TEMPORAL MUSIC CONTEXT IDENTIFICATION WITH USER LISTENING DATA

Cameron Summers
Gracenote
csummers@gracenote.com

Phillip Popp
Gracenote
ppopp@gracenote.com

ABSTRACT

The times when music is played can indicate context for listeners. From the peaceful song for waking up each morning to the traditional song for celebrating a holiday to an up-beat song for enjoying the summer, the relationship between the music and the temporal context is clearly important. For music search and recommendation systems, an understanding of these relationships provides a richer environment to discover and listen. But with the large number of tracks available in music catalogues today, manually labeling track-temporal context associations is difficult, time consuming, and costly.

This paper examines track-day contexts with the purpose of identifying relationships with specific music tracks. Improvements are made to an existing method for classifying Christmas tracks and a generalization to the approach is shown that allows automated discovery of music for any day of the year. Analyzing the top 50 tracks obtained from this method for three well-known holidays, Halloween, Saint Patrick's Day, and July 4th, precision@50 was 95%, 99%, and 73%, respectively.

1. INTRODUCTION

With the ever increasing amount of recorded music, structured metadata is important to organize it. For holiday music, there is some metadata that indicates an association with a music track, often Christmas [1], but comprehensive labeling for other holidays is still lacking. One reason for this is the varying nature of holiday music. Across geographies, cultures, and time, what music is used to celebrate holidays changes dramatically. There is a bit of a paradox as to whether a holiday track is so because the artist recorded it for that purpose or the listeners use it to celebrate [1] . Given this complex landscape of holiday music, manual labeling of a large number of music tracks is difficult, time consuming, and costly. Methods for automated labeling are desirable for large scale organization, further improving the capabilities of music search and recommendation systems.

One automated approach is using text search of track names or album names for keywords also associated with the target holiday [2]. For example, tracks with the keywords ”winter” or ”spooky” may be likely associated with Christmas or Halloween, respectively. This approach has drawbacks, however. First, it requires experts to create keywords lists, which can be costly or difficult, particularly for music in different languages. Second, the keywords do not guarantee correct track-holiday association, particularly for ambiguous words like ”whiskey”, which could be linked contextually to Saint Patrick's Day or simply drinking beverages. This problem is compounded when using multiple keywords, as is required for a comprehensive set of tracks.

Another automated approach for labeling holiday music is through user crowdsourcing. LastFM (last.fm), for example, allows users to add their own tags to music tracks, which include tags for some holidays like Halloween [3]. This has the advantages of outsourcing the work of labeling and getting a better representation of the holiday music preferences of a larger number of listeners. But this too has drawbacks. Users tend to only label popular tracks and artists, leading to imbalanced coverage. The quality of these tags can suffer to due misspellings, synonyms, biases, or dishonest labeling. And the users providing tags are still typically a small subset of the total users of a service [4].

Alternatively, leveraging user listening data avoids the quality issues associated with user tagging and keyword association, can utilize the entire user base, and is language agnostic. Researchers have studied temporal dynamics of user data previously to understand context. [5] examined temporal context to improve biosurveillance. [6] and [7] classified web search queries using features in the popularity signal over time and in music, [8], [9], and [10] show the usefulness of temporal analysis in recommendations systems. An approach proposed by [11] exploits user listening data to automatically label tracks as associated with Christmas. However, the approach performs poorly for other holidays in our experiments. In this paper we show that the methodology in [11] can be improved and generalized to discover tracks associated with other holidays throughout the year.

[1] The interpretation of the authors of what is truly holiday music is the latter.

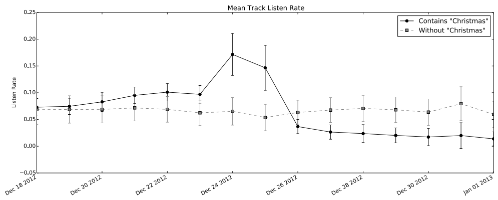

Figure 1. Mean listen rates R for tracks above 1,500 total listen threshold with ”Christmas” in track or album name (solid line) and tracks without (dotted line) for December 18, 2012 - January 1, 2013.

