Automatic Beat Alignment of Rap Lyrics

Sang Won Lee[1], Jieun Oh[2]
[1]Department of Management Science and Engineering
[2]Center for Computer Research in Music and Acoustics
{sangwlee, jieun5}@stanford.edu

Abstract *Rap is characterized by highly rhythmic delivery of words; the subtle ways in which syllables in the lyrics fit into the beats in the music lead to expressivity in rap music. Unfortunately, given a lack of notated score for the genre of rap, the task of determining the rhythm to existing rap performances must be carried out manually. In this paper, we explore a method of automatically aligning rap lyrics to beats using logistic regression. Specifically, we select top linguistic and audio features and compare the two resulting models. The audio model yielded about 3% higher accuracy than the linguistic model but at the cost of heavy computations in the data preparation stage. Potential applications to our technique include automatic rap rhythm transcription and style characterization and imitation of rap artists.*

1. INTRODUCTION

Rapping is a primary ingredient of hip-hop music that is characterized by highly rhythmic delivery of words with relatively small variation in pitch. The alignment of the syllables in the lyrics to the beats in the music thus defines the "sound and feel" of a particular rap song.

Unfortunately, the rhythm of rap is rarely notated, and the task of determining how the words in the lyrics fit wit the musical beats must be carried out manually through careful listening.

A technique using dynamic programming [1] tries to align lyrics and audio of popular music at a paragraph-to-segment level using hand-labeled lyrics; this algorithm finds the minimum-cost alignment path between lyrics and audio and further adjusts the results using vocal and non-vocal classifiers. This approach works on a higher structural level, so it has an average alignment error of 3.50 seconds and a standard deviation of 6.76 seconds, which is too coarse for our purpose of aligning words to music at the syllabic level (lasting only a fraction of a second).

One approach to this problem would be to take the audio signal of the entire song, perform source-separation between vocals and instrumental parts, and—assuming that the vocal part can be cleanly extracted—segment the audio at every syllabic onsets to perform analysis in the time- and frequency-domain. Unfortunately, extracting only the vocal part [2] and automatically segmenting the resulting speech audio at every syllable onsets [3,4] are challenging signal-processing maneuvers, even using state of the art techniques. So, we have manually created a data set resembling what would likely result from successful source separation and syllable segmentation, to serve as an input to our learning algorithm (*audio model*) and to compare the resulting alignment prediction against that of an easier alternative method.

The alternative approach under consideration is based on taking linguistic features in the symbolic domain (i.e. not relying on the audio signal). The merits of using features in the symbolic domain—and in particular linguistic features taken from the lyrics—have previously been explored to perform music genre classification [5]. Taking this idea, we also create a *linguistic model* with which to train parameters and calculate the probability of a given syllable in the lyrics falling on the musical beats. We compare this result with the *audio model*.

2. METHOD

2.1 Data

We chose *I'll be Missing You* by Puff Daddy (featuring Faith Evans) as the music for training and testing.[1]

Since our goal is to determine the syllables to which the musical beats fall, we broke down each word in the lyrics[2] into their constituent syllables.

[1] Ideally, parameters should be trained and/or tested on multiple songs to check robustness of our method. The extent to which parameters found on Puff Daddy's *I'll be Missing You* applies to (i) other renditions of the same song by a different artist, (ii) other songs by Puff Daddy, and (iii) a completely different style of rap, would be an interesting study to conduct as a next step.

[2] *I'll be missing you* comprised of 2 verses, each having 16 measures of 4/4 meter.

For example, the word *notorious* was broken down into four syllables: *no-to-ri-ous*.

2.2 Response Variable $y^{(i)}$

The song has a 4/4 time signature, so a phrase of text is spread across four major beats. Thus, we design a response variable $y_j^{(i)}$ with four components $1 \le j \le 4$, indicating whether syllable i falls on beat j. Optionally, syllables that do not fall on any of the four beats can be categorized as j=5.[3]

2.3 Feature Vector $x^{(i)}$

For each syllable i, we prepare features in the following three categories. (See Section 3 for a description of our feature selection method to choose the most informative features to make up our linguistic model and audio model.)

