. REFERENCES

[1] T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011, Miami, Florida, USA*, A. Klapuri and C. Leider, Eds., 2011, pp. 591–596.

[2] J. Serrà, A. Corral, M. Boguñá, M. Haro, and J. L. Arcos, "Measuring the evolution of contemporary western popular music," *CoRR*, vol. abs/1205.5651, 2012. [Online]. Available: http://arxiv.org/abs/1205.5651

[3] U. Shalit, D. Weinshall, and G. Chechik, "Modeling musical influence with topic models," in *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, ser. JMLR Workshop and Conference Proceedings, vol. 28. JMLR.org, 2013, pp. 244–252.

[4] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," in *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France*, E. Gómez, X. Hu, E. Humphrey, and E. Benetos, Eds., 2018, pp. 334–340.

[5] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," *CoRR*, vol. abs/2211.08553, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2211.08553

[6] A. M. Demetriou, A. Jansson, A. Kumar, and R. M. Bittner, "Vocals in music matter: the relevance of vocals in the minds of listeners," in *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018*, E. Gómez, X. Hu, E. Humphrey, and E. Benetos, Eds., 2018, pp. 514–520.

[7] M. Bürgel, L. Picinali, and K. Siedenburg, "Listening in the mix: Lead vocals robustly attract auditory attention in popular music," *Frontiers in Psychology*, vol. 12, 12 2021.

[8] M. Votter, M. Mayerl, G. Specht, and E. Zangerle, "Novel datasets for evaluatingsong popularity prediction tasks," *IEEE International Symposium on Multimedia*, p. 289, 2021.

[9] A. Schindler, R. Mayer, and A. Rauber, "Facilitating comprehensive benchmarking experiments on the million song dataset," in *Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR 2012, Mosteiro S.Bento Da Vitória, Porto, Portugal, October 8-12, 2012*, F. Gouyon, P. Herrera, L. G. Martins, and M. Müller, Eds. FEUP Edições, 2012, pp. 469–474.

[10] H. Schreiber, "Improving genre annotations for the million song dataset," in *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015, Málaga, Spain, October 26-30, 2015*, M. Müller and F. Wiering, Eds., 2015, pp. 241–247.

[11] M. Mauch and S. Dixon, "Pyin: A fundamental frequency estimator using probabilistic threshold distributions," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 659–663.

[12] B. McFee, A. Metsai, M. McVicar, S. Balke, C. Thomé, C. Raffel, F. Zalkow, A. Malek, Dana, K. Lee, O. Nieto, D. Ellis, J. Mason, E. Battenberg, S. Seyfarth, R. Yamamoto, viktorandreevichmorozov, K. Choi, J. Moore, R. Bittner, S. Hidaka, Z. Wei, nullmightybofo, D. Herénú, F.-R. Stöter, P. Friesch, A. Weiss, M. Vollrath, T. Kim, and Thassilo, "librosa/librosa: 0.8.1rc2," https://doi.org/10.5281/zenodo.4792298, May 2021.

[13] M. Panteli, R. Bittner, J. P. Bello, and S. Dixon, "Towards the characterization of singing styles in world music," *EEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[14] M. Ohriner, "Analysing the pitch content of the rapping voice," *Journal of New Music Research*, vol. 48, pp. 413 – 433, 2019.

[15] R. Komaniecki, "Vocal pitch in rap flow," *Intégral*, vol. 34, pp. 25–46, 2020.