2. METHODOLOGY

[11] hypothesized that the listening signals of two tracks, one associated with a holiday and one not, will have differing and detectable patterns on and around that holiday. This can readily be seen for Christmas tracks and non-Christmas tracks in Figure 1. In this section we show the methodology in [11] for detecting Christmas tracks and propose improvements.

2.1 Listening Rates

The form of the raw data is listening events in which a known user has listened to a known track at a date and time. If a track is associated with a day, we expect users to engage more relative to other time periods. The signal used in [11] can be described as user engagement,

$$E_{ij} = \sum_{k=1}^{U} c_{ijk} \tag{1}$$

which is the total number of listens for all users for track i in time period j.

In Eqn (1), c_{ijk} is an element of C, and $C \in \mathbb{R}^{T \times W \times U}$ where T is the number of tracks, W is the number of time periods, and U is the number of users. To account for differences between popularity of tracks, the E was normalized across the periods of time as described by

$$R_{ij} = \frac{E_{ij}}{\sum_{l=1}^{W} E_{il}} \tag{2}$$

which were the listen rates used to train the Christmas model.

We propose a new signal based on absolute user engagement, $\hat{E}$. Given the function

$$f(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

$\hat{E}_{ij}$ is the number of users who listened to a track i in time period j and is calculated by

$$\hat{E}_{ij} = \sum_{k=1}^{U} f(c_{ijk}) \; . \tag{4}$$

Number of Records	4,819,992,847
Number of Users	1,648,796
Number of Tracks	13,227,376
Date Range	January 2012 - February 2013

Table 1. Dataset of listening records.

This is similarly normalized across time periods to get new listen rates

$$\hat{R}_{ij} = \frac{\hat{E}_{ij}}{\sum_{l=1}^{W} \hat{E}_{il}} \; . \tag{5}$$

The intuition is to limit the effect of repeat plays among individual users. In estimating cultural preferences, there is likely more information gained when 100 users listen to a track once than when one user listens to a track 100 times.

2.2 Detection

For detection, [11] fit a multi-variate Gaussian with listen rates of only Christmas tracks with parameters $\theta_{Christmas}$ composed of mean, μ_{R_j}, and covariance matrix, Σ_{R_j}. Given a new track with listen rates for the same time periods, $\mathbf{x} = [R_1, R_2, ..., R_W]$, the metric for detection was

$$P(\mathbf{x}|\theta_{Christmas}) = \prod_{j=1}^{W} \mathcal{N}(\mathbf{x}|\mu_{R_j}, \Sigma_{R_j}) \; . \tag{6}$$

We propose another metric using the posterior probability from a direct application of Bayes' Rule in

$$P(\theta_c|\mathbf{x}) = \frac{P(\mathbf{x}|\theta_c) * P(c)}{P(\mathbf{x}|\theta_c) * P(c) + P(\mathbf{x}|\theta_n) * P(n)} \tag{7}$$

where c subscript represents Christmas and n subscript represents non-Christmas. This includes a model for non-Christmas tracks and prior probabilities $P(c)$ and $P(n)$, which represent the proportion of Christmas and non-Christmas tracks in matrix C, respectively. The priors in particular are important because of the small number of Christmas tracks in the dataset. $P(\mathbf{x}|\theta_n)$ is calculated from Eqn (6) where μ_{R_j} and Σ_{R_j} are calculated using all non-Christmas tracks.

2.3 Dataset

This study uses the same internal Gracenote dataset of online radio listening records in North America as [11]. Some basic information is shown in Table 1. Each record in the dataset represents one listen of a track by one user and provides User ID, Date, Time, and Track ID. Track metadata is also available such as track name, album name, and artist name.