2.3.1 Relative position in phrase (A)

We created four features to convey the temporal order of syllables in a given phrase. To do so, we first calculated $rp_{ij} \in (0, 1]$, the relative position of syllable *i* in phrase:

$$rp_i = \frac{\text{index of syllable i in phrase}}{\#\text{ of syllables in phrase}}$$

From this, we calculated four features:

$$x^{(i)}_1 = |\, rp_i - 0.25 \,|$$
$$x^{(i)}_2 = |\, rp_i - 0.50 \,|$$
$$x^{(i)}_3 = |\, rp_i - 0.75 \,|$$
$$x^{(i)}_4 = |\, rp_i - 1.00 \,|$$

2.3.2 Linguistic Features (B)

We also annotated the following eleven linguistic features. Note that these features are relatively easy to create, based on a simple dictionary and phonetic transcription lookups. Thus, it would be quite feasible to automate this process to annotate a large volume of data.

1. **number of syllables***:* denotes the number of syllables in the word that the syllable came from. (i.e. assign 4 to each of the syllables in no-to-ri-ous).
2. **lexical accent**: encoded as 1 if the syllable is accented, and 0 otherwise.
3. **rhyme:** encoded as 1 if the syllable functions as a rhyme, and 0 otherwise[4]. This was hand-annotated, but could be automated by performing string matching on phonetic transcription, and through other sophisticated techniques for detecting internal and imperfect rhymes [6].
4. **parts of speech:** nine features indicating whether the word to which the syllable belongs is a noun, pronoun, verb, preposition, article, adjective, adverb, or conjunction.

2.3.3 Audio Features (C)

Finally, we added three audio features. Ideally, these features would be calculated using the vocal part that has been extracted from the original recording of the song. But because audio source separation is a difficult problem[5], we created a data set resembling what would likely result from successful source-separation, in order to get a rough idea of the prediction accuracy that can be achieved by these means.

Specifically, we created a recording of the rap part as closely as possible to Puff Daddy's performance. Then we segmented this audio at every syllable boundaries, and calculated the following features for each syllable.

1. **duration***:* length of the syllable in seconds
2. **power**: average power of signal (the mean square of a real signal)
3. **pitch**: average frequency of the syllable in hertz, normalized to [0,1]

	Feature vector $x^{(i)}$ = (A) + (B) + (C)							Response variable $y^{(i)}$				
	Relative Position (A)					Linguistic (B)	Audio (C)					
	Rel.Pos.	$x^{(i)}_1$	$x^{(i)}_2$	$x^{(i)}_3$	$x^{(i)}_4$	...	...	1	2	3	4	offbeat (5)
seems	1/11	.159	.409	.659	.909	Linguistic features: Selected from number of syllables per word, lexical accent, rhyme, and parts of speech	Audio features: Selected from duration, power, and pitch	0	0	0	0	1
like	2/11	.068	.318	.568	.818			1	0	0	0	0
yes	3/11	.023	.227	.477	.727			0	0	0	0	1
ter	4/11	.114	.136	.386	.636			0	0	0	0	1
day	5/11	.205	.045	.295	.545			0	1	0	0	0
we	6/11	.295	.045	.205	.455			0	0	0	0	1
used	7/11	.386	.116	.114	.364			0	0	1	0	0
to	8/11	.477	.227	.023	.273			0	0	0	0	1
rock	9/11	.568	.318	.068	.182			0	0	1	0	0
the	10/11	.659	.409	.159	.091			0	0	0	0	1
show	11/11	.750	.500	.250	.000			0	0	0	1	0

Figure 1: Syllable-Beat Mapping and Feature Vector

[3] This design would assign a class to every syllable, and thus would also allow the use of softmax regression as an alternative machine-learning algorithm.

[4] Based on the work of [6], we could also consider encoding rhyme as a continuous value between 0 and 1 such that half rhymes or internal rhymes are assigned a smaller number compared to perfect rhymes.

