LEARNING SHARED VECTOR REPRESENTATIONS OF LYRICS AND CHORDS IN MUSIC

Timothy Greer, Karan Singla, Benjamin Ma, and Shrikanth Narayanan

Signal Analysis and Interpretation Lab, University of Southern California, USA

ABSTRACT

Music has a powerful influence on a listener's emotions. In this paper, we represent lyrics and chords in a shared vector space using a phrase-aligned chord-and-lyrics corpus. We show that models that use these shared representations predict a listener's emotion while hearing musical passages better than models that do not use these representations. Additionally, we conduct a visual analysis of these learnt shared vector representations and explain how they support existing theories in music. This work adds to our understanding of how lyrics and chords interact with one another in music and bears applications in music emotion recognition tasks and music information retrieval.

Index Terms— distributed representations, text classification, music emotion recognition

1. INTRODUCTION

Oftentimes, music is called a language of emotion. Can music bestow a certain emotional quality to words that words alone lack? Do we feel a different emotion when we hear an instrumental passage than when we hear that same passage with lyrics? Do songwriters couple lyrics and chords to elicit a particular mood in music listeners? We use techniques in natural language processing (NLP) to address these questions in this paper.

Studying chords and their patterns is useful in music emotion understanding [1, 2]. Studies show that consonant music activates different brain structures than dissonant music and indicate that music can affect our emotion [1]. Additionally, there exist several studies on emotional valence using NLP techniques [3]. Other studies apply NLP techniques to investigate if song lyrics can be predictive of music listeners' emotions [4]. It remains a topic of interest if information contained within a lyrical modality and other feature modalities, like chords, can be complementary for this prediction task [5].

Predicting how a music listener experiences emotion is a research interest in music information retrieval [6], psychology [7], and neuroscience [8]. Automatically identifying emotion in music can be used to tag music or provide insights into the mechanisms of human cognition [9]. However, this is a challenging task because of the subjective nature of emotion.

Several approaches have been taken to predict emotion during music listening. Researchers have used lyrics [10, 4] or multimodal approaches [11, 12] in music emotion recognition (MER). However, these studies have been limited to making predictions on emotions of people listening to a cappella [13] or instrumental music [14] or have otherwise not captured and used lyrical and chordal data in combination. Authors in [5] assert that the most accurate MER systems, like [15] and [16], apply large-scale machine learning algorithms to relatively short musical selections, using vast feature sets that span multiple domains.

Learning distributed word representations is a heavily-researched topic in NLP [17, 18, 19]. Recently, [20] applied the widely-used "word2vec" architecture to chord progressions. Other research has extended this architecture to a bilingual scenario [21, 19]. In this paper, we apply a bilingual approach to two "languages" in music: lyric sequences and chord progressions.

We hypothesize that learning shared representations—that is, embedding words from lyrics and chords in a shared vector space—capture how chord progressions and lyrics affect each other. We propose a novel emotion classification task to show the utility of these shared embeddings in predicting emotion. We also visualize these shared representations and explain how they support concepts from music theory.

2. RELATED WORK

A number of emotion classification tasks for music listening exist in the literature [14, 22, 23]. However, these tasks do not use datasets that have lyrics and chord information aligned together or otherwise use datasets with a limited number of songs with English lyrics. To the best of our knowledge, there is no existing dataset with chords and English lyrics side-by-side, so we curated our own dataset using online tablatures and used it on a pilot task related to MER (see Section 4 for details). Researchers have created visualisations of latent spaces [24]. In this paper, we extend this work to music, showing and analyzing visualisations of embedding spaces that use both chords and lyrics.

Learning word representations from text is a widely-studied topic in NLP in recent years [17, 18]. Recently, [20] and [25] have shown the utility of learning chord representations in predicting chord sequences. We adapt an architecture motivated from word2vec for creating bilingual word embeddings, similar to [21].

