Towards a music-language mapping

Michele Berlingerio, Francesca Bonin
IBM Research Ireland
{mberling, fbonin}[at] ie.ibm.com

Abstract

We explore a novel research idea, that we call Musical Language Processing (MLP), which investigates the possibility of a musical input to speech interaction systems. We present the first attempts at finding a mapping between musical pieces and dialogues, based on the frequency of musical patterns. Our findings on one possible alignment between classical piano compositions and dialogues from popular TV series are encouraging, and open the way to further investigations along this line of research.

Keywords: Language generation, lexical analysis, data-driven methods

1. Introduction

Natural Language Processing applications are becoming more and more pervasive in our every day lives. Dialogue systems like Amazon Echo, Google Home, Microsoft Cortana, or Apple Siri, are assisting us in many tasks like searches, purchases, simple calculations, etc. As of today, all these technologies are language-dependent, must be re-trained for each different language and they imply the knowledge of the input language for any form of communication. In this work, we investigate the possibility of musical input to dialogue systems. While we appreciate that musical input certainly requires the knowledge of the basis of musical theory for allowing communication, it presents several advantages: *i)* it is universal among different languages; *ii)* it would facilitate the communication with dialogue system for persons with linguistics disabilities, that have been proven to have particular musical skills (Heaton et al., 1998; Mottron et al., 2006; Happé, 1999); *iii)* it would be simple to learn for a large class of users, i.e. musicians. We call this line of research Musical Language Processing (MLP).

The fist step is to find a mapping between musical pieces and dialogues, in order to investigate the possibility of an alignment. To this aim, in this preliminary work, we study one possible mapping between classical piano compositions and dialogues from popular TV series, by investigating the frequency distributions of words in dialogues and chords in music. From this analysis we aim at creating a lexical mapping between chords and the English vocabulary, which would allow, at a later step, to investigate possible syntactical mappings (n-chords and n-grams), and semantic mappings, where the musical language will be able to express simple meaningful sentences.

Note that, instead of trying to map a chord to a word, and then forcing users to learn the mapping and play the right sequence of chords to form a sentence, our final goal is to "reverse-engineer" already existing music, and find a mapping that any person with "some" knowledge of music would be able to reproduce, without having to learn a new language, or a new kind of music. This is a what-if analysis of what would happen if, for example, we would map existing Bach's compositions, chord by chord: would they form a meaningful sentence?

While there are works involving music and Natural Language Processing, to the best of our knowledge this is the first attempt at mapping a musical language to a spoken one.

The paper is structured as follows: in Section 2. we present previous works; in Section 3. we describe the methodology; Section 4. describes the datasets used; preliminary experiments are presented in Section 5. and conclusions are drawn in Section 6.

2. Related work

Previous works have been focussing on the correlation between music and language. (Longfellow, 1835) defines music as the universal language of humanity, showing how music has always been considered a means of communication, sharing many characteristics with languages.

Approaches to apply NLP techniques to music have been conducted by (Bod and others, 2001). The authors apply syntactic parsing to musical compositions noticing how ambiguities is a common problem, hence an interesting similarity between music and language.

From a prosodic point of view, (Patel and Daniele, 2003) showed the relations between rhythm in language and music.

Some works have been conducted on the relation between music and emotions, in sentiment analysis, (Mihalcea and Strapparava, 2012; Strapparava et al., 2012), exploiting the music and the lyrics of songs.

(Davis and Mohammad, 2014) created a system to generate music from text, using a mechanism to determine sequences of notes that capture the emotional activity in text.

In light of the similarities emerged between music and language in previous works, in this paper we investigate the possibility to map musical chords and words based on their frequency. Differently from (Davis and Mohammad, 2014), we do not create music from text, but we explore the first steps needed to produce the best possible mapping between English lexicon and chords of six classical composers.

3. Methodology

The purpose of this paper is to show a what-if analysis of what would happen if we were to try to translate music into spoken language, in such a way that it should be possible to control an ad-hoc smart personal assistant.