Min. Listens	All Tracks	Christmas Tracks
1,500	338,406	4,732
500	767,116	10,647
200	1,397,032	18,170
100	2,087,863	26,134
10	5,906,307	68,582
1	10,207,335	118,515

Table 2. Track distribution at each threshold of minimum listens.

3. CHRISTMAS

3.1 Experiments

In these experiments, we compared the performance of the signals and prediction metrics in Section 2. As in [11], we generated a ground truth of Christmas tracks by searching for keyword "Christmas" in track names and album names. We defined the window radius, r_w, as the number of consecutive days before and after the target holiday, December 25, 2012, such that the window length $W = 2 * r_w + 1$. Since [11] showed an increase in performance with increasing popularity of tracks, we use the same thresholds of minimum listens in the dataset (1,500, 500, 200, 100, 10, 1) for direct comparison. Table 2 shows the distribution of tracks for each threshold.

For each listen rate in Section 2.1, a matrix was constructed from the dataset using tracks above the specified threshold, all users, and W days. We varied W by choosing r_w ranging 1 to 30 to capture the signal up to one month before and a month after December 25. The matrix was randomized and split into train (60%) and test (40%) sets on the first dimension. Two single component Gaussian Mixture Models, for Christmas and non-Christmas, were trained with the training set in a supervised manner with each training example a track and features the listen rates for each day in the signal window. Classification was performed with each metric in Section 2.2 on the test set and the area under the Receiver Operating Characteristic (AUROC) was calculated for evaluation.

3.2 Results

Figure 2 and Figure 3 show the AUROC against the window radius for the proposed listen rate, $\hat{R}$, and each prediction metric. Observing the difference in y-axis scale, the most notable difference between the figures is an increase in performance across all thresholds and signal lengths for the posterior probability. In particular, the lowest two thresholds have quite large increases of about 0.15 at each signal length.

Among tracks with the strongest listening signals, there is a small decay with increased window length. In contrast, the weakest listening signals show a large boost in performance with increased signal length. Similar plots for listen rates R are not shown because they track very closely and mostly just below the trends for $\hat{R}$. Lastly, Table 3 shows the maximum AUROC value for each threshold across signal lengths as a measure of overall performance. The proposed signal and prediction metric give the highest AUROC for the top four thresholds, and the signal from [11] with the proposed prediction metric have slightly higher AUROC for the bottom two thresholds.

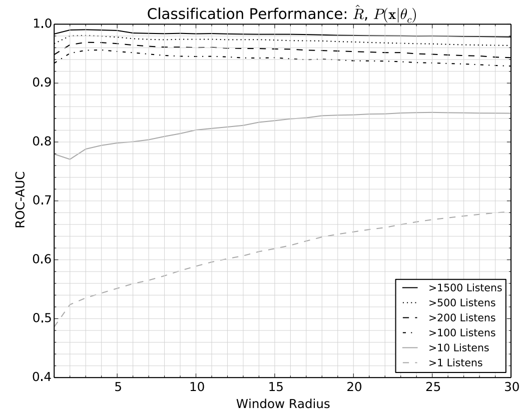

Figure 2. AUROC for each listen threshold for listen rate $\hat{R}$ and prediction metric $P(\mathbf{x}|\theta_c)$

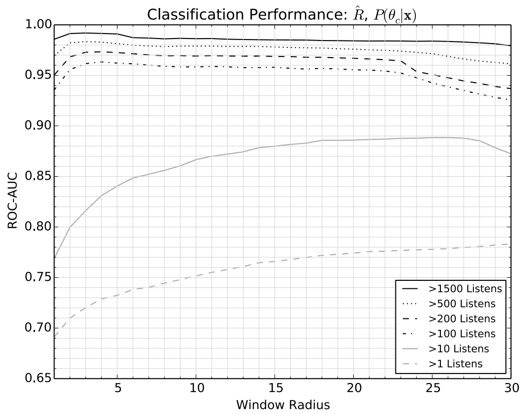

Figure 3. AUROC for each listen threshold for listen rate $\hat{R}$ and prediction metric $P(\theta_c|\mathbf{x})$

3.3 Analysis

The posterior probability performs better than the likelihood because the inclusion of a non-Christmas model provides additional discriminative information. There is a lot of complexity in the non-Christmas tracks that is not modeled well by a single Gaussian with a mostly uniform distribution as shown in Figure 1. This suggests that incorporating models for other common signal shapes such as those of newly released tracks might further improve performance.

