Evaluating the Flow of Robotic Rappers Against their Human Counterparts

MUSI 4677/6001 Final Experimental Research Report

Alec Burge, Robert Kimelman, Will Martin, Thomas Ottolin, Joshua Smith, Jace Walden

Table of Contents

Abstract

This experiment was created by a group of students taking the Music Perception and Cognition course at the Georgia Institute of Technology. Many of the students involved in the group have musical experience and have made music in some aspect whether with physical instruments or digital composition. The researchers wanted to understand how individuals would perceive different rap flows with two distinct characteristics, a digital (robotic) flow and a real (acoustic) flow. The experiment was conducted by using two songs of similar styles, A Milli by Lil Wayne and Welcome to the Party by Pop Smoke, with 10 samples each. The samples were differentiated by set millisecond increments, and whether it was the original version or a

robotically-dubbed version of the song. The participants were asked which flow was the “best”, or which flows were the most comfortable to their ear through two Qualtrics surveys. From the experiment it was found that most participants seemed to agree that the digital samples suffered worse flows than acoustic samples. This was done through multiple two-tailed t-tests between the flow ratings for the robotic version of A Milli and the flow ratings for the acoustic version of Welcome to the Party resulting in the conclusion that it is not possible for a “robotic rapper” to rise to fame as technology stands today.

Introduction

Motivation

Our group was initially interested in exploring how musicality and expression could be perceived differently across acoustic and digital mediums of creating music. As the music industry digitizes and popular music becomes increasingly more reliant on technology and futuristic sounds, the group wanted to know if people perceived them to be any more musically expressive than music that contained clear human elements. As the group discussed this research interest more, it became difficult to comprehend how they would successfully conduct an experiment that tested this accurately, given the time and resource constraints of the project. Once the group started considering what tools they had at their disposal, one member shared he had a Python-driven software that produced a robotic vocal output of lyrics inputted into a database. He wanted to know how listeners would judge the similarity between these robotic outputs and their original versions. After listening through a few previous efforts he had done to emulate a similar verse, the group decided to test the theory that with enough effort, listeners would not be able to tell much of a difference in which version, digital or acoustic, had a better flow. After discussing this question with professor Nat Condit-Schultz, he recommended that we include another factor that could impact a perception of a rapper’s flow, which is how they thought to bring in timing relative to the downbeat of the song. The researchers defined flow as the rhythm and rhymes of a song's lyrics and how they interact with the beat. Good flow can be characterized by a smooth, steady stream of lines that fit into the groove in the song.

Literature

The group was able to find a variety of studies that investigated microtiming, rhythm, and flow across different genres of music. Estefanía Cano and Scott Beveridge published *Micotiming Analysis in Traditional Shetland Fiddle Music* highlighting that they found, “these microtiming variations (in various samples of Shetland Fiddle music) dictate the rhythmic flow of a performed melody, and contribute, among other things, to the suitability of this music as an accompaniment to dancing” (Cano and Beveridge 1). In *Rhythm and Flow in Hip-Hop Music: A Corpus Study,* Adam Waller stated, “I would argue, however, that, the number of specific devices that can be described in terms of the metrical treatment of accent is actually quite rich, and furthermore that no other device (with the possible exception of rhyme scheme) is as significant to the way flow

is perceived. If our concern is with the perception of rhythmic complexity, it would be hard to argue that any factor is more important to this than the placement of syllables, which describes, ultimately, the rhythm itself!" (Waller 29). Both of these studies seemed to support that if this research team could properly establish syllabic timing in the robotic samples, the microtiming shifts across robotic and acoustic samples may have the stronger impact on flow. Olivier Senn, Lorenz Kilchenmann, Richard von Georgi, and Claudia Bullerjahn stated in their work *The Effect of Expert Performance Microtiming on Listeners' Experience of Groove in Swing or Funk Music* that, "exaggerated microtiming deviations diminish groove. But whether listeners consider microtiming magnitudes to be adequate or exaggerated seems to depend on musical genre and on the musical expertise of the listener. We suspect that the patterning of microtiming deviations is relevant, and we propose to study this aspect further in the future" (Senn et al. 9). They followed a very similar experimental design as this team envisioned by slightly manipulating the same verse to achieve different microtiming variations. Since it seemed that participant's perception would be a large influencer on how they perceive flow, the experiments must include an opportunity for participants to explain or justify their perceptual rationale