[5] Source separation on multi-channel audio using ICA works well if each channel represents input from a microphone placed in distinct locations. But the problem becomes very difficult when the different microphone inputs have been combined, flattened, and digitally manipulated to create effects.

2.4 Learning Algorithm

For each output component j (representing musical beats) and feature k, we performed logistic regression using batch gradient ascent. We updated $\theta_{j,k}$ until convergence (α=0.01, 3000 iterations):

$$\theta_{j,k} := \theta_{j,k} + \alpha(y^{(i)}_{j,k} - h_\theta(x^{(i)}))(x^{(i)}_{j,k})$$

We then made our hypothesis,

$$h_{\theta,j}(x^{(i)}) = g(\theta_j{}^T x) = \frac{1}{1+e^{-\theta_j{}^T x}}$$

Train error and test error calculations were made by summing up the correct predictions; that is, instances in which ($h_j(x^{(i)})>0.5$ *and* $y^{(i)}==1$) *or* ($h_j(x^{(i)})<0.5$ *and* $y^{(i)}==0$). But we used a separate heuristics for making the final syllables-to-beat alignment prediction, the details to which are presented in Section 4.

3. FEATURE SELECTIONS

To determine the extent to which each feature improves our hypothesis, we carried out feature selection using a hybrid method of filter selection and forward search, based on the feature's average correlation score across output components.

3.1 Feature Selection Score

For each output component j, we calculated the feature selection score $S_j(k)$ to be the correlation between x_k and y_j to measure how informative each feature x_k is about the class label y_j. We then sorted features in decreasing order of $S(k) = \frac{1}{5} \cdot \sum_j |S_j(k)|$.

	rhyme	pronoun	verb	article	noun	prep.
$S_1(k)$	-0.042	-0.086	-0.033	-0.076	-0.001	0.072
$S_2(k)$	-0.009	-0.057	-0.105	-0.076	0.158	-0.115
$S_3(k)$	0.023	-0.142	0.233	-0.076	-0.054	-0.040
$S_4(k)$	0.682	-0.081	0.072	-0.075	0.112	-0.113
$S_5(k)$	-0.406	0.230	-0.105	0.190	-0.135	0.123
S(k)	0.232	0.119	0.110	0.099	0.092	0.092

accent	adjective	adverb	conjunction	# syllables
0.057	-0.037	0.109	0.075	0.046
-0.015	0.156	0.072	-0.059	0.046
0.092	0.060	-0.040	-0.059	-0.081
0.097	-0.034	0.038	-0.058	-0.044
-0.145	-0.092	-0.112	0.063	0.020
0.081	0.076	0.074	0.063	0.047

Figure 2: Linguistic Features: Selection Scores

	duration	pitch	power
$S_1(k)$	0.099	0.290	0.233
$S_2(k)$	0.116	0.158	0.070
$S_3(k)$	0.186	-0.009	-0.084
$S_4(k)$	0.594	-0.285	-0.223
$S_5(k)$	-0.622	-0.100	0.000
S(k)	0.323	0.122	0.168

Figure 3: Audio Features: Selection Scores

Figure 2 and **Figure 3** show the selection scores for linguistic features and audio features, respectively, in decreasing order of $S(k)$.

3.2 Feature Selection Method

We use a hybrid of filter feature selection and forward search to pick features to insert in our linguistic model and audio model.

Specifically, having ranked our n features (n=11 for linguistic; n=3 for audio) in decreasing order *of* $S(k)$, we construct a set_l consisting of top l features, for $1 \le l \le n$. In other words, for each linguistic and audio domain, the l^{th} most informative feature gets added to set_l. We desire to determine the optimal l^* for the linguistic model (l_l^*) and audio model (l_a^*) based on the training error and test error computed.

