# A COMPARATIVE ANALYSIS OF POETRY READING AUDIO: SINGING, NARRATING, OR SOMEWHERE IN BETWEEN?

*Kahyun Choi*[1] *and Minje Kim*[2‡]

[1]Indiana University, Luddy School of Informatics, Computing and Engineering, Bloomington, IN, 47408
[2]University of Illinois at Urbana-Champaign, Department of Computer Science, IL, USA 61801

## ABSTRACT

This paper provides a computational analysis of poetry reading audio signals at a large scale to unveil the musicality within professionally-read poems. Although the acoustic characteristics of other types of spoken language have been extensively studied, most of the literature is limited to narrative speech or singing voice, discussing how different they are from each other. In this work, we develop signal processing methods, which are tailored to capture the unique acoustic characteristics of poetry reading based on their silence patterns, temporal variations of local pitch, and beat stability. Our large-scale statistical analyses on three big corpora, each of which consists of narration (LibriSpeech), singing voice (Intonation), and poetry reading (from The Poetry Foundation), discover that poetry reading does share some musical characteristics with singing voice, although it may also resemble narrative speech.

***Index Terms*—** Poetry reading, singing voice, pitch tracking, beat tracking, musicality

## 1. INTRODUCTION

Recently, poetry reading has gained increased popularity. According to the 2017 survey from the National Endowment for the Arts, the poetry readership in the United States increased by 76% between 2012 and 2017 [1]. A follow-up survey in 2022 also reported high poetry readership: 11.5% of adults in the US engaged in poetry either by reading or listening, while 4.8% of US adults engaged in poetry *listening* [2], highlighting a large audience for poetry audio. Recently, the Recording Academy introduced a new, dedicated category for "Best Spoken Word Poetry Album" to the Grammy Awards [3], as a response to calls from the broader spoken word community for more equitable representation of poetry reading. However, compared to the other oral performance types, e.g., narrations and singing voice, computational analysis of poetry reading's acoustical characteristics has rarely been done in the literature.

Similarly to any other speech, poetry reading must also contain a certain level of musicality coming from the pitched voice and the pace of reading [4]. However, it also has a deep-rooted connection with vocal music or singing voice [5]. Historically, poetry has been conveyed through oral performances, or recitations, across various cultures. For instance, ancient Greek poets often performed lyric poems accompanied by a lyre; traditional Chinese poems were often sung to accompany a musical instrument, such as the pipa or the guqin [6, 7]. Also, poetry often employs musical devices to enhance its rhythm. For example, rhyme creates correspondence in sounds between words, while meter dictates the rhythmic structure through patterns of stressed and unstressed syllables. Additionally, poetry reading exhibits pitch variations as a form of spoken language. However, it is not widely agreed upon how to interpret the pitched signal as a form of "melody" or how its characteristics differ from those in singing voice or narrations [8].

Extensive research has been conducted on the acoustic features of narrated speech (e.g., spoken language that reads non-poetic text or conversational speech) and singing voice. For example, compared to narrating, singing tends to use higher pitch, slower temporal rate, and more stable pitches and rhythm [9, 10, 11, 12]. However, few studies have attempted to quantify where poetry reading falls within this spectrum. Hence, there is a significant gap in our understanding while poetry reading appears to exhibit a unique blend of characteristics typical of both narrating and singing. Meanwhile, computational analyses of poetry in the literature have primarily been done from a natural language understanding perspective, e.g., understanding the semantics, themes, and meanings of the text [13, 14, 15, 16, 17, 18], missing the acoustic aspect of poetry reading.

In this paper, we conduct a large-scale quantitative analysis of three different types of oral performance: singing voice, narration, and poetry reading. To this end, we propose a few signal processing algorithms that allow us to scale up the experiments to the degree that was previously infeasible. Specifically, we begin with an analysis of silence patterns of the three categories, and then provide insights coming from their pitch variation patterns, e.g., whether the poetry reading has steady-pitched areas as in singing voice. Finally, we also provide an analysis of the beat stability of the three types and present our findings that the poetry reading has a certain level of established beat patterns. We perform the experiments on three large-scale, publicly-available datasets: poetry readings collected from The Poetry Foundation[1], the LibriSpeech audiobook dataset [19], and the Intonation set with isolated singing voice recordings [20]. We ensured that other non-vocal sound sources, e.g., musical instruments, sound effects, etc., are not included in our datasets, while some recordings could contain moderate environmental noise. To our best knowledge, this paper reports the first large-scale quantitative analysis of poetry reading in comparison to narrations and singing voices. The proposed algorithms are developed to provide the best-effort analysis of the poetry reading, while we hope that deeper and broader quantitative studies follow our work for a better understanding of poetry reading. Whilst the study is at an unprecedented scale, it is still limited to Western pop music. Moreover, 93% of the poems were written after 1980, confirming a strong contemporary focus. Since poems in this era use more flexible and experimen-