3. MODEL

We start this section by reviewing the standard skip-gram neural network architecture of Mikolov et al. [17]. Given a text corpus, skip-gram aims to induce word representations that are useful for predicting the context words surrounding a target word. The autoencoder maximizes the (monolingual) objective function:

$$MONO_W = \frac{1}{T}\sum_{t=1}^{T}\sum_{-l\leq j\leq l, j\neq 0} \log(p(w_{t+j}|w_t)) \qquad (1)$$

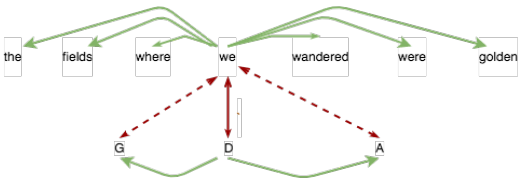

Fig. 1. Each word and chord predicts its word and chord context. This is similar to the architecture used by [21]

Fig. 2. Screenshot from UkuTabs showing a song excerpt. The "x2" indicates that the interlude section is repeated

where w_1, w_2, ..., w_T are words in the training corpus W and l is the size of the window around target word w_t, which is also from corpus W. Our proposed model aims to induce representations for lyrics and chords together. To this end, we implement a bilingual adaptation of the standard skip-gram, introduced by [21]. Specifically, this approach predicts the neighbors of given a chord c in a chord vocabulary if it is aligned with a word w in an English vocabulary and vice versa. Effectively, we train a single skip-gram model with a joint vocabulary on parallel corpora in which we enrich the training examples with pairs of words from both chords and lyrics instead of from lyrics or chords alone. As a result, this bilingual method learns embeddings for chords that are dependent on co-occurring lyrics and vice versa. The training objective function is $MONO_W + MONO_C + CROSS_{WC} + CROSS_{CW}$, where C and W are the corpora for chords and lyrics, respectively, and $CROSS_{WC}$ is defined as

$$CROSS_{WC} = \frac{1}{T_w}\sum_{t=1}^{T_w}\Big(\sum_{-l_c \le j \le l_c} \log(p(c_{k+j}|w_t))\Big) \quad (2)$$

and $CROSS_{CW}$ is defined as

$$CROSS_{CW} = \frac{1}{T_c}\sum_{t=1}^{T_c}\Big(\sum_{-l_w \le j \le l_w} \log(p(w_{k+j}|c_t))\Big) \quad (3)$$

In these cross-lingual objectives $CROSS_{CW}$ and $CROSS_{WC}$, the target index k is found by computing $[t * L_t/L_s]$ where L_t and L_s are the sentence lengths of the target language and source language, respectively. Figure 1 shows an example alignment of chords and lyrics.

We use stochastic gradient descent [26] with a learning rate of 0.01 and exponential decay of 0.98 after 10k steps (1 step = 256 word pairs), and negative sampling with 64 samples. A skip-gram window of size five is used for lyrics and a skip-gram window of size one is used for chords. Although there are more lyric tokens than chord tokens, we sample equal number of monolingual and cross-lingual word pairs to make a mini-batch at every step. The embedding space is 200-dimensional.

4. DATA

We curated a dataset from Ukutabs arrangements [27]. This website gives users direct access to an archive of over 5,500 popular songs from the 20th and 21st century. UkuTabs is sourced by users and systematically verified for quality by moderators. Each song is arranged in individual lines, with each line containing a matching chord and lyrical passage. Although other websites–such as ultimate-guitar.com, e-chords.com, and chordie.com–offer more songs, they are not verified for accuracy or do not have a standard format, making them unsuitable for automatically collecting high-quality data.

4.1. Data Collection

We retrieved the text data from every song in UkuTabs that was listed as a chord tablature [27]. For each musical passage that contained chords and lyrics (which we will call a "clip"), we lined up the chords with the lyrics.

We developed a chord caster, which converts all chords in the dataset into one of the four basic chord types: major, minor, dominant 7th, and diminished. This chord caster changed 15,020 of the 360,936 chords in the corpus (4.2%).