Kyle Adams examined the "most significant metrical techniques that constitute a rapper's flow." (Adams 1) in *On the Metrical Techniques of Flow in Rap Music.* He stated, "In rapping…that musical support is found primarily in the rhythmic, metrical, and syntactic arrangement of syllables—that is, in the components of what rap musicians call flow" (Adams 3). Oliver Kaunty seemingly would agree with this finding, as he investigated a similar topic in *Lyrics and flow in rap music.* He shared, "The pitched and stretched accents play around the beat and interrupt the fast flow of syllables, which results in a permanent and quick alternation between accelerating and slowing down, like a rider modulating the tempo of his horse, speeding over a course, slowing down before jumping over a fence and speeding up afterwards again, etc. The temporal shifts correspond to the effect of advanced and deleted syncopations. It is important to note that the rhymes are not only a means of high-lighting accents. They also provide the vocal layer with additional coherence, following their own phonetic and temporal logic" (Kaunty 112). He also provided "Some scholars believe that microtiming and metrical divergences are crucial for an effective (and pleasurable) perception of rhythm. The interplay of deviation and norm seems to correspond very often with the dialectic perception of tension and resolution, variety and stasis, excitement, and predictability" (Kaunty 107). The focus on syllabic timing and control over nuanced deviations seemed to be large determinants of flow. These would be important to emulate in the robotic samples in order to achieve a perception of flow from the research participants.

Hypothesis

The group's null hypothesis is there is no different baseline rating flow between acoustic and digital versions of the same song. This would be explicitly tested in the experiment methodology explained below. This means the research team would be testing to prove an insignificant difference between the acoustic and digital ratings of each samples. The group would also receive data on how people perceive flow, as well as how microtiming may augment the perception of flow in one song. However, the studies would need to be augmented to

capture more data about these two latter subject matters in order to create an additional hypothesis to prove or disprove.

Methodology

Participants in the study mainly consisted of students of the Music Perception and Cognition class. Many of them have musical experience and have played instruments or made music. Due to the limited number of participants, permission was granted to share the study to members outside of class. The amount of musical training is unknown in this group.

Participants were asked to listen to various samples of music. The songs “A Milli” by Lil Wayne and “Welcome to the Party” by Pop Smoke were chosen. They were chosen because they were representative of the kind of flow being studied, and their vocal and instrumental stems were easy to get a hold of without the use of source separation, which may have led to noise when editing the tracks. Each of the tracks were then modified. Five samples were created with differing timings between the vocal and instrumental tracks. The vocals were offset from the original tracks by -4, -2, +2, and +4 millisecond increments in a DAW. This is one of our independent variables. This led to slight, but noticeable micro-timings in the tracks due to the prevalent drum patterns in the instrumental tracks.

Another variable we modified was that an additional track was created of each of the songs in which the vocals were swapped with a robotic version of the lyrics. The program was written by one of the researchers to be able to recreate rap verses with rhythmic precision, so the timings of the lyrics would match with the original. After creating the robotic mix with the correct lyrical timings, the same millisecond offsets as the first set was applied to the lyrics and the instrumental tracks.