*This work was supported by RE-252382-OLS-22 from the Institute of Museum and Library Services.

†The authors appreciate Smule, Inc. for providing the Intonation dataset and Sumitha Vellinalur Thattai's help on the initial data collection effort.

‡Work done at Indiana University.

[1]https://www.poetryfoundation.org

tal forms, e.g., free verse, [21], the signal processing algorithms are more challenged to find a distinguishable feature. We open-source our project and provide necessary metadata needed to reproduce the results: `https://github.com/kc82/poetry-reading`.

## 2. METHODOLOGY

In the literature, both narration and singing voice have been acoustically examined through a variety of features, including pitch range, pitch stability, rhythm, and tempo [9, 10, 11, 12]. However, these studies were on small datasets and subjective measures such as user responses or author annotations for assessment. To scale up, we devise quantifiable metrics and apply them to more than 1,000 audio files for each category. We aim to position poetry reading within the spectrum that spans narration and vocal music. Although we analyze audio from a local timeframe to a global histogram perspective, we leave the consideration of long-term patterns to future work.

### 2.1. Preprocessing via Transcription

We preprocess the audio signals using WhisperX to distinguish between voiced and silent segments, detect language, and select clean audio files [22]. Whisper [23] is a state-of-the-art automatic speech recognition (ASR) package renowned for its low word error rate (WER), whose successor WhisperX provides improved performance. We opt for WhisperX because it offers more accurate word boundary detection through its phoneme-based ASR, which provides the beginning and ending timestamps of each word. Additionally, it is equipped with voice activity detection (VAD), making it useful for identifying silent intervals within audio files.

**Silence Detection Based on Word Boundaries**: Audio files often contain silence either at the beginning, end, or interspersed between voiced segments. These patterns of silence can vary across different audio genres. To consider or mitigate the impact of silence on some of our analysis algorithms, we conduct studies on both silence-removed and contained versions of the audio. A simple-minded silence detection approach, e.g., a decision based on a threshold amplitude, is not a robust approach as some of the recordings contain different levels of noise. More appropriate methods are based on the VAD model trained from noisy speech signals, such as the one WhisperX provides. Instead of using the VAD output directly, we further process the signal to compute the word boundaries using WhisperX's ASR module. Hence, we define silence by the areas outside of the word boundaries. This is at the cost of excluding some spoken words that ASR fails to detect or filler sounds with no meanings, making the result after silence removal more conservative.

**Language Detection**: Each language has its distinct acoustic characteristics. To control for this variable and align with the English LibriSpeech dataset, our analysis is limited to English content. This is done by WhisperX's language detection score, i.e., if it is greater than 0.93 for English. However, we plan to include multiple languages in future studies.

### 2.2. Local Pitch Variability

A well-known qualitative feature of the singing voice is its steady pitch that remains within a short period of time, i.e., the concept of musical notes. On the contrary, in narration, the vowel sound often varies its pitch over time, creating an unstable pitch contour [9, 10]. In this paper, we empirically capture this local pitch variation.

First, we estimate the pitch value at every 64 ms-long audio frame by using the probabilistic YIN (pYIN) algorithm [24]. First of all, we ignore the silent areas defined by the ASR model as described in Sec. 2.1. Yet, for some unvoiced frames, pYIN can fail and result in "not a number" (NaN). Hence, when we compute the local pitch statistics of 12 consecutive frames (96 ms), we make sure that (a) there are less than seven NaN values (b) at least five consecutive pitch values are found within the 12-frames window. Otherwise, we disregard that 12-frame chunk. These local statistics (i.e., standard deviation of those pitch values) will tell us whether pitch varies too much within a short period of time or sustains. We will perform statistical tests to verify that the poetry reading is more likely to contain sustained pitches than narration.