If a song's lyrics were less than 50% English words, the song was not included in the dataset. In addition, if a particular section of a song was repeated, the lyrics and chords were repeated in the dataset. See Figure 2 for an example of a repeated section.

To create useful representations of chords, it is necessary to find a chord's relation to a song's tonal center, or key. [28] uses hidden Markov models to estimate musical key for Beatles songs using chord symbols; however, they only use major and minor chords in their study and tested their model on only 110 songs from one artist in one genre. We developed a simple method to estimate the key of every song in our dataset. We created a 48-dimensional vector, with the count of the casted chords that are in a song as the entries of the vector. Then, for all twelve potential major keys, we tallied the number of chords that are in the scale of that key. The potential key with the highest such tally was selected as the estimated key. In the case of a tie, we summed the number of I, IV, V and vi chords of the tied keys and estimated the key to be the key with the highest sum.

Analysis of 50 random songs from the dataset revealed that this method for calculating the key of a song is effective: the method was 98% accurate on these songs. The key was estimated incorrectly for one song because that song contained a key change. This song was removed from the dataset.

Some statistics of the final dataset are given in Table 1.

5. EMOTION CLASSIFICATION TASK

Two annotators listened to music clips like the one represented in the verse in Figure 2 and labeled if they felt positive emotion, negative emotion, or neither. Both annotators are male musicians and songwriters, ages 18 and 26.

Table 1. Statistics of chords- and lyrics-aligned dataset

Total Sample Points	159,427
Number of Songs	4,417
Average Chords per Sample	2.8
Average Words per Sample	6.5
Number of Artists	1,611

5.1. Collecting Annotations

1,000 music clips were randomly chosen without replacement from the 206,315 clips and presented to two annotators. The annotators listened to the songs on YouTube from three seconds before the selected clip to three seconds after the selected clip. Past research has shown that this method gives enough time for emotion to stabilize [29]. The annotators noted if the music clip elicited positive emotion (happiness, love, excitement, elation, etc.), negative emotion (sadness, anger, heartbreak, wistfulness, etc.), or neither. We computed the inter-annotator agreement using a weighted Cohen's Kappa metric, which penalizes opposite polarity annotations [30]. Using this metric, we found $\kappa_w = .645$., which is consistent with another emotion detection task [31]. In the literature, κ values between .6 and 1.0 are considered "substantial" and "good" [32]

The annotators met after the initial annotation session to resolve discrepancies. Of the 308 samples that had disagreement, 237 were agreed-upon after discussion, giving a total of 929 samples with unanimous labels and $\kappa_f = .894$. The 71 other samples were discarded, so only samples with agreed-upon labels were used. This method of discrepancy resolution is used in other emotion tagging tasks like [33].

5.2. Results

We compare our systems to three baseline models. For our first baseline, we used a classifier that "chooses" the majority class (the neutral emotion class). For the other two baselines, we first created word 3-grams for chord progressions and lyrics separately, removing stopwords for the lyrics. Then, we trained a separate logistic regression classifier on the chord 3-grams and lyric 3-grams. Because the dataset contained English lyrics, a classifier like the one in [4] could not be used as a baseline.

We learned monolingual embeddings for chords and lyrics separately to use as two additional models using the monolingual word2vec architecture. To create clip-level features, we summed up the word or chord embeddings. Then, a one-vs.-rest logistic regression model without regularization was trained with 10-fold cross-validation on these features.

Table 2 shows the results for the emotion classification task. The *Chords Only* model and *Lyrics Only* model refer to a one-vs.-rest logistic regression model without regularization that uses embeddings learnt only using chord progressions and lyric sequences, respectively. The *Chords & Lyrics* model uses word embeddings learnt jointly using lyrics word sequences and chord progressions, as described in Section 3.