The signal length effects the performance in different ways. Tracks with the strongest listening signals perform best more with localized time window. We believe this is due primarily to higher variability of listening rates leading up to Christmas. The Christmas holiday is celebrated

	$P(\mathbf{x}\|\theta_c)$		$P(\theta_c\|\mathbf{x})$	
Threshold	R	$\hat{R}$	R	$\hat{R}$
1,500	0.987	0.991	0.989	**0.992**
500	0.975	0.981	0.978	**0.983**
200	0.964	0.969	0.969	**0.973**
100	0.950	0.956	0.958	**0.963**
10	0.851	0.850	**0.892**	0.888
1	0.680	0.682	**0.784**	0.783

Table 3. Best AUROC for any signal window.

for many days before, and the signals during this time may be less stable than nearby December 25th. Tracks with the weakest listening signals perform best with a larger time window performs. This is likely because there is more information available with a longer signal, even if a small amount. Tracks with 50 plays in the dataset average only one play in ten days so capturing enough discriminatory information for detection requires a longer signal window.

The proposed signal of user counts, $\hat{R}$, has a smoothing effect over the signal of play counts, R, boosting performance. With tracks of stronger signals this appears to be more discriminating as shown in Table 3. But tracks with weaker signals, this seems to remove some useful information, which would explain why the play counts, R, perform slightly better at the two lowest thresholds. No single configuration appears to give optimal performance for this task.

4. HOLIDAY GENERALIZATION

We are interested in detecting track-temporal context associations for many days other than Christmas. Directly repeating the procedure in Section 3 for other holidays produced poor results on the dataset in Section 1. We believe this is because the ground truth generated from keywords is much less clean. Since many other holidays have a smaller music repertoire than Christmas, discriminative keywords like the holiday names generate too few tracks with strong listening signals for model training. And less discriminative keywords inadvertently include tracks not associated with the holiday, similarly compromising training.

Instead, the Christmas model in Section 2 can be reinterpreted as a holiday model with parameters $\theta_{holiday}$ composed of the same mean, μ_{R_j}, and covariance matrix, Σ_{R_j} as Eqn (6). Now a new track with listen rate signal of the same length, W, centered on a *different target holiday*, $\mathbf{x} = [R_1, R_2, ..., R_W]$, can be detected with Eqn (6) or Eqn (7).

4.1 Experiments

In these experiments, we show the performance of detection on three other holidays. Since the dataset in Section 1 is from users in North America, we chose Halloween, Saint Patrick's Day, and U.S. Independence Day as they are well-known holidays in North America and likely to have music associations. For the best results, we use only tracks with

Saint Patrick's	U.S. Independence	Halloween
95%	73%	99%

Table 4. Average precision@50 for holiday track detection.

strong listening signals - above 1,500 total listens in the dataset - and the best performing listen rate and prediction metric from Section 3, $\hat{R}$ and Eqn (7).

We constructed the training set feature matrix using $W = 15$, implying $r_w = 7$ and $\hat{R}_{i8}$ is the listen rate for track i on December 25, 2012. Again, we trained two single component Gaussian Mixture Models, holiday and non-holiday, in a supervised manner. We constructed the test set feature matrix similarly with $W = 15$, meaning $\hat{R}_{i8}$ is the listen rate for track i on the target holiday.