### 2.3. Beat Stability

The rationale behind the beat stability feature is that the beat tracking effort (either made by human listeners or machines) should be mitigated if the audio signal contains a regular beat pattern. Otherwise, such as in narration, it takes more exceptions to identify candidate beat locations that are off from the perceived regularity.

To this end, we employ a well-known dynamic programming (DP) algorithm to track down the beats faithfully [25], which is implemented as the `librosa.beat.beat_track` function in `librosa` [26]. Specifically, its objective function provides a systematic way to quantify the rhythmic irregularity, which we use to compare the three categories. The objective is defined over the onset signal $O(\cdot)$ as input, which records the abrupt changes of audio, and then finds the beat sequence $\{t_i\}$ that maximizes the score function:

$$C(\{t_i\}) = \sum_{i=1}^{N} O(t_i) + \alpha \sum_{i=2}^{N} F(t_i - t_{i-1}, \tau_p). \quad (1)$$

The objective function can be understood from two perspectives. Since perceived beats are based on sound events, if the onset function's values at the beat positions are large, the beats match the salient sound events' locations. The first summation in the objective function quantifies this: the larger the sum is, the better the match is.

Meanwhile, the objective function also allows some beat locations that are slightly off from the regularity. The beat tracking algorithm first finds the global tempo from the onset signal, and then tries to keep the number of beats per minute (BPM) the same as the global tempo. Here, the second term regularizes the optimization by enforcing a particular beat interval, $\tau_p$, to all inter-beat time differences $t_i - t_{i-1}$ using a pre-defined penalty function: $F(\Delta t, \tau) = -\left(\log \frac{\Delta t}{\tau}\right)^2$. Hence, if it were not for the second term, the first term only quantifies the sum of the onset function values and can produce a trivial solution, e.g., beat locations match the loudest sound events regardless of the periodicity.

$\alpha$ is an important hyperparameter, which controls the "tightness" of the beat estimates compared to the target tempo. A small $\alpha$ will result in less regular beat patterns, as the DP optimizer relies only on the large onset values (the first term of eq. (1)). On the other hand, a too-large $\alpha$ will estimate a periodic sequence of beat locations that are irrelevant to the perceived beats defined by the sound events.

In this work, we quantify the beat stability by using the DP algorithm's performance depending on the different choices of $\alpha$: if a steady beat pattern is observed, the objective function $C(\{t_i\})$ will be affected less by a smaller choice of $\alpha$, and vice versa. Therefore, for a given recording, we compare the total beat tracking scores by changing $\alpha$ from 1,000 to 1 (the default value is 100). As for the onset detection, we use the spectral fluctuation feature [27]. We denote the optimal objective function value of the audio signal by

$$C^* = \arg\max_{\{t_i\}} C(\{t_i\}). \quad (2)$$

## 3. DATASETS

Table 1 summarizes the basic information of the datasets we use in this paper. The datasets are closely matched in terms of both the number of files and word counts.

**Poetry Reading Dataset**: The poetry reading dataset is sourced from `www.poetryfoundation.org`. The Poetry Foundation stands as a distinguished platform boasting thousands of audio recordings, offering a broad spectrum of data as well as the reliability of the dataset when analyzing the vocal nuances of poetry. We only consider the poems that are accompanied by audio, i.e., a recorded poetry recitation. Starting from the initial 1,300 crawled audio files, we refine the dataset by choosing the ones that WhisperX detects their language as English with 93% certainty. Although it does not necessarily mean that those excluded samples were written in other languages, in this way, the selected examples are favored by the ASR model, improving the word boundary detection performance. This selection step reduces the final dataset down to 1,058 audio files. We provide the URLs of the poems we used in this study. We reduce their sample rate to 16kHz.