Using n-gram models did not significantly outperform choosing the majority class baseline. However, using only chord embeddings to predict the labels of this emotion classification task outperforms the majority class baseline ($p = .027$, using a one-tailed, two-sample t-test). Using embeddings performs significantly better than the majority class, but not significantly better than using n-gram models at $\alpha = .05$. The model which uses shared representations of chords and lyrics performs significantly better than the Majority Class model ($p = .001$) and the N-gram models ($p<.025$), but not significantly better than the model using only chord embeddings or only lyric embeddings ($p>.131$).

Table 2. Our model outperforms language models of chords or lyrics only, and the models that use embeddings outperform n-gram models. The p-values listed are based on a one-sided, two-sample t-test between each model and the majority class

Model	**Accuracy**	**p-value**
Baselines		
Majority Class	55.32 %	N/A
Chord n-grams	57.83 %	.138
Lyrics n-grams	57.31 %	.194
Our models		
Chords only	59.74 %	.027
Lyrics only	60.52 %	.012
Chords & Lyrics	62.28 %	.001

6. OBSERVATIONS

Many observations can be made from the embedding spaces that we create. We analyze the space that uses only chords (Figure 3) and the bilingual embedding space (Figure 4). We transposed every song into the key of C major and implemented a T-SNE Visualisation of the embeddings [24]. Diminished chords are not included in the visualisation to reduce clutter. The other three chord types are color-coded.

6.1. Submediants

In Figures 3 and 4, we observe that submediant chords are close in space to each other. The Dm chord is near the F chord and the Am chord is near the C chord, for example. Submediant chords share two of the same notes, so it is feasible that these chords may create a similar "feel" or affect.

6.2. Blues

Dominant chords are also clustered in space. C7, D7, F7, and G7 are near one another. This may be the result of the presence of bluesy songs in UkuTabs. Blues songs are generally built on repetition. These seventh chords may tend to contain similar lyrics to each other because blues artists tend to repeat lyrical lines when playing these chords.

6.3. "Unexpected chords"

Chords that are "outside" or non-enharmonic to the key are close in space in both T-SNE Visualisations. For example, Db, Bbm, and Ebm contain at least two notes that are not in the key of the songs, and these chords are close in space in Figures 4 and

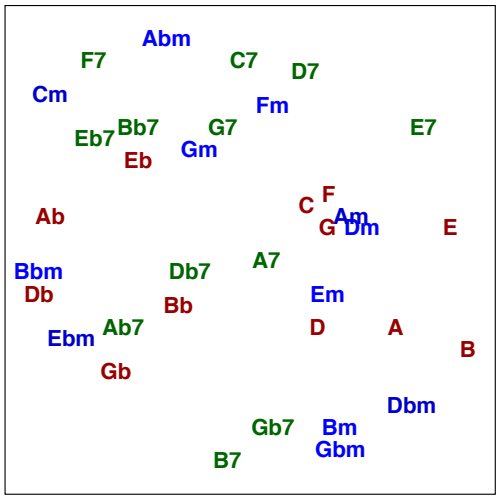

Fig. 3. Visualisation of the chord embedding space

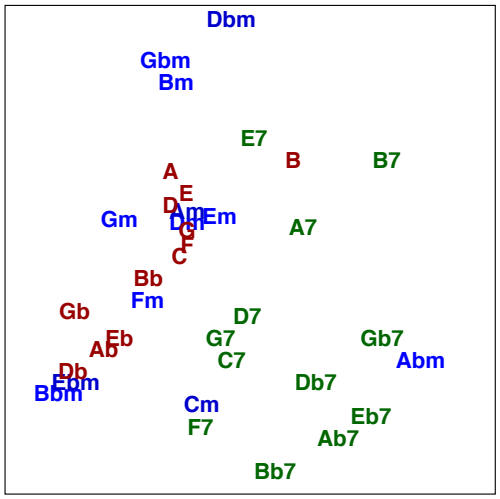

Fig. 4. Visualisation of the chords and lyrics embedding space

5. In both figures, enharmonic chords (C, F, G, Dm, Em, and Am) are also close in space. It may be that lyricists tend to write atypical lyrical lines when the music is drastically different from the tonal center of a song to reflect the chord's aberrance in the key. Music with "unexpected" chords tend to have lyrics that are similarly "unexpected."