A simple interpretation of this would be to train a musician to play "words" instead of notes, or chords. Our approach is the opposite, and takes advantage of the large availability of music on the Internet: we start from existing music, and reverse-engineer a possible music-to-language mapping, by means of a data-driven approach.

We looked at a large set of musical pieces, trying to map notes, or chords, to words, and find a *good* mapping, so that any musician would have to invest little effort in adapting to the mapping, starting from the music (s)he already knows. However, defining a good mapping is not an easy task. We should try to align music and language not only at the lexicon level, but also the syntactical and the semantical one. In this paper, we make the first attempt in finding such a mapping directly from available music, starting from the lexical level. In future work, we will refine our methodology and focus also on other linguistics levels.

To focus on the lexicon, one of the first steps is exploring possible manners to compose a word out of music. There are several options: we could map a note to a phoneme, a note to a word, a sequence of notes to a word, etc. The first solution presents a problem: in music, several notes can be played at the same time (which is the definition of a *chord*), while, in linguistics, phonemes are uttered in sequence. The second solution, mapping notes to words, besides suffering from the problem above, would also dramatically limit our vocabulary: a piano keyboard has 88 keys, while there are hundreds of thousands words in modern English.

To solve this issue, one could consider also the chords, to broaden the space. If we allow for any combination of the 88 keys of a piano keyboard to correspond to a word, this would give us $2^{88} = 2.09 * 10^{26}$ possibilities, which is more than enough for any given language in the world. However, not all these combinations are used by composers, as well as not all the words of a vocabulary are used in a given conversation. Actually, in both music and language, we can compute the frequency of appearance of a particular combination of notes or words. From there, we can also compute n-grams (i.e. sequences of chords), chords co-occurence, and evolve into a syntactical and semantic mapping.

Moreover, we could add punctuation and map musical *phrases* to linguistic sentences: John White defines a musical phrase as "the smallest musical unit that conveys a more or less complete musical thought. Phrases vary in length and are terminated at a point of full or partial repose, which is called a cadence." (White, 1976).

In Section 5. we show the results of applying this kind of mapping, hereafter *chord-word* mapping, to six music datasets presented in Section 4.. Our first step is to compare high level statistics of the musical datasets with 6 textual dataset (described in 4.): number of different chord-words used by a composer (or, in a musical piece) vs number of different words used in a given TV series, frequency distribution of the resulting chord-words (unigrams) as well as bi and trigrams, and chord-words per minute vs words per minute in a TV series. The aim is to assess if even basic high-level properties of the spoken language, such as the Zipf's law of the word frequencies, are properties of our mapping as well.

At a high level, given a music file, our procedure is as follows:

1. we scan the music score and keep track of the notes being played at the same time: every time these change, we record a "chord" (even if this is composed by less than three notes, for simplicity)
2. we assign an id to each resulting chord
3. we repeat the procedure for each piece from the same composer, forming the composer's vocabulary
4. we sort the vocabulary by descending frequency by composer, and map it to different words according to the descending frequency in different samples of spoken language.

Data	# pieces	Avg # of notes per piece
Bach	153	2747.82
Beethoven	17	8311.53
Chopin	49	4388.08
Grieg	17	2812.88
Schubert	30	9482.50
Schumann	25	2955.64

Table 1: Basic statistics for the music datasets

Data	# episodes	Avg # of words per ep.
BB	58	3033.71
GOT	51	3786.96
HIMYM	188	2635.03
HOC	41	5037.59
MF	170	3248.56
S	77	5902.74

Table 2: Basic statistics for the TV series datasets

4. Datasets

We used two kinds of data: MIDI[1] music files of classical piano compositions, and TV series subtitles for dialogues. In particular, we used compositions from: Bach[2], Beethoven, Chopin, Grieg, Schubert, and Schumann[3]. As TV series, we used the subtitles from all available episodes of six popular shows: Breaking Bad (BB), Game of Thrones (GOT), How I Met Your Mother (HIMYM), House of Cards (HOC), Modern Family (MF), and Suits (S)[4]. The aim of this broad choice was to try to differentiate the styles, both of the musical pieces and of the dialogues, by varying composers (from different eras), and topics and setting for the TV series. We acknowledge that, in future work, an even broader and more elaborate choice should be made to reduce potential bias. Moreover, standard datasets should be added for spoken dialogues, to be able to assess the results against well known properties of those datasets.