**Narration Dataset**: LibriSpeech [19] provides a collection of 1,101 unique books translated into audio files with 16 kHz sample rate, serving as a representation of narrative speech. This is primarily because the content is drawn from full-length books, which inherently encompass continuous and extended narratives, both in fiction and non-fiction genres. The natural flow of language, varying tones, cadences, and expressive modulations intrinsic to book readings capture the essence of narrative speech. Furthermore, as these recordings are derived from diverse authors, styles, and periods, they collectively offer a comprehensive portrayal of storytelling and narrative techniques across a broad spectrum. We carefully concatenated the segmented audio signals to be slightly over 90 seconds to increase the length of the signals in our experiments, while preserving their sequence and continuity of the narrative. We brought data only from the clean data fold, leaving out non-clean data.

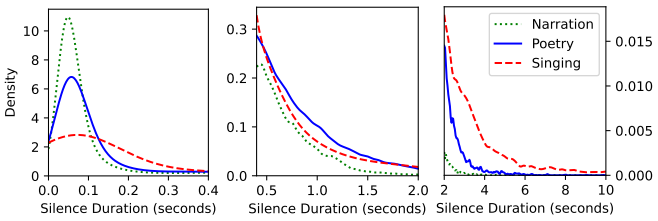
**Singing Voice Dataset**: We use the Intonation dataset [20] as a representative of singing voice. Intonation contains 4,702 audio files sourced globally using a karaoke app serviced by Smule, Inc. This ensures a broad spectrum of singing styles, techniques, and nuances inherent to diverse cultures and traditions. The user-generated nature of Intonation brings a raw authenticity to the collection, encompassing both trained voices and natural, untrained vocal expressions. In addition, Intonation is suitable for our purposes because the recordings do not contain the other accompanying musical instruments. Once again, we reduce it down to 1,050 English-language songs using WhisperX's language detection with the same criteria applied to select poetry audio. We resample the signals at a 16kHz rate.

## 4. EXPERIMENTS

### 4.1. Silence Patterns

In all three categories, it is common to observe silent regions in between words. However, their lengths and functions in the oral performance have different meanings. For example, in narration, there tends to be a pause between sentences. On the other hand, in singing voice, long pauses are also common in Western pop music, e.g., during the interlude. We are interested in the intentional pauses in poetry reading, which are often expected after a line or stanza [28] to maintain the rhythms or make poems more song-like [29].

To capture this, we draw histograms of the lengths of the silent periods, which are acquired from the silence removal process in Sec. 2.1 as a byproduct. Fig. 1a, shows three histograms of relatively short silent chunks (from 0 to 0.4 second). First, very short pauses (below 0.1 second) are the most common type across the three categories; however, their prevalence varies, being most frequent in narration, moderately common in poetry reading, and comparatively less frequent in singing voice. In Fig. 1b, we see more pauses in poetry reading, indicating that poetry reading contains those medium-length (around 1 second) pauses, more than other categories. Finally, In Fig. 1c, the density of singing voice surges, showcasing its frequent long pauses. Likewise, we can see that moderate-length pauses are relatively common in poetry reading, which is a unique feature that differentiates it from other categories.




(a) (b) (c)

**Fig. 1**: Histograms of the (a) short (b) medium (c) long silent segments.

**Table 1**: Summary of the Datasets

| | LibriSpeech | Poetry Reading | Intonation |
|---|---|---|---|
| Number of Files | 1,101 | 1,058 | 1,050 |
| Spoken Duration (min) | 1,079 | 1,318 | 1,889 |
| Word Count | 285,233 | 307,039 | 269,567 |
| Total Silence (min) | 653 | 1,142 | 1,450 |
| Words Per Min. | 164.73 | 124.88 | 80.76 |
| Std. Silence (sec) | 0.25 | 0.41 | 1.25 |

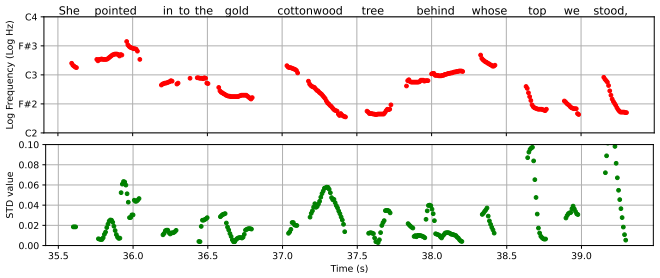
### 4.2. Local Pitch Variability

The top figures in Fig. 2 illustrate five second-long representative pitch contours (in red) of the three categories. Once again, in general, poetry reading and narration share similar patterns, i.e., less steady pitch contours compared to singing voices'. However, poetry reading sometimes contains more elongated vowel sounds and more steady-pitch areas than narration. The bottom graphs (in green) show the corresponding local pitch variation values, i.e., std of the 12 consecutive pitch values, where the high std values are correlated with the steep changes of the pitch contours (in narration), while the low-std regions are associated with the sustained musical notes (in singing voice). Poetry reading shows less drastic changes in its contour, leading to lower std than narration.