We calculated the probability of each track in the test set with Eqn (7) and ranked the tracks from highest probability to lowest for analysis. We chose precision@k to provide a general measure of relevance of tracks. We set k=50 with the assumption that there are at least 50 tracks truly associated with each holiday in order to better attribute errors in detection to the methodology. Three music content experts examined each list and labeled tracks as relevant to the holiday or not. We averaged the results to get a single value of precision@50 for each holiday.

4.2 Results

Table 4 shows the average precision@50 of holiday track detection as indicated by three music experts. Halloween and Saint Patrick's Day had high values at 99% and 95%, respectively, and U.S. Independence Day was lower at 73%. The mean probability of the all tracks according to the holiday model was 99.9%. The distribution of incorrect tracks for U.S. Independence Day is skewed toward the bottom of the list.

The top 10 tracks for each holiday are shown in Section 4.2.1 - Section 4.2.3 to further characterize the results. All of these tracks had a probability of 1.0 according to the holiday model. The ordering for each track is track name, artist name.

4.2.1 Top 10 Saint Patrick's Tracks

1. When Irish Eyes Are Smiling, Bing Crosby
2. Maloney Wants A Drink, The Clancy Brothers
3. Sally MacLennane, The Pogues
4. When Irish Eyes Are Smiling, The Irish Folk
5. Danny Boy, Irish Drinking Songs
6. Water Is Alright In Tay, The Clancy Brothers
7. Whiskey In The Jar, The Clancy Brothers
8. A Pair of Brown Eyes, The Pogues
9. The Black Velvet Band, Irish Drinking Songs
10. Grace, Jim McCann

4.2.2 Top 10 U.S. Independence Tracks

1. America, Barry White
2. Independence Day, Elliott Smith

3. Proud to be an American, Tiki
4. Stars And Stripes Forever, John Philip Sousa
5. Justice And Independence '85, John Mellencamp
6. 4th of July, X
7. Our Country (Rock Version), John Mellencamp
8. America the Beautiful, Blake Shelton and Miranda Lambert
9. This Is My Country, The Impressions
10. God Bless The U.S.A., Lee Greenwood

4.2.3 Top 10 Halloween Tracks

1. Purple People Eater, Halloween Hit Factory
2. "Dr. Who" Theme Song, Mannheim Steamroller
3. Graveyard Of The Living Dead, Halloween Sound Effects
4. Werewolves - Scary Halloween Sound Effects, Halloween Sound Effects
5. Dracula's Organ - Scary Halloween Sound Effects, Halloween Sound Effects
6. Creatures Of The Night (Original Mix), Mannheim Steamroller
7. This Is Halloween, The Countdown Kids
8. Hall of Screams - Scary Halloween Sound Effects, Halloween Sound Effects
9. Scary Halloween Haunted House, Sound Fx
10. Grimly Fiendish (Album Edit Version), The Damned

5. ANALYSIS - HOLIDAY

The high values for precision@50, particularly those of Halloween and Saint Patrick's, show that a model trained with user data around Christmas is effective in identifying daily music-temporal context associations. The lower precision@50 of U.S. Independence Day and the incorrect tracks being skewed towards the bottom of the list suggests that our assumption of at least 50 associated tracks for the holiday may be incorrect. Flaws in the methodology could also be the cause.

In particular, the assumption that Christmas listening signals have a distribution that matches those other holidays closely is likely flawed. Looking at Christmas signal in Figure 1 and the Saint Patrick's signal in Figure 4, they are similar but do not match exactly. The Christmas tracks have two peaks, December 24 and December 25, and the Saint Patrick's tracks have a single peak on March 17. With shorter signal lengths, this difference is pronounced and gives poor results for detecting other holiday tracks. This is why the experiments in Section 4.1 used $r_w = 7$, and not the optimal from Section 3, $r_w = 3$.