7. DISCUSSION

The current dataset is limited by the coverage of UkuTabs' data, which has a bias towards pop music, but the method we present can be performed on any dataset that contains lyrics and chords in parallel. While our representations were useful for the particular task we formulated, we want to evaluate the utility of our representations on other tasks that already exist in the literature. Many tasks on emotion classification in music have been performed in the past; given that joint chordal and lyrical data can be accessed, it would be valuable to determine if our model outperforms the models proposed for other datasets. This way, we can know the generalization power of our embeddings. The chord caster we developed is untested and may be inaccurate for as much as 4.2% of the chords in the UkuTabs corpus. Using chord detection and speech-to-text algorithms available, our system can be used on any song for which the user contains the audio. Our key estimator performed well on the 50 songs that were tested. However, more investigation is necessary to determine if this simple estimator generalizes well. If it does, this estimation method may be a valuable, computationally-inexpensive way to estimate musical key.

Using a model that captures the temporal, dynamic behavior of chords and lyrics may be helpful in MER. A model like a recurrent neural network, for example, may outperform the logistic regression model that we present here.

8. CONCLUSIONS

We obtained a dataset that contains 159,427 musical segments from 4,477 pop songs, with lyrics and corresponding chords. Using this data, we developed a shared vector representation of the lyrics and chords together. We tested our representation on an emotion classification task by using a logistic regression model on the sum of the embeddings to predict listener emotion. We developed three models to predict the emotion experienced by the listeners on 929 musical clips: a model using only chord embeddings, a model using only lyric embeddings, and a model using joint chord-and-lyric embeddings. The models that use embeddings significantly outperformed the baseline models, but the chords-and-lyrics model did not significantly outperform the other models that use embeddings.

Visualising these representations, we found evidence of certain musical theories. Submediant substitutions are common in popular music, and seventh chords tend to be used in similar musical contexts. When lyrics and chords are embedded together, the seventh chords continue to be clustered together, and chords that are "outside" a key are clustered together in space. We can apply this work to many areas, including multimodal MER, music visualization, and music information retrieval.

9. REFERENCES

[1] S Koelsch, "Investigating emotion with music: neuroscientific approaches.," *Annals of the New York Academy of Sciences*, pp. 412–418, 2005.

[2] K J Pallesen, E Brattico, C Bailey, A Korvenoja, J Koivisto, A Gjedde, and Söve Carlson, "Emotion processing of major, minor, and dissonant chords," *Annals of the New York Academy of Sciences*, pp. 450–453, 2005.

[3] M M Bradley and P J Lang, "Affective norms for english words (anew): Instruction manual and affective ratings," 1999.

[4] Y Xia, L Wang, and KF Wong, "Sentiment vector space model for lyric-based song sentiment classification," *International Journal of Computer Processing Of Languages*, vol. 21, no. 04, pp. 309–330, 2008.

[5] Y E Kim, E M Schmidt, R Migneco, B G Morton, P Richardson, J Scott, J A Speck, and D Turnbull, "Music emotion recognition: A state of the art review," in *Proc. ISMIR*. Citeseer, 2010, pp. 255–266.

[6] Y Song, S Dixon, and M Pearce, "Evaluation of musical features for emotion classification.," in *ISMIR*. Citeseer, 2012, pp. 523–528.

[7] M Zentner, D Grandjean, and K R Scherer, "Emotions evoked by the sound of music: characterization, classification, and measurement.," *Emotion*, vol. 8, no. 4, pp. 494, 2008.

[8] E Brattico, V Alluri, B Bogert, T Jacobsen, N Vartiainen, S K Nieminen, and M Tervaniemi, "A functional MRI study of happy and sad emotions in music with and without lyrics," *Frontiers in psychology*, vol. 2, pp. 308, 2011.