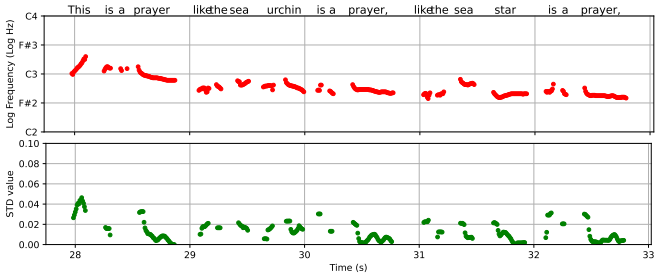
The histograms of the three std categories are shown in Fig. 3. Once again, poetry reading shows a similar trend to narration, while it has more density towards zero. We conduct a two-sample Kolmogorov-Smirnov (KS) test [30] to verify the moderate but statistically significant difference between poetry reading and narration. Their KS statistic of 0.089 is relatively small, while the p-value is 0.0 (below the machine precision): the chance of the two std sets being from the same distribution is improbable.

### 4.3. Beat Stability

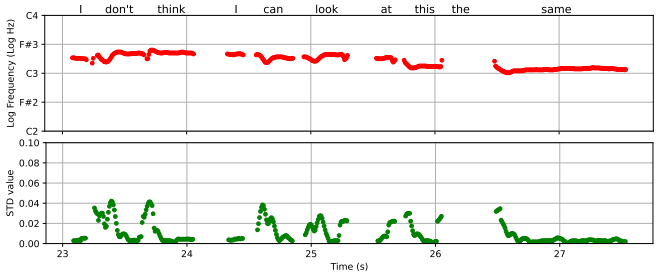
We extract the final score $C^*$ as defined in eq. (2), which indicates the beat-tracking algorithm's fitness for the signal. We consider two potential issues to interpret the scores. First, $C^*$ keeps increasing as the algorithm finds more beats from the signal, so we divide the final score by the number of found beats to normalize it.




(a) Narration




(b) Poetry reading




(c) Singing voice

**Fig. 2**: Pitch contours (top) and local standard deviation (bottom).

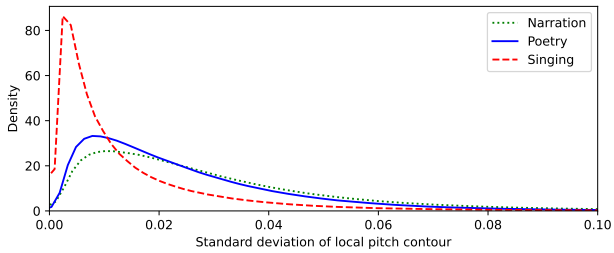
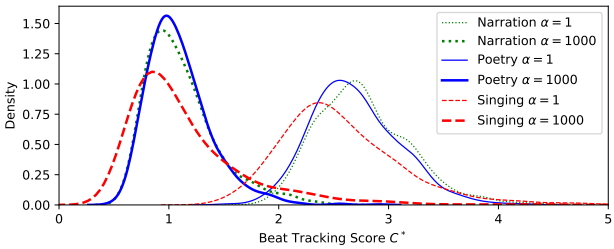
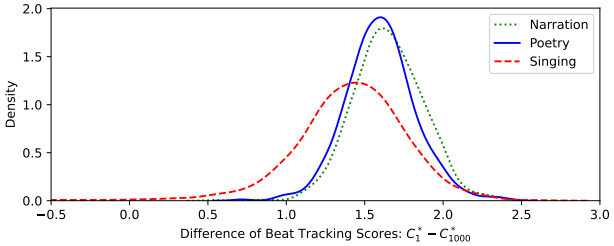