Tables 3 and 4 show some basic statistics of our datasets: for each musical composer, we report the number of pieces

[1] https://www.midi.org/specifications

[2] http://www.bachcentral.com/midiindexcomplete.html

[3] http://www.piano-midi.de

[4] http://www.tvsubtitles.net

Data	# distinct chord-words	Avg distinct c.-w./piece
Bach	16727	109.33
Beethoven	15218	895.18
Chopin	18388	375.27
Grieg	3708	218.12
Schubert	19945	664.83
Schumann	8160	326.40

Table 3: Distinct chord-words - music datasets

Data	# distinct words	Avg distinct w./ep.
BB	10325	178.02
GOT	8381	164.33
HIMYM	21461	114.15
HOC	10802	263.46
MF	20776	122.21
S	15257	198.14

Table 4: Distinct words - TV series datasets

in our collection, and the average number of notes per piece; for each TV series, we report the number of episodes in our collection and the average number of words per episode.

5. Experiments

We implemented our chord-word mapping in Python, using the mido library for handling MIDI files[5].

Tables 3 and 4 show the number of distinct chord-words and words in different composers or TV series, respectively, as well as the average number of them by piece or episode. Note that we did not apply any kind of normalization, while we are aware that episodes from different series have different lenghts, and this is true also within the pieces of a given composer. However, we think of a single piece or episode as a story *per se*, therefore we are interested in studying the language used to tell *one* story. Different normalizations are possible here (for example, number of words or chord-words per minute), and they will be studied in future work.

Figures 1 and 2 show the distributions of frequency of the n-grams (n from 1 to 3) in three different TV series and three different composers, respectively.

We see that both the distributions of the n-grams in the analysed compositions and TV series follow a Zipf's Law (Powers, 1998). Moreover, we notice that the change of α exponent between unigrams, bigrams and trigrams in the composers is similar to the one in the TV series. This interesting parallelism poses the bases for a possible lexical mapping based on chords frequency.

Note that we had datasets of different lengths for different composers, therefore the data for Grieg and Schumann appear much sparser. However, being Schumann a composer from practically the same musical era of all the other ones, his distribution does not differ from the others. Grieg, instead, was from a different era, the romanticism, and his language is indeed different. This may suggest that composers of even more recent eras, like the jazz one, may present much different languages. Jazz, for example, is well known to follow non-repetitive, improvised, schemas, by continuously introducing and relieving musical *tension*. This is obtained by playing notes and chords that are far from the current harmony, or by delaying or anticipating notes. It would be interesting to compare jazz to particular types of dialogues, such as quarrels, or jokes.

[5] https://github.com/olemb/mido

Rank	Unigram	Bigram	Trigram
1	i	you know	i don't know
2	you	i don't	you want to
3	the	in the	oh my god
4	to	this is	what are you
5	a	and i	what do you

Table 5: Most frequent n-grams in HIMYM

Rank	Unigram	Bigram	Trigram
1	rest	A#4 rest	G5-A#5-C#6 F#5-A5-A#5-D6 F#5-A5-D6
2	A#4	G4 rest	D#4-B4 D4-D#4-G#4-A#4-B4 D4-G#4-A#4
3	C6	C5 rest	D#4-G#4-B4 D#4-B4 D4-D#4-G#4-A#4-B4
4	F6	F5 rest	G5-A5-A#5-C#6 G5-A#5-C#6 F#5-A5-A#5-D6
5	D#6	D#5 rest	D4-D#4-G#4-A#4-B4 D#4-G#4-B4 D#4-B4

