Query-by-Multiple-Examples: Content-Based Search in Computer-Assisted Sound-Based Musical Composition

Tiago Fernandes Tavares, Jônatas Manzolli
Interdisciplinary Nucleus of Sound Communication
Music Department – Institute of Arts
University of Campinas
Campinas - SP - Brazil
tiago@nics.unicamp.br

ABSTRACT

We propose a search method, namely Query-by-Multiple-Examples, that is able to search, within an audio sample database, for a particular sonic characteristic. The characteristic is learned on-the-fly by means of multiple examples provided by a human user, thus avoiding ambiguities due to manual labelling. We evaluate four variations of the proposed method using ground truth provided by three musicians. It is shown that, for queries based on sonic characteristics, the query modelling process yields more correct results than if several single-example queries were executed in parallel using the same input data.

1. INTRODUCTION

Sound-based music is that in which the main discourse is based on the evolution of sonic characteristics [1]. This category comprises genres such as Electroacoustic, some kinds of Electronics, Acousmatic, and Mixed Music. A frequent part of the composition process in these genres is recording sound samples from diverse sources and using them as material – either raw [2], processed [3] or as inspiration [4] – for the construction of a piece.

Although it is common that a musician has a personal, well-known sample database, the process of recording new samples may be time-demanding. Collaborative databases allow a composer to benefit from its peer's recording work, providing quick access to many more sounds than it would be feasible to personally record. We propose a search method that allows semi-automatic search using personalized criteria, allowing composers to find new, interesting sounds in sample databases that are too big for careful listening.

Many content-based search methods rely on semantic tags [5, 6, 7, 8, 9, 10] or other contextual data, like user ratings or popularity [11], but these approaches are of limited use for composers as they are often interested in sonic characteristics of an audio sample, not the identification or perceived quality of the recorded object. It is important to note that sonic characteristics are often multi-dimensional, and composers are often interested in nuances [12], such as "noisiness" or "brightness", which may assume different meanings depending on the context [13]. Therefore, we propose a data-driven system as a solution for this problem.

In a data-driven search system [14, 15], sound samples are mapped into a $\mathbb{R}^N$ vector space defined by low-level features calculated from audio samples. These features aim at encoding the multiple dimensions related to sound perception, which means perceptually similar sounds are likely to be closer to each other [16] according to some distance measure [17]. However, it is impossible to know, from a single element, what perceptual characteristics are desired by the composer and what are not important; hence, the search system requires, as input, two or more examples so that it is possible to know which dimensions should be considered or disregarded in the search process.

We propose a novel search method, namely *Query-by-Multiple-Examples*, in which multiple examples are used to train a search machine regarding what perceptual characteristics are desired by the user and what other characteristics may be disregarded. This aims at providing a high level of customization in the search criteria. Thus, the composer's perception is quickly modelled and extended, allowing the retrieval of sound samples in a big database according to personal criteria.

This paper is organized as follows. The implemented search methods are described in Section 2. The evaluation method, as well as the results, are shown in Section 3. Section 4 brings further discussion and Section 5 concludes the text.

2. PROPOSED METHOD

The proposed system is built as to merge two different sources of information, as depicted in Figure 1. The first source is the composer, which provides audio examples of the sonic characteristic that is desired to be found (a query). The second source is the computer, which aims at extending, for all the database, the criteria applied to the construction of the query.

By combining the objective, vector representation of audio samples and the subjective, perceptually-driven query, the system builds a model for what it detects as the characteristic sought by the composer. This model is, then, extrapolated, so that other audio samples that correspond

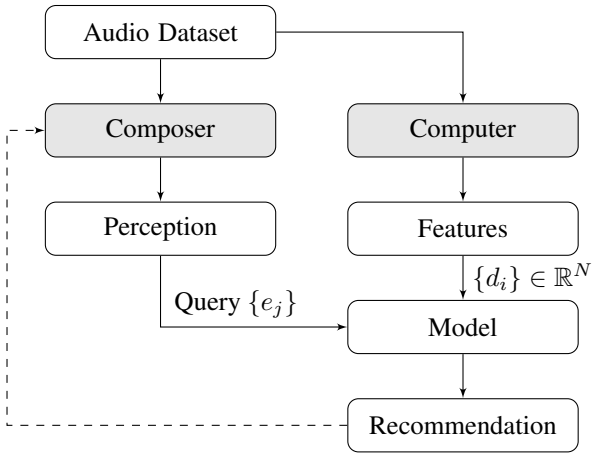


Figure 1. Block diagram for the proposed system.

to that characteristic, are found. Then, the system yields a recommendation, which can be used by the composer.

In our work, the calculation of features (yielding representations in the $\mathbb{R}^N$ vector space) follows the same general structure used in previous work [16, 17], as described in Section 2.1. However, we experimented several different methods for modelling queries. This will be described in details in Section 2.2.

2.1 Feature calculations

The process of obtaining a vector representation in the feature space begins with calculating a framewise Short-Time Fourier Transform, using 23ms frames, with a 50% overlap ratio, multiplying each frame by a Hanning window and then calculating the absolute value $X_q[k]$ of the DFT of each frame. For each frame q, a set of acoustic features are calculated, as described in Table 1. Also, the first and second-order differentials of each features are calculated.

For each feature and its differentials, we calculate a set of statistics. This set comprises mean and variance, which give an idea of the general behaviour of these features. We also obtain the slope (considering a linear regression) and the value and time location of the maximum and minimum values, to depict the evolution of features over time.

This process associates each audio sample to a descriptive vector, which we expect to encode its perceptual characteristics. As it will be seen, there are many ways to model multiple-example queries. This will be discussed in the next section.

2.2 Models for queries

The modelling process for queries aims at detecting relevant sonic characteristics as described by the composer using examples e_j. This process assumes that these characteristics are encoded within the objective $\mathbb{R}^N$ feature space defined by the calculated features (as described in