**Fig. 3**: Histograms of the std values of the local pitch contours.




(a) Change of the beat tracking score $C^*$ by varying the tightness value $\alpha$




(b) Histogram of the difference of the beat tracking scores $C_1^* - C_{1000}^*$

**Fig. 4**: Histograms of the beat tracking scores.

Second, due to the variations among different recordings, a direct comparison of the scores computed from two different examples may not be robust. For example, a soft song with a regular beat pattern could have a lower score than a series of drum attacks with no periodicity due to their differences in the onset peaks. To this end, we propose to compute two beat-tracking scores per recording by using two different tightness setups, $\alpha = 1$ and $1,000$, i.e., $C_1^*$ and $C_{1000}^*$. Our assumption is that, even for the same audio, $C_1^*$ will go up because beat tracking is free to find the largest onset peaks, ignoring the inter-beat temporal regularity. However, the same audio's $C_{1000}^*$ could be poor as there is less chance for the beat tracking algorithm to find the onset peaks at the rigidly regular beat intervals, unless the audio signal already contains a regular beat pattern.

Fig. 4a presents two histograms from $C_1^*$ (thin line) and $C_{1000}^*$ (thick line) per category. In all three categories, we see that $C_1^*$'s distribution is overall with higher scores than $C_{1000}^*$ as expected. However, their differences vary depending on the category. For example, the difference between $C_{1000}^*$ and $C_1^*$'s empirical distributions for the singing voice category is the smallest, with a Wasserstein distance [31] of 1.41. In other words, the singing voice tends to contain more regular beat patterns, which can be found more easily by the beat tracking algorithm, even with the very tight regularization. On the other hand, narration's distributions change more drastically with a higher Wasserstein distance of 1.65: the algorithm exploits less regularization to find higher onset peaks. What is interesting is the poetry reading distributions: their beat tracking results differ more widely than singing voice's by varying $\alpha$, but their difference is less than that of narration (1.60). It means the very rigid enforcement of a steady tempo still finds more meaningful beat locations from poetry reading than narration.

In Fig. 4b, we compute the difference between $C_1^*$ and $C_{1000}^*$ of each recording, and draw their histograms. It re-emphasizes our findings (a) singing voice is less affected by the beat tracking algorithm's rigidness than narration (b) poetry reading contains slightly, but significantly more beat patterns. The KS test between the narration and poetry distributions from this graph reports a relatively low score of 0.1059, reflecting their resemblance. However, the p-value is significantly low ($9.896 \times 10^{-6}$), thus clearly rejecting the hypothesis that they originate from the same distribution.

## 5. CONCLUSION

In this research, we applied computational methodologies to analyze the acoustic parameters of professionally-read poetry, a domain previously less scrutinized than narrative speech and singing voice. Utilizing advanced signal processing techniques, we specifically addressed the characteristics of poetry reading based on silence patterns, pitch temporal variations, and beat stability. The analysis results from three comprehensive corpora indicated that poetry reading exhibits intermediate characteristics between narrative and singing. To our best knowledge, this is the first large-scale poetry reading analysis. We open-sourced our project for successive research.

## 6. REFERENCES


[1] S. Iyengar, B. Nichols, P Moore Shaffer, M. Menzer, E. Grantham, H. Santoro, A. Moyseowicz, and E. Hall, "US trends in arts attendance and literary reading: 2002–2017," 2018.

[2] S. Iyengar, "New Survey Reports Size of Poetry's Audience, Streaming Included," 2023, Accessed: September 5, 2023.

[3] Author's Name, "Behind The Grammy: Best Spoken Word Poetry Album Roundtable, New Category," 2023.

[4] S. Nooteboom et al., "The prosody of speech: melody and rhythm," *The handbook of phonetic sciences*, vol. 5, pp. 640–673, 1997.

[5] N. Francis, "Verbal art across language and culture: poetry as music," *Neohelicon*, vol. 48, no. 2, pp. 539–552, 2021.

[6] A. E. Harvey, "The classification of Greek lyric poetry," *The Classical Quarterly*, vol. 5, no. 3-4, pp. 157–175, 1955.

[7] Y. Hou and A. Frank, "Analyzing sentiment in classical chinese poetry," in *Proceedings of the 9th SIGHUM workshop on language Technology for Cultural Heritage, social sciences, and humanities (LaTeCH)*, 2015, pp. 15–24.