Among the incorrect U.S. Independence tracks, nearly one-third were from a single album by electro-punk band Frittenbude. This highlights other possible reasons for increased engagement such as marketing pushes. This album appears to have been released in the summer of 2012 and the synchronized rise and fall of the album's initial listening could be one explanation. In this case and other one-time events, album releases happen just once and could be separated from the more cyclical holiday listening with multiple years of data.

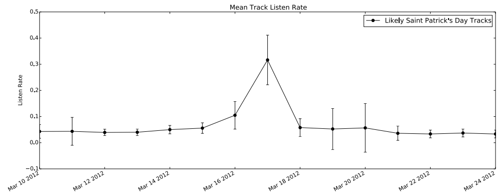

Figure 4. Mean listen rates R for top 100 predicted Saint Patrick's Day tracks from March 10, 2012 to March 24, 2012.

The effectiveness of general holiday detection implies one obvious commercial application: an automated, "always-on" seasonal radio station. With multiple years of data, the results likely could be improved and characterized by their change over time. Also, the addition of location data could highlight geographical differences for improved recommendations. For example, since our dataset is primarily North America, the tracks in Section 4.2.1 may be poor recommendations to users celebrating Saint Patrick's in Ireland or other parts of the world.

6. FUTURE WORK

The issues with matching the signal shapes of Christmas tracks to other holidays suggest room for improvement. Artificial templates or hand labeling a holiday ground truth could estimate the target distributions more accurately. Although labeling track-temporal context associations with user data has advantages over the other automated methods as outlined in Section 1, combining these methods could produce superior results. Lastly, applying this methodology at additional time resolutions (e.g hours, weeks, months) or exploring how these contexts interact with user data (e.g. age, geography, personality) could further enrich the user listening experience.

7. CONCLUSION

This study showed improvements to previous method for detecting Christmas tracks from user listening data and generalized the method to detect tracks for other holidays. The proposed improvements showed small increases of about 0.01 maximum AUROC for the most popular tracks but larger improvements of about 0.1 maximum AUROC for less popular tracks. Detection of Halloween, Saint Patrick's Day, and July 4th tracks was promising with precision@50 at 95%, 99%, and 73%, respectively.

8. REFERENCES

[1] Christmas comes early to The Echo Nest (The Echo Nest Blog) http://blog.echonest.com/post/35845347430/christmas-comes-early-to-the-echo-nest

[2] Top 50 Love Songs of All Time (Billboard)

http://www.billboard.com/articles/list/1538839/top-50-love-songs-of-all-time

[3] Last.fm (Last.fm) http://www.last.fm/charts/toptags

[4] P. Lamere. ”Social Tagging and Music Information Retrieval.” *Journal of New Music Research*, Vol. 37 No. 2 pp. 101-114, 2008.

[5] Reis, Ben Y., Marcello Pagano, and Kenneth D. Mandl. ”Using temporal context to improve biosurveillance.” Proceedings of the National Academy of Sciences 100.4 (2003): 1961-1965.

[6] Kulkarni, Anagha, et al. ”Understanding temporal query dynamics.” Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011.

[7] M. Shokouhi: Shokouhi, Milad. ”Detecting seasonal queries by time-series analysis.” Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. ACM, 2011.

[8] Park, Chan Ho, and Minsuk Kahng. ”Temporal dynamics in music listening behavior: A case study of online music service.” Computer and Information Science (ICIS), 2010 IEEE/ACIS 9th International Conference on. IEEE, 2010.

[9] Carneiro, Mrio Joo Teixeira. ”Towards the discovery of temporal patterns in music listening using Last.fm profiles.” Dissertation, 2012.

[10] Hidasi, Balzs, and Domonkos Tikk. ”Context-aware recommendations from implicit data via scalable tensor factorization.” arXiv preprint arXiv:1309.7611, 2013.

[11] C. Summers, P. Popp. ”Large Scale Discovery of Seasonal Music with User Data.” *User Modeling, Adaptation and Personalization: 23rd International Conference, UMAP 2015, Extended Proceedings*. Springer, 2015.