The study was conducted in the form of a Qualtrics survey. Two surveys were available to be chosen. The between-subjects approach was chosen to compare the effect of the human versus robot voices on each of the tracks. Each contained a human and a robotic version of one of the tracks. The beginning of the surveys included our definition of flow, which was how rhythms and rhymes interact with a beat, a smooth flow of lines being a good flow. The participants were asked to listen to five, 51-second long clips of one of the songs, each clip having a different microtiming. Participants were then asked to rate the clips in order of best to worst flow. Afterwards, they were asked to rate each clip individually as to how good they felt the flow was for each one separately on a scale from 1-5. Next, they were asked to perform the same ratings to the other song, but with a robotic voice (one survey would have the original “A Milli” and a robotic “Welcome to the Party”, and the other vice versa). Finally, participants of each survey would be asked after rating each of the clips why they had come to their ratings. They were asked how they were able to determine which samples had good flow and how some samples had better flow than others. This was done because although we had given a definition of flow, the matter is still one that is subjective, and information about what our participants think of what constitutes good flow could help us with correcting our models. An error was made in

one of the surveys in which participants were not able to select individual flow ratings for a robot version of “Welcome to the Party”. Participants finished the survey regardless. Additionally, the samples included were limited due to time constraints and limited crossover of our robotic voice generation software and separated stems of the same songs.

Both surveys can be accessed here: Group 1 and Group 2. The following bulleted list shows the microtiming and song sampled in each sample:

Group 1: ‘A Milli’ Acoustic, ‘Welcome to the Party’ Digital

- Set A: ‘A Milli’ Acoustic
    - Sample 1: Early 1 beat
    - Sample 2: Late 2 beats
    - Sample 3: Early 2 beats
    - Sample 4: Normal
    - Sample 5: Late 1 beat
- Set B: ‘Welcome to the Party’ Digital
    - Sample 6: 4ms back
    - Sample 7: Actual
    - Sample 8: 2ms forward
    - Sample 9: 2ms back
    - Sample 10: 4ms forward

Group 2: ‘Welcome to the Party’ Acoustic, ‘A Milli’ Digital

- Set C: ‘Welcome to the Party’ Acoustic
    - Sample 11: Early 1
    - Sample 12: Late 2
    - Sample 13: Late 1
    - Sample 14: Normal
    - Sample 15: Early 2
- Set D: ‘A Milli’ Digital
    - Sample 16: 2ms forward
    - Sample 17: 4ms forward
    - Sample 18: actual
    - Sample 19: 2ms back
    - Sample 20: 4ms back