Table 6: Most frequent n-grams in Beethoven

Tables 5 and 6 show the five most frequent unigrams, bigrams and trigrams, for HIMYM and Beethoven, respectively (due to lack of space, we are not reporting this information for the others). As our aim was to map the entire language, we did not perform any filtering of stop words, nor applied lemmatization. It is not a surprise, then, that the top words in the TV series are usually considered stop words. It is interesting to note that their musical counterpart could be considered *musical stop words* as well: the most frequent chords are a pause, and then single notes in the central area of the human voice. Note that, due to the way we processed MIDI files and subtitles, we do have pauses in the music datasets, but we did not process the silences in the TV series. In the latter, in fact, they are also affected by non-dialogue scenes, while pauses in music are usually brief. Nevertheless, this result in music confirms composer Stravinski's view of the musical language, when he said that music is composed mainly of silence (Cantoni, 2014).

6. Conclusions and future work

We have reported the preliminary investigation of a music-language mapping, to pose the basis for a new language to be used in speech interaction systems. We have shown interesting parallelisms between classical music and the lexicon of dialogues in TV series. We intend to proceed with the analysis between music and the syntactic level of the spoken language as well as investigating different musical and linguistic genres. Finally we will be looking at the semantic level, to understand how to ensure that a so constructed musical piece would *make sense* in the common language.

7. Bibliographical References

Bod, R. et al. (2001). Probabilistic grammars for music. In *Belgian-Dutch Conference on Artificial Intelligence (BNAIC)*. Citeseer.

Cantoni, A. (2014). *The Language of Stravinsky*. Musikwissenschaftliche Publikationen. Olms, Georg.