The rationale for using this hybrid model is that it involves comparing error rate across just n models, as opposed to $O(n^2)$ required by forward search. But because we choose l^* based on our calculation of training and test errors (rather than arbitrarily deciding on it beforehand, as is done in filter feature selection), we are able to make a computationally efficient decision about the number of features that will our models will comprise of. [6]

We use k-fold cross validation (k=2): we (i) train on verse 1 and test on verse 2, and (ii) train on verse 2 and test on verse 1, and sum the two results. **Figure 4** and **Figure 5** show comparisons between training and error rates as we append l^{th} feature in our models.

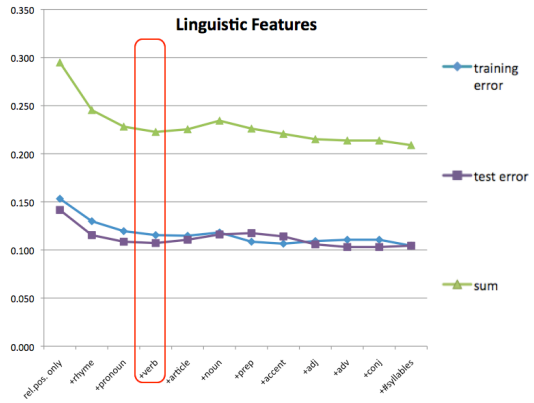

Figure 4: Linguistic Features Training and Test Error Comparisons

[6] We confirmed using the audio set (n=3) that our hybrid selection method (involving comparisons of 3 different sets) yielded the same feature-selection order as forward search (involving comparisons of 6 different sets).

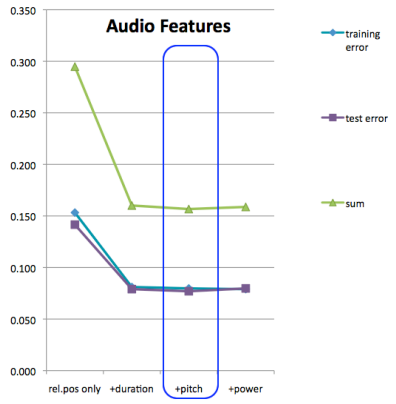

Figure 5: Audio Features Training and Test Error Comparisons

Based on test and training error comparisons, we chose l_l*=3 (rhyme, pronoun, verb) and l_a*=2 (duration, pitch). Thus, the features that make up our models are as follows:

Linguistic Model:
[*relative position in phrase*] + [rhyme, pronoun, verb]

Audio Model:
[*relative position in phrase*] + [duration, pitch]

4. ALIGNMENT RESULT

Given a phrase of lyrics to be spread across four beats of a measure, we assigned syllables to beats using the following heuristics:

For each measure 1≤m≤16:
  For each beat 1≤j≤4:
    Assign to beat *j* the syllable *i* with highest $h(x^{(i)})$,

$$h_{\theta,j}(x^{(i)}) = g(\theta_j^T x) = \frac{1}{1+e^{-\theta_j^T x}}$$

Consequently, we are guaranteed to have exactly one syllable assigned to beat j, even if $h_{\theta j}(x^{(i)})$ happens to be less than 0.5 for all syllables *i* in the measure. **Figure 6** compares the accuracy of alignments attained from our models. See Appendix for a comparison between the linguistic and audio models' alignment result for verse 1 and verse 2.

	Features	Train Error	Test Error	Alignment Accuracy
Linguistic Set	A (all) + B (rhyme, pronoun, verb)	11.5 %	10.7 %	208 / 291 syllables
Audio Set	A (all) + C (duration, pitch)	8.0 %	7.7 %	229 / 291 syllables

Figure 6: Accuracy Comparisons

5. DISCUSSION

Our audio model yielded higher alignment accuracy than our linguistic model. The train error difference was 3.6%, and the test error difference was 3.0%. However, the difference in accuracy may not be worth the technical trouble of creating the audio model. Audio features—such as power and pitch—require heavier computations in the signal-processing domain. But even a greater challenge with preparing the audio features lies in obtaining a clean source-separated vocal part that has been segmented at every syllable onsets. These are time-consuming processes to carry out manually, and have dissatisfying results when performed automatically using even state of the art techniques.