Analysis

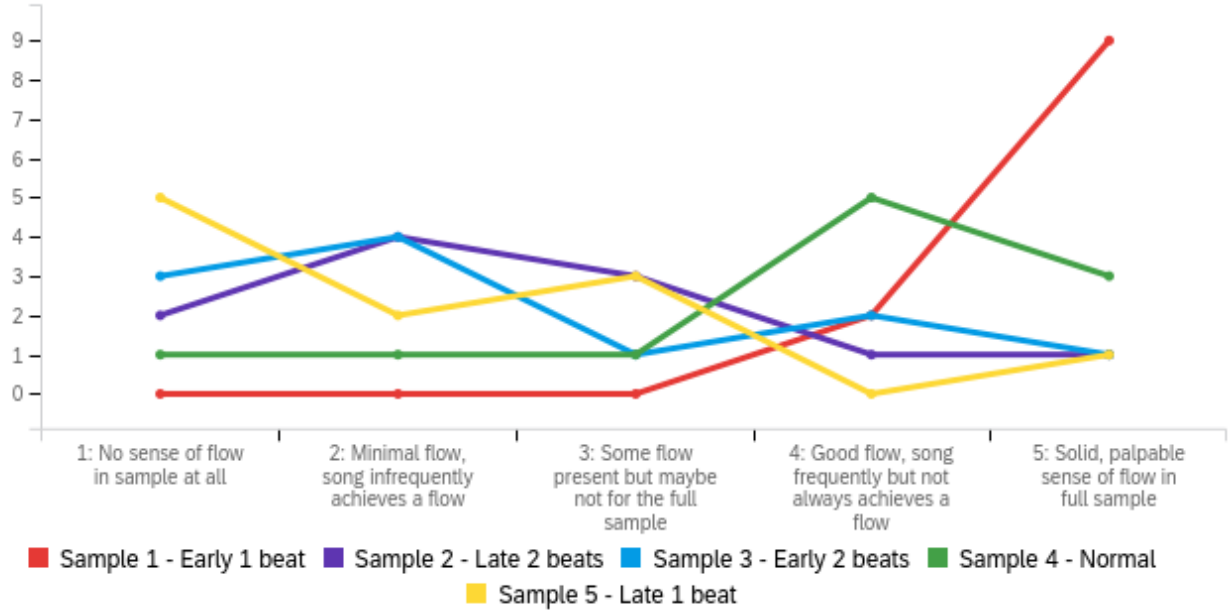
Summary description

The following data visualizations depict how samples were ranked against each other (pie graph) and individually scored (line graph) in terms of which flow was the best. Group 1 had 20 finished responses, and Group 2 had 12 finished responses that were incorporated into data

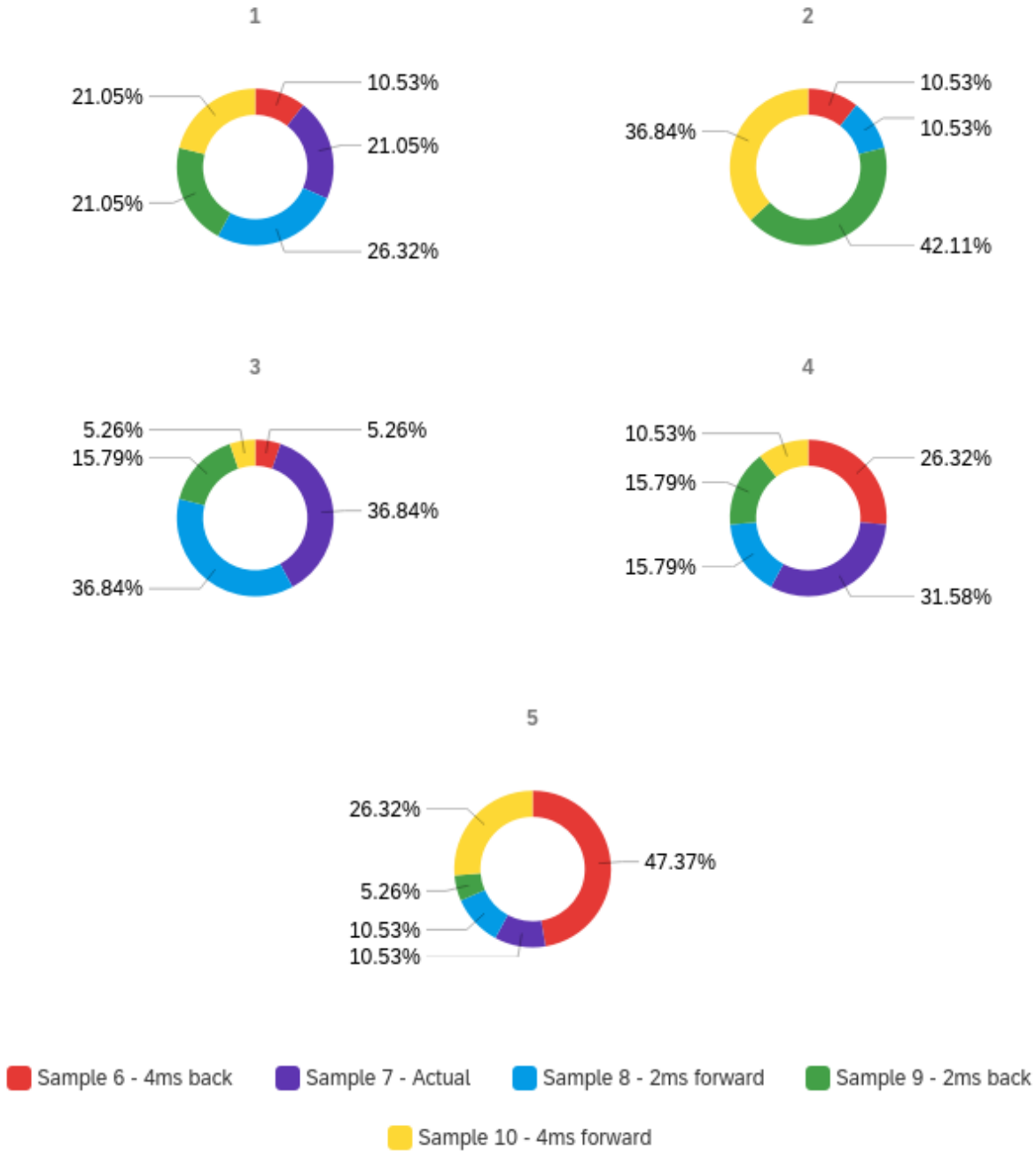
analysis. The graphs are listed in order of how they were listed across the two surveys (Set A, B, C, D):