[8] M. Scharinger, V. Wagner, C. A. Knoop, and W. Menninghaus, "Melody in poems and songs: fundamental statistical properties predict aesthetic evaluation," *Psychology of Aesthetics, Creativity, and the Arts*, vol. 17, no. 2, pp. 163, 2023.

[9] C. M. Vanden Bosch der Nederlanden, X. Qi, S. Sequeira, P. Seth, J. A. Grahn, M. F. Joanisse, and E. E. Hannon, "Developmental changes in the categorization of speech and song," *Developmental Science*, vol. 26, no. 5, pp. e13346, 2023.

[10] Y. Ozaki, A. Tierney, P. Pfordresher, J. Mcbride, E. Benetos, P. Proutskova, G. Chiba, F. Liu, N. Jacoby, S. Purdy, et al., "Similarities and differences in a global sample of song and speech recordings [Stage 1 Registered Report]," 2022.

[11] G. List, "The boundaries of speech and song," *Ethnomusicology*, vol. 7, no. 1, pp. 1–16, 1963.

[12] A. D. Patel, J. R. Iversen, and J. C. Rosenberg, "Comparing the rhythm and melody of speech and music: The case of British English and French," *The Journal of the Acoustical Society of America*, vol. 119, no. 5, pp. 3034–3047, 2006.

[13] D. M. Kaplan and D. M. Blei, "A computational approach to style in american poetry," in *Seventh IEEE International Conference on Data Mining (ICDM 2007)*. IEEE, 2007, pp. 553–558.

[14] J. Kao and D. Jurafsky, "A computational analysis of style, affect, and imagery in contemporary poetry," in *Proceedings of the NAACL-HLT 2012 workshop on computational linguistics for literature*, 2012, pp. 8–17.

[15] G. Rakshit, A. Ghosh, P. Bhattacharyya, and G. Haffari, "Automated analysis of Bangla poetry for classification and poet identification," in *Proceedings of the 12th International Conference on Natural Language Processing*, 2015, pp. 247–253.

[16] J. Kaur and J. R. Saini, "Designing Punjabi poetry classifiers using machine learning and different textual features," *Int. Arab J. Inf. Technol.*, vol. 17, no. 1, pp. 38–44, 2020.

[17] K. Choi, "Computational thematic analysis of poetry via bimodal large language models," in *the Association for Information Science and Technology (ASIS&T)*, 2023.

[18] A. Singhi and D. G. Brown, "Are poetry and lyrics all that different?," in *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, 2014, pp. 471–476.

[19] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[20] S. Wager, G. Tzanetakis, S. Sullivan, C.-I. Wang, J. Shimmin, M. Kim, and P. Cook, "Intonation: A dataset of quality vocal performances refined by spectral clustering on pitch congruence," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019.

[21] T. Baumann, H. Hussein, and B. Meyer-Sickendiek, "Style detection for free verse poetry from text and speech," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 1929–1940.

[22] M. Bain, J. Huh, T. Han, and A. Zisserman, "WhisperX: Time-Accurate Speech Transcription of Long-Form Audio," *in Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2023.

[23] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28492–28518.

[24] M. Mauch and S. Dixon, "PYIN: A fundamental frequency estimator using probabilistic threshold distributions," 2014, pp. 659–663.

[25] D. P. W. Ellis, "Beat tracking by dynamic programming," *Journal of New Music Research*, vol. 36, no. 1, pp. 51–60, 2007.

[26] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "Librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015, vol. 8, pp. 18–25.

[27] P. Masri, *Computer Modeling of Sound for Transformation and Synthesis of Musical Signals*, Ph.D. thesis, University of Bristol, UK, 1996.

[28] G. L. Dillon, "Clause, pause, and punctuation in poetry," 1976.

[29] Z. Lissa, "Aesthetic functions of silence and rests in music," *The Journal of Aesthetics and Art Criticism*, vol. 22, no. 4, pp. 443–454, 1964.

[30] F. J. Massey, "The Kolmogorov-Smirnov Test for Goodness of Fit," *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[31] L. V. Kantorovich, "Mathematical methods of organizing and planning production," *Management Science*, vol. 6, no. 4, pp. 366–422, 1960.