[9] S Koelsch, T Fritz, K Müller, A D Friederici, et al., "Investigating emotion with music: an fMRI study," *Human brain mapping*, pp. 239–250, 2006.

[10] RH Chen, ZL Xu, ZX Zhang, and FZ Luo, "Content based music emotion analysis and recognition," in *Proc. of 2006 International Workshop on Computer Music and Audio Technology*, 2006, vol. 68275, p. 2.

[11] B Schuller, J Dorfner, and G Rigoll, "Determination of nonprototypical valence and arousal in popular music: features and performances," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, no. 1, pp. 735854, 2010.

[12] X Hu, J S Downie, and A F Ehmann, "Lyric text mining in music mood classification," *American music*, vol. 183, no. 5,049, pp. 2–209, 2009.

[13] E Parada-Cabaleiro, A Baird, A Batliner, N Cummins, S Hantke, and B W Schuller, "The perception of emotion in the singing voice: The understanding of music mood for music organisation," in *Proc. of the 4th International Workshop on Digital Libraries for Musicology*, 2017, pp. 29–36.

[14] J Fan, K Tatar, M Thorogood, and P Pasquier, "Ranking-based emotion recognition for experimental music," in *International Symposium on Music Information Retrieval*, 2017.

[15] D Turnbull, L Barrington, D Torres, and G Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 467–476, 2008.

[16] K Bischoff, C S Firan, R Paiu, W Nejdl, C Laurier, and M Sordo, "Music mood and theme classification-a hybrid approach.," in *ISMIR*, 2009, pp. 657–662.

[17] T Mikolov, I Sutskever, K Chen, G S Corrado, and J Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[18] J Pennington, R Socher, and C D Manning, "Glove: Global vectors for word representation.," in *EMNLP*, 2014, vol. 14, pp. 1532–1543.

[19] S Gouws, Y Bengio, and G Corrado, "Bilbowa: Fast bilingual distributed representations without word alignments," in *International Conference on Machine Learning*, 2015, pp. 748–756.

[20] S Madjiheurem, L Qu, and C Walder, "Chord2vec: Learning musical chord embeddings," 2016.

[21] T Luong, H Pham, and C D Manning, "Bilingual word representations with monolingual quality in mind," in *Proc. of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, 2015, pp. 151–159.

[22] K Trohidis, G Tsoumakas, G Kalliris, and I P Vlahavas, "Multi-label classification of music into emotions.," in *ISMIR*, 2008, pp. 325–330.

[23] B Han, S Rho, S Jun, and E Hwang, "Music emotion classification and context-based music recommendation," *Multimedia Tools and Applications*, vol. 47, no. 3, pp. 433–460, May 2010.

[24] L Maaten and G Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[25] C A Huang, D Duvenaud, and K Z Gajos, "Chordripple: Recommending chords to help novice composers go beyond the ordinary," in *Proc. of the 21st International Conference on Intelligent User Interfaces*, 2016, pp. 241–250.

[26] Herbert Robbins and Sutton Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.

[27] "Ukutabs," `http://ukutabs.com`, Accessed: 2017-10-27.

[28] K C Noland and M B Sandler, "Key estimation using a hidden markov model.," in *ISMIR*, 2006, pp. 121–126.

[29] Z Xiao, E Dellandréa, W Dou, and L Chen, "What is the Best Segment Duration for Music Mood Analysis ?," in *International Workshop on Content-Based Multimedia Indexing, CBMI 2008*, 2008, pp. 17–24.

[30] J Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit.," *Psychological bulletin*, vol. 70, no. 4, pp. 213, 1968.

[31] M Waterman, "Emotional responses to music: Implicit and explicit effects in listeners and performers," *Psychology of Music*, vol. 24, no. 1, pp. 53–67, 1996.

[32] J R Landis and G G Koch, "The measurement of observer agreement for categorical data," *Biometrics*, 1977.

[33] A Abbasi, H Chen, S Thoms, and T Fu, "Affect analysis of web forums and blogs using correlation ensembles," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 9, pp. 1168–1180, 2008.