Set A: ‘A Milli’ Acoustic

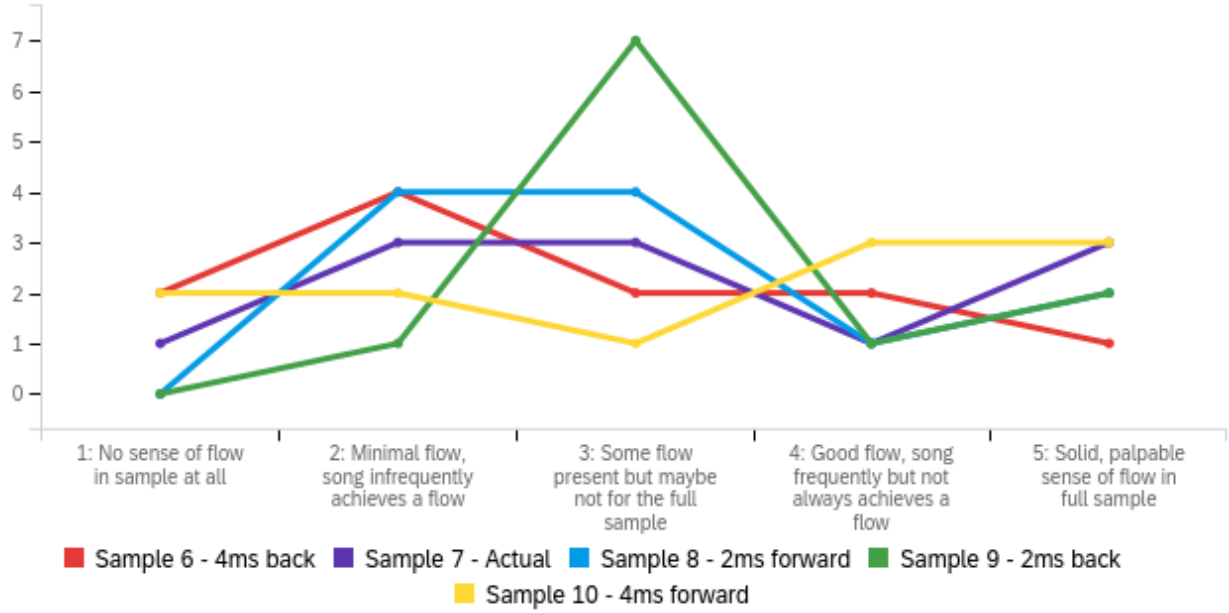
Set A: Rating Flow 1-5



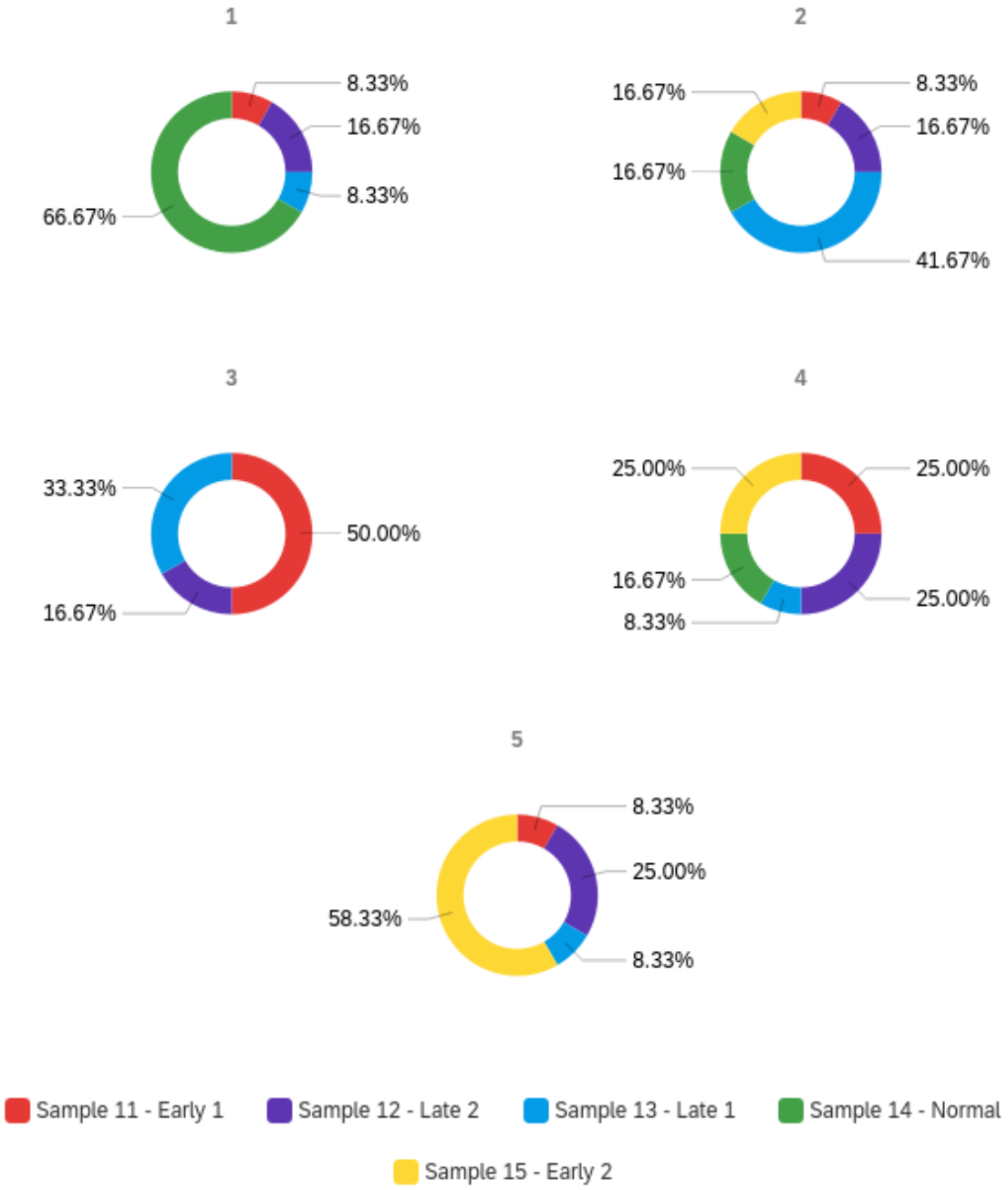
Set B: ‘Welcome to the Party’ Digital



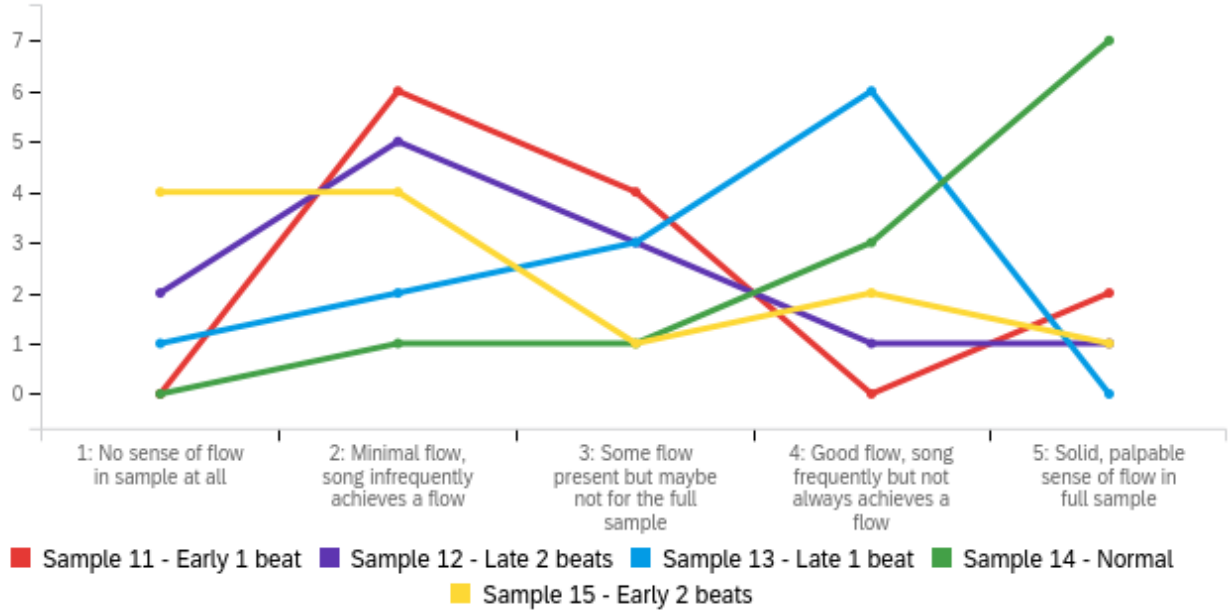
Set B: Rating Flow 1-5



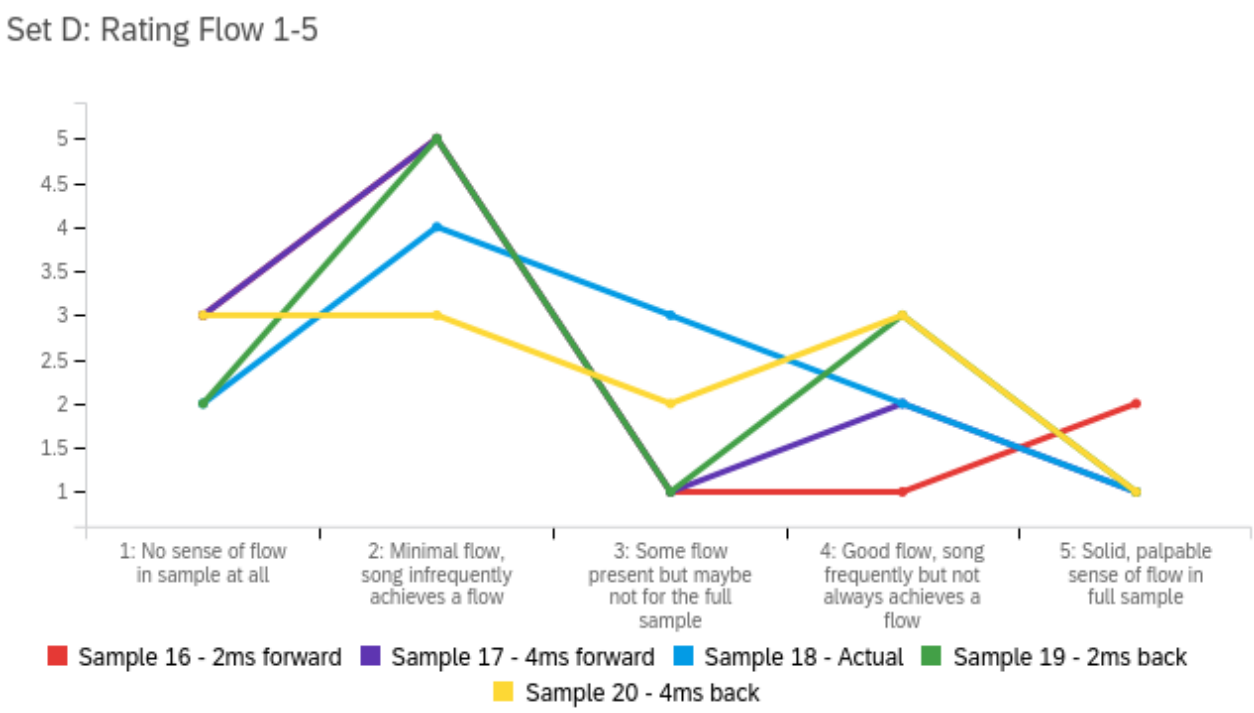
Set C: ‘Welcome to the Party’ Acoustic



Set C: Rating Flow 1-5



Set D: ‘A Milli’ Digital



The data represents how varied the participants' perception of flow was throughout both experiments. For the acoustic samples, while Set C's undoctored version was ranked to be the best flow, Set A's undoctored sample, on average, was ranked to be the second best behind the sample that was early one beat. It is difficult to identify any trends or shared findings from either digital sample. Most participants seemed to agree that the digital samples suffered worse flows than acoustic samples. More data would be needed in order to understand what relationship exists between these different variables for the robotic voices.