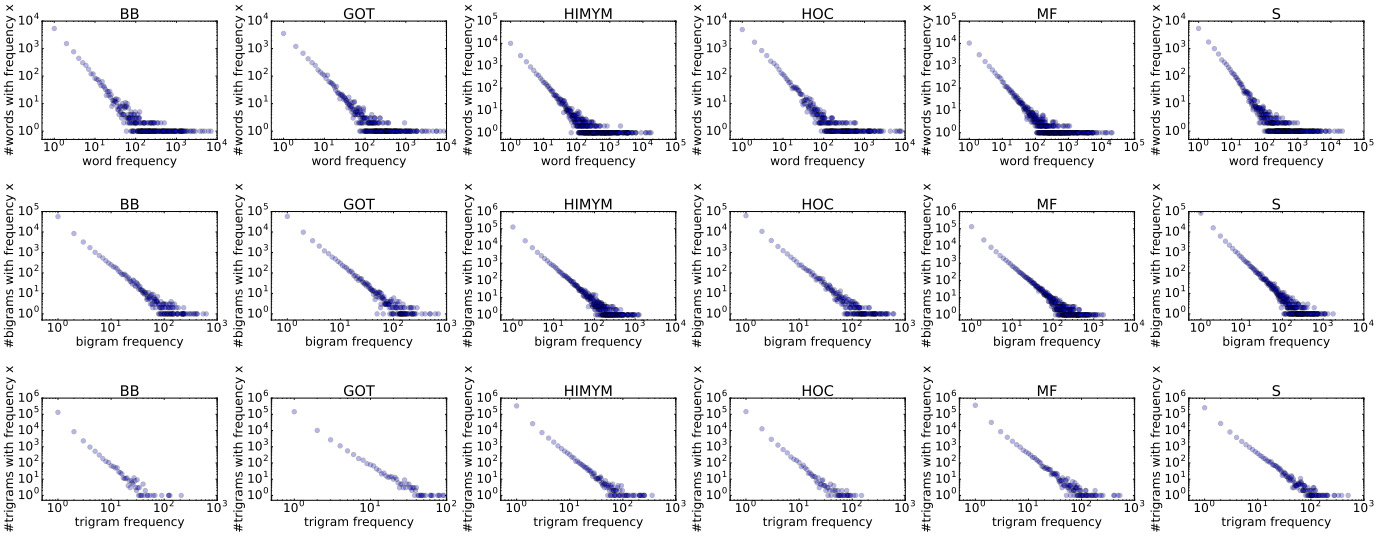

Figure 1: N-grams frequencies for the TV series

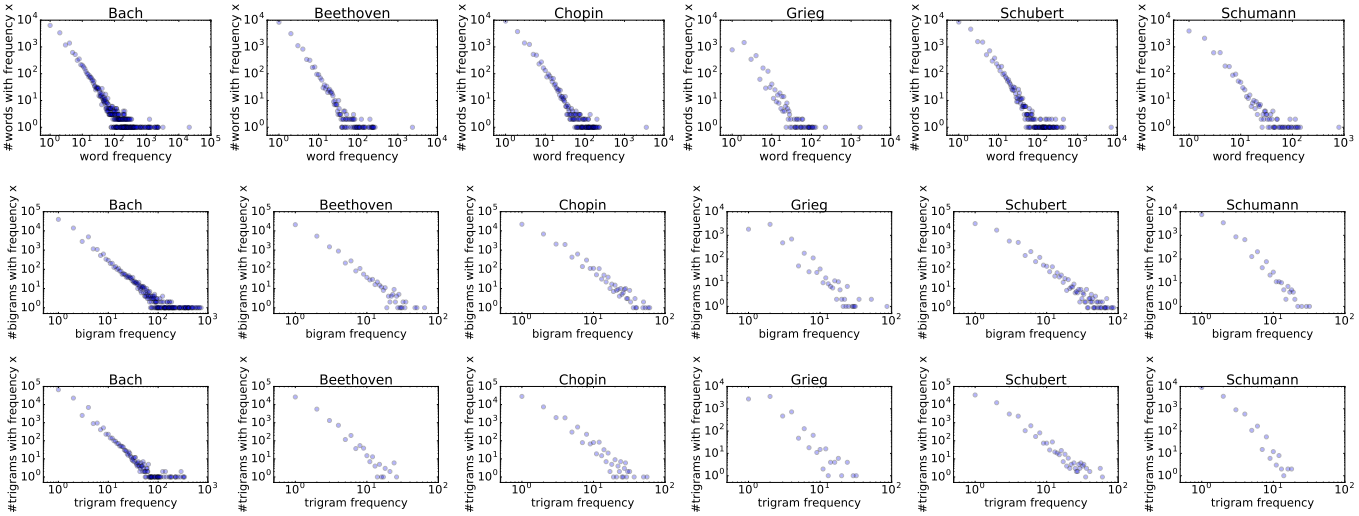

Figure 2: N-grams frequencies for the composers

Davis, H. and Mohammad, S., (2014). *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, chapter Generating Music from Literature, pages 1–10. ACL.

Happé, F. (1999). Autism: cognitive deficit or cognitive style? *Trends in cognitive sciences*, 3(6):216–222.

Heaton, P., Hermelin, B., and Pring, L. (1998). Autism and pitch processing: A precursor for savant musical ability? *Music Perception: An Interdisciplinary Journal*, 15(3):291–305.

Longfellow, H. (1835). *Outre-mer: A Pilgrimage Beyond the Sea*. Number v. 1-2 in Outre-mer: A Pilgrimage Beyond the Sea. Harper.

Mihalcea, R. and Strapparava, C. (2012). Lyrics, music, and emotions. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 590–599. ACL.

Mottron, L., Dawson, M., Soulieres, I., Hubert, B., and Burack, J. (2006). Enhanced perceptual functioning in autism: An update, and eight principles of autistic perception. *Journal of autism and developmental disorders*, 36(1):27–43.

Patel, A. D. and Daniele, J. R. (2003). An empirical comparison of rhythm in language and music. *Cognition*, 87(1):B35 – B45.

Powers, D. M. W. (1998). Applications and explanations of zipf's law. In *Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning*, NeMLaP3/CoNLL '98, pages 151–160, Stroudsburg, PA, USA. ACL.

Strapparava, C., Mihalcea, R., and Battocchi, A. (2012). A parallel corpus of music and lyrics annotated with emotions. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. European Language Resources Association (ELRA).

White, J. D. (1976). *The analysis of music*. Englewood Cliffs, N.J. : Prentice-Hall. Includes bibliographical references and index.