In contrast, linguistic features are relatively easy to prepare using a dictionary and phonetic transcriptions. Given that rhyme was the most informative linguistic feature in predicting beat alignment, we may be able to improve our accuracy significantly by using automatic rhyme-detection techniques that can spot even the subtle internal rhymes and imperfect rhymes [6]. This would be a good next step for improving our current design.

Additionally, a separate model on anacrusis prediction (i.e. determining whether the first syllable of phrase falls on the downbeat or not) should be integrated into our design to improve alignment of beat 1 and beat 2.

Finally, future studies should evaluate the extent to which our model can be applied to songs by different rap artists, or even to different schools of rap. This will open up the possibilities of characterizing styles of rap artists or performances based on the temporal-rhythmic flow of rap.

Potential applications for our model include automatic creation of extended LRC files, which contain song lyrics that are time-stamped at the word-level, for the genre of rap. These files are currently manually created for use in karaoke, but with the ability to align rap lyrics to musical beats based on the characteristic style, creation of these files can be largely automated. Online music streaming services can also use our technique to display lyrics right as they are being played.

6. ACKNOWLEGMENT

The authors wish to thank Professor Andrew Ng and the teaching assistants of cs229 for their feedback and guidance.

7. REFERENCES

[1] Lee, Kyogu, and Markus Cremer: "Segmentation-Based Lyrics-Audio Alignment Using Dynamic Programming," *Proceedings of the 9th International Conference for Music Information Retrieval* (ISMIR 2008).

[2] Vincent, Emmanuel, Hiroshi Sawada, Pau Bofill, Shoji Makino, and Justinian Rosca: "First Stereo Audio Source Separation Evaluation Campaign: Data, Algorithms and Results," in *Proc. Int. Conf. Independent Compon. Anal. Signal Separation (ICA'07)*, 2007, CD-ROM.

[3] Shastri, L. Chang, S. and Greenberg, S.: "Syllable Detection and Segmentation using Temporal Flow Model Neural Networks," *Proc. XIVth Int. Cong. Phon. Sci.*, pp. 1721-1724, 1999

[4] Ying, D.W., W. Gao, and W.Q. Wang: "A New Approach to Segment and Detect Syllables from High-Speed Speech," in Eurospeech 2003, 765-768.

[5] Mayer, Rudolf, Robert Neumayer, and Andreas Rauber: "Rhyme and Style Features for Musical Genre Classification by Song Lyrics," *Proceedings of the 9th International Conference for Music Information Retrieval* (ISMIR 2008).

[6] Hirjee, Hussein and Daniel G. Brown: "Automatic Detection of Internal and Imperfect Rhymes in Rap Lyrics," *Proceedings of the 10th International Society for Music Information Retrieval Conference* (ISMIR 2009).

8. APPENDIX

Actual Beat **Ling. Predict** **Audio Predict** **Ling. Predict & Audio Predict**

Seems like yesterday we used to rock the show
I laced the track, you locked the flow
So far from hanging on the block for dough
Notorious, they got to know that
Life ain't always what it seem to be
Words can't express what you mean to me
Even though you're gone, we still a team
Through your family, I'll fulfill your dream
In the future, can't wait to see
If you open up the gates for me/ Reminisce
some time, the night they took my friend
Try to black it out, but it plays again
When it's real, feelings hard to conceal
Can't imagine all the pain I feel
Give anything to hear half your breath (half your breath)
I know you still living your life, after death

(verse 2)

It's kinda hard with you not around
Know you in heaven smiling down
Watching us while we pray for you
Every day we pray for you
Till the day we meet again
In my heart is where I'll keep you friend
Memories give me the strength I need
to proceed/ Strength I need to believe
My thoughts Big I just can't define
Wish I could turn back the hands of time
Us in the Six, shop for new
clothes and kicks/ You and me taking flicks
Making hits, stages they receive
you on/ Still can't believe you're gone
Give anything to hear half your breath (half your breath)
I know you still living your life, after death