We qualitatively coded the free text responses of each of the questions that asked about participants' rationale behind their rankings and scores. This process requires researchers to identify words and themes that are reused across different participants' answers. There were a variety of ways that participants described how they tried to perceive flow. Many touched on the relationship between the lyrics and the beat, describing good flow as samples that were "more harmonious", "most natural sounding with the beat", and had "spaces in between verses…where the beat and the speed felt right." Bad flow was characterized as "the lyrics did not match up with the rhythm of the beat", "did not stay on beat", and unnatural pauses according to the beat. At least 4 participants admitted to guessing to perceive the differences between each sample.

We performed a two-tailed t-test between the flow ratings for the robotic version of A Milli and the flow ratings for the acoustic version of Welcome to the Party and found no significant difference between the means (H0: true difference in means = 0, HA: true difference in means ≠ 0; $p > 0.05$). We could not perform a t-test between any other two samples because the portion of the block 1 survey with the other two audio samples that asked for flow ratings was not functional for a period the survey was published; there was then too little data in order to

conduct a proper t-test. This provides evidence for an insignificant baseline difference in perceived flow quality between our robotic versions and the original acoustic versions.

Conclusions

Although it is intuitive to think that a robot could never rap as well or with as much flow as a real rapper, as one of our survey respondents commented, our data shows that there is no significant difference between the perceived flows of the live and robotic rappers. Additionally, on average, our participants seemed to favor the recordings that were either on-beat or 2 milliseconds before the beat. These findings mean we are unable to reject our null hypothesis. In a future experiment, we would perfect our survey and ideally have more participants to have a better understanding of our data. As mentioned earlier, four of our participants admitted to guessing occasionally on our survey questions. This could be because they were unclear of what “flow” means or because they were unable to tell significant differences between the off-beat tracks and the original. A widening of the offset range would likely produce a more apparent trend in our data. Our stimuli could have been more extremely offset to potentially define a range of microtiming where perceptual differences become insignificant. Additionally, the error in the first survey that didn’t enable participants to give objective scorings to the first two sets of samples could have given us more data to interpret, potentially changing our findings. As technology continues to advance it is possible that a robotic rapper could take these findings and become more popular than a real life rapper, but as technology exists current day, it is unlikely a robotic rapper will rise to fame.