SMM19, Workshop on Speech, Music and Mind 2019
14 September 2019, Vienna, Austria

Using Shared Vector Representations of Words and Chords in Music for Genre Classification

Timothy Greer[1], *Shrikanth Narayanan*[1,2]

[1]Department of Computer Science, [2]Department of Electrical Engineering
University of Southern California, Los Angeles, USA

timothdg@usc.edu, shri@sipi.usc.edu

Abstract

With so much music readily available for consumption today, it has never been more important to study music perception. In this paper, we represent lyrics and chords in a shared vector space using a phrase-aligned lyrics-and-chords corpus and show that models that use these shared representations can predict musical genre of songs—a perceptual construct of music listening—better than models that do not use these representations. This work adds to our understanding of how lyrics and chords interact with one another in music and has applications in multimodal perception and music information retrieval.

Index Terms: distributed representations, music perception, text classification, automatic genre classification

1. Introduction

Music is a complex, multifaceted, multimodal, perceptual experience that is challenging to analyze. Can a musical structure bestow a new quality to lyrics? How do we perceive music when we hear an instrumental passage and when we hear that same passage with lyrics? How would we categorize a song with lyrics related to pop music, but with chords steeped in the R&B tradition? We use techniques in natural language processing (NLP) to address these questions in this paper.

Many different ways to study music perception, and judgment, exist. Music emotion is one such lens through which to look at human perception of song. Another way to study music perception is through categorization: it is necessary to make a judgment (perception) about a stimulus (music) in order to classify the genre of a song. We use a music genre classification task as a way to investigate how humans perceive musical stimuli. One of the interesting questions is the cross-over and multi-attribute aspects of musical genre that do not necessarily fit clearly into one category. Among other things, this work has implications for how music content is marketed and consumed.

Studying chords and their patterns is useful in automatic genre classification [1, 2]. Additionally, there exist several studies on automatic genre classification using NLP techniques [3]. Other studies combine these NLP techniques with audio information to determine if a multimodal approach is helpful for genre prediction [4, 5]. It remains a topic of interest to determine if symbolic information contained within a lyrical modality and another modality can be complementary for genre classification [6].

Analyzing how a listener perceives music is a research interest in music information retrieval [7], psychology [8], and affective computing [9]. Automatically classifying musical genre can be used for music tagging or providing insights into the mechanisms of human cognition [10]. This is a challenging and interesting task because of the subjective nature of experience.

Learning distributed word representations is a heavily researched topic in NLP [11, 12, 13]. Recently, [14] applied the widely used "word2vec" architecture to chord progressions. Other research has extended this architecture to a bilingual scenario [15, 13]. 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 create a genre classification task using Billboard listings[1] to show the utility of these shared embeddings in predicting musical genre.

2. Related Work

A number of genre classification tasks exist in the literature [16, 17, 18]. However, these tasks do not use datasets that have lyrical and chordal information aligned together. In previous work, we curated our own dataset with chords and English lyrics side-by-side and used it on a pilot task related to music emotion recognition [19]. In this paper, we collect data from the same online source and use this data for a pilot task related to multi-label music genre classification (see Section 5 for details).

The subjective nature of music perception asks for a more objective, agreed-upon metric for genre classification. For this reason, we use the Billboard charts for musical genre labels. Billboard determines genre by "key fan interactions with music, including album sales and downloads, track downloads, radio airplay and touring as well as streaming and social interactions on Facebook, Twitter, Vevo, Youtube, Spotify and other popular online destinations for music" [20]. Social tags provided by users of online music streaming sites have been shown to be an effective method for classifying music based on their emotional content [21].

Learning word representations from text has been a widely-studied topic in NLP in recent years [11, 12]. [14] and [22] 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 [15].

We then use these embeddings in a multi-label classification task: classifying musical genre. Many techniques have been used for multi-label classification, including k-nearest neighbors (k-NN) classifiers [23], decision trees [24, 25], and neural networks [26]. In this paper, we focus on k-NN classifiers because of their simplicity, fast training times, and shown utility in multi-label classification tasks [27].

[1]https://www.billboard.com/charts

10.21437/SMM.2019-10

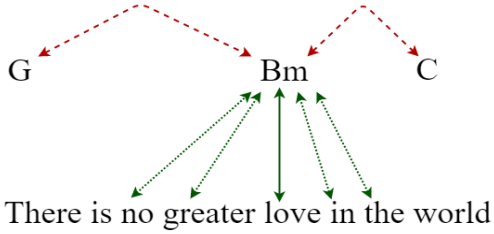

Figure 1: *Each lyric and chord predicts its lyric and chord context. Here, the B minor chord (Bm) predicts chords around it (dashed lines) and lyrics that are sung during and around the B minor chord (solid and dotted lines). The B minor chord is aligned with the lyric "love" because they are played and sung at the same time, respectively. This is similar to the architecture used by [15].*

3. Model

We begin this section by reviewing the standard skip-gram neural network architecture of Mikolov et al. [11]. Given a text corpus, a skip-gram model aims to induce word representations that are useful for predicting the context words surrounding a target word. The autoencoder maximizes the monolingual objective function:

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

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 two symbolic languages together: lyrics and chords. To this end, we implement a bilingual adaptation of the standard skip-gram, introduced by [15].

Specifically, this approach predicts the neighbors of a given chord c in a chord vocabulary C if it is aligned with a word w in a vocabulary W 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. $CROSS_{WC}$ is defined as

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

and $CROSS_{CW}$ is defined as

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

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 with lyrics.

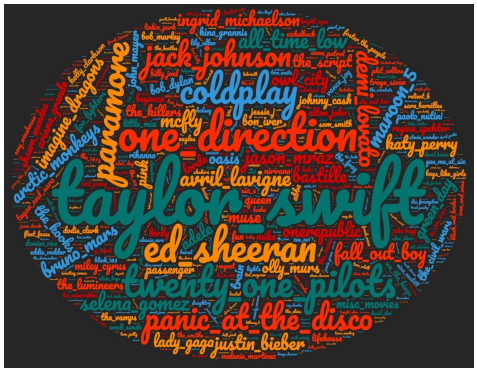

Figure 2: *A wordcloud of major artist names in our dataset.*

We use stochastic gradient descent [28] 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 resulting embedding space is 200-dimensional.

4. Data

We curated a dataset from Ukutabs arrangements [29]. This website gives users direct access to an archive of over 5,500 popular songs from the 20th and 21st centuries. The wordcloud in Figure 2 shows some of the most prominently featured artists on this website. 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 chordal 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 [29]. 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 17,602 of the 428,544 chords in the corpus (4.1%).

If a song's lyrics were less than 30% 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 3 for an example of a repeated section. Some statistics of the final dataset are given in Table 1.

To create useful representations of chords, it is necessary to find a chord's relation to a song's tonal center, or key. [30] 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, identi-

Table 1: *Statistics of chords- and lyrics-aligned dataset*

Total Sample Points	190,165
Number of Songs	5,304
Average Chords per Sample	2.2
Average Words per Sample	8.0
Number of Artists	1,770

Figure 3: *Screenshot from UkuTabs showing a song excerpt. The "x2" indicates that the interlude section is repeated.*

cal to [19]. 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[2].

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 final dataset.

5. Genre Classification Task

After learning representations for these musical passages, we created a task that involved classifying musical genre: a proxy for music perception. After all, categorization and perception are thought to be strongly linked [31]. We used the Billboard charts to provide ground truth for a song's genre.

5.1. Collecting Billboard Songs

We collected the song titles listed on the Latin, Rock, R&B/Hip-Hop, and Pop Billboard charts from the last 20 years [32]. These songs were then matched to songs from the dataset collected from Ukutabs by song title[3]. Table 2 lists the number of songs from each chart found in the dataset. There were 850 unique songs found.

We computed the embeddings for each of these songs by summing the embeddings of every token in each song and dividing by the number of tokens in that song. This 200-dimensional embedding was used as a feature vector for prediction.

[2]"I", "IV", "V", and "vi" refer to the Roman numeral notation of four common chords in diatonic music. In the key of C major, this refers to C major, F major, G major, and A minor, respectively.

[3]As only song titles were used for matching in this study, many songs that were found in Ukutabs were covers of their "matching" Billboard songs. These songs were included because cover songs usually have similar or identical chords and lyrics to those of the original song. Matching songs only if they shared the same song title *and* artist name produced similar results to those shown in this paper.

Table 2: *Number of songs in each Billboard chart in the last 20 years also found in the Ukutabs dataset. The diagonal entries refer to the number of songs that contain a label in a particular genre. The number of "crossover" songs, or songs listed in more than one chart, are entries in the off-diagonals. RnB/HH stands for R&B/Hip-Hop.*

	Latin	Country	Pop	Rock	RnB/HH
Latin	79	7	67	12	21
Country	7	342	101	64	61
Pop	67	101	780	198	190
Rock	12	64	198	620	88
RnB/HH	21	61	190	88	344

Table 3: *A list of models used for multi-label genre classification and their performance in three metrics. The* Chords & Lyrics *model performs second-best in the harsh Exact Match Ratio (EMR), but has better label accuracy (Accuracy), and label-based, micro-averaged f1-score (f1-score) than baseline models and models that use only chords or only lyrics.*

Model	**EMR**	**Accuracy**	**f1-score**
Baselines			
Most Common Set	**23.9 %**	.350	.413
Bag of Words	16.8 %	.298	.392
Our models			
Chords Only	16.5 %	.350	.370
Lyrics Only	18.5 %	.362	.402
Chords & Lyrics	20.4 %	**.379**	**.428**

5.2. Results

We compared our systems to two baseline models. For our first baseline, we used a classifier that "chooses" the most common labelset. This classifier predicted that every song belonged to only the pop genre. We created a Bag of Words classifier, treating chords and lyrics as one language. While this baseline uses both lyrical and chordal modes, it does not have a notion of chord progressions, lyric sequences, or chordal and lyrical interaction.

We learned monolingual embeddings for chords and lyrics using the monolingual word2vec architecture to use as two additional models. To create song-level features, we averaged the embeddings. A k-NN classifier was trained on these features, after reducing the dimensionality using principal component analysis [33]. We empirically set the post-PCA dimensionality to be three for all classifiers (this captured >90% of the variability in the data) and used five-fold cross-validation in all experiments.

Table 3 shows the results for the emotion classification task. The *Chords Only* model and *Lyrics Only* model refer to a k-NN model 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.

The Exact Matching Ratio (EMR) metric is the number of test examples that have labelsets that exactly match the predicted labelsets, divided by the number of test examples:

$$EMR = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(Y_i = Z_i) \quad (4)$$

where $\mathbb{1}(\cdot)$ is the indicator function, Y_i is the true labelset, and

Z_i is the predicted label.

The Label Accuracy metric rewards correctly predicted labels and penalizes incorrectly predicted labels. Concretely:

$$H = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|} \quad (5)$$

Label-based, micro-averaged f1-score involves aggregating the contributions of all classes to determine precision and recall measures and computing the f1-score from these aggregated measures [34]. The stringent EMR metric favored the *Most Common Set* classifier. However, in another common metric for multi-label classification (label accuracy), the *Lyrics only* model outperformed the baseline models, and the chords-and-lyrics embeddings model outperformed all other models in both label accuracy and f1-score. That the embeddings-based models performed well according to the label accuracy metric suggests that these models are predicting correct labels for songs even if the predicted labelset does not perfectly match the actual labelset. The *Chords & Lyrics* model's outperfoming other models in label accuracy and f1-score shows the utility of using a multimodal approach to study music perception.

6. Discussion

The current dataset is limited by the coverage of UkuTabs' data, which has a bias towards music that is playable by ukelele musicians, but the method we present can be performed on any dataset that contains accurate lyrics and chords in parallel. While our representations are useful for genre classification, we want to evaluate the performance of our models with respect to models that use auditory features. Fusing audio information (such as rhythm and harmony) with our symbolic representations of these songs may result in a model that can better classify musical genre, providing a deeper understanding of how we perceive music. If audio features do not contribute to a model, it suggests that music perception may be better modeled using symbolic representations of songs[4]. State-of-the-art models for music genre classification, such as RAkEL models, may also demonstrate better performance than the k-NN classifiers used in this study [27].

The chord caster we developed is untested and may be inaccurate for as much as 4.1% of the chords in the UkuTabs corpus. Using chord detection algorithms (like those mentioned in [35]) and state-of-the-art speech-to-text algorithms, our system could be used on any song for which the user contains the audio.

While our objective function for creating embeddings placed equal emphasis on all terms, different weight coefficients may be used to emphasize the mono-lingual lyrical or chordal terms or the cross-lingual components of the objective function. Choosing to do this would change the resultant features used in the genre classification task presented. To avoid using embeddings that were overly influenced by any of the four terms in the objective function, we opted to set the weight coefficients for each term to 1. Further study is necessary to determine if emphasizing certain terms in the objective function results in shared representations that are better-suited for musical genre prediction.

Discarding songs that contained 30% or less words in English did not have a great effect on the performance of the classification models. In fact, only one song listed in the Billboard Latin charts and the Ukutabs dataset contained Spanish lyrics ("Feliz Navidad" by José Feliciano.) 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.

[4] Our representations can be trained much faster than computing auditory features for these songs: our model took three minutes to train, whereas a model trained using auditory features would likely take orders of magnitude longer.

7. Conclusions

We obtained a dataset that contains 190,165 musical segments from 5,304 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 a genre classification task by using a k-NN classifier on the average of the embeddings to predict genre labels given by the Billboard charts. We developed three models to predict genre: a model using only chord embeddings, a model using only lyric embeddings, and a model using joint chord-and-lyric embeddings. The model that uses joint embeddings significantly outperformed the baseline models and monolingual embedding models in three multi-label classification metrics, demonstrating the utility of taking a multimodal approach to music perception. We can apply this work to many areas, including multimodal human perception, automatic genre classification, and music information retrieval.

8. Acknowledgements

The support for USC SAIL from NSF and NIH, and from Google for its Center for Computational Media Intelligence, is gratefully acknowledged. We would also like to thank Professor Morteza Dehghani at the University of Southern California for encouraging this work and Benjamin Ma for his insights and help in data collection.

9. References

[1] H.-T. Cheng, Y.-H. Yang, Y.-C. Lin, I.-B. Liao, and H. H. Chen, "Automatic chord recognition for music classification and retrieval," in *2008 IEEE International Conference on Multimedia and Expo*. IEEE, 2008, pp. 1505–1508.

[2] A. Anglade, R. Ramirez, S. Dixon *et al.*, "Genre classification using harmony rules induced from automatic chord transcriptions." in *ISMIR*, 2009, pp. 669–674.

[3] R. Mayer, R. Neumayer, and A. Rauber, "Rhyme and style features for musical genre classification by song lyrics." in *Ismir*, 2008, pp. 337–342.

[4] ——, "Combination of audio and lyrics features for genre classification in digital audio collections," in *Proceedings of the 16th ACM international conference on Multimedia*. ACM, 2008, pp. 159–168.

[5] C. Laurier, J. Grivolla, and P. Herrera, "Multimodal music mood classification using audio and lyrics," in *2008 Seventh International Conference on Machine Learning and Applications*. IEEE, 2008, pp. 688–693.

[6] C. McKay, J. A. Burgoyne, J. Hockman, J. B. Smith, G. Vigliensoni, and I. Fujinaga, "Evaluating the genre classification performance of lyrical features relative to audio, symbolic and cultural features." in *ISMIR*, 2010, pp. 213–218.

[7] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. P. Vlahavas, "Multi-label classification of music into emotions." in *ISMIR*, vol. 8, 2008, pp. 325–330.

[8] K. Siedenburg, I. Fujinaga, and S. McAdams, "A comparison of approaches to timbre descriptors in music information retrieval

and music psychology," *Journal of New Music Research*, vol. 45, no. 1, pp. 27–41, 2016.

[9] A. Schindler and A. Rauber, "An audio-visual approach to music genre classification through affective color features," in *European Conference on Information Retrieval*. Springer, 2015, pp. 61–67.

[10] 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.

[11] 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.

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

[13] 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.

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

[15] 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.

[16] T. Li, M. Ogihara, and Q. Li, "A comparative study on content-based music genre classification," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003, pp. 282–289.

[17] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on speech and audio processing*, vol. 10, no. 5, pp. 293–302, 2002.

[18] N. Scaringella, G. Zoia, and D. Mlynek, "Automatic genre classification of music content: a survey," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 133–141, 2006.

[19] T. Greer, K. Singla, B. Ma, and S. Narayanan, "Learning shared vector representations of lyrics and chords in music," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3951–3955.

[20] "Billboard-legend," https://www.billboard.com/biz/billboard-charts-legend, accessed: 2019-05-27.

[21] Y. Song, S. Dixon, and M. Pearce, "A survey of music recommendation systems and future perspectives," in *9th International Symposium on Computer Music Modeling and Retrieval*, vol. 4, 2012.

[22] 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.

[23] M.-L. Zhang and Z.-H. Zhou, "Ml-knn: A lazy learning approach to multi-label learning," *Pattern recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.

[24] H. Blockeel, L. Schietgat, J. Struyf, S. Džeroski, and A. Clare, "Decision trees for hierarchical multilabel classification: A case study in functional genomics," in *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, 2006, pp. 18–29.

[25] W. Yi, M. Lu, and Z. Liu, "Multi-valued attribute and multi-labeled data decision tree algorithm," *International Journal of Machine Learning and Cybernetics*, vol. 2, no. 2, pp. 67–74, 2011.

[26] M.-L. Zhang and Z.-H. Zhou, "Multilabel neural networks with applications to functional genomics and text categorization," *IEEE transactions on Knowledge and Data Engineering*, vol. 18, no. 10, pp. 1338–1351, 2006.

[27] C. Sanden and J. Z. Zhang, "Enhancing multi-label music genre classification through ensemble techniques," in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 2011, pp. 705–714.

[28] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.

[29] "Ukutabs," http://ukutabs.com, accessed: 2019-05-27.

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

[31] P. Kay and W. Kempton, "What is the sapir-whorf hypothesis?" *American anthropologist*, vol. 86, no. 1, pp. 65–79, 1984.

[32] "Billboard," https://www.billboard.com/charts, accessed: 2019-05-27.

[33] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[34] M. S. Sorower, "A literature survey on algorithms for multi-label learning."

[35] J. Pauwels and G. Peeters, "Evaluating automatically estimated chord sequences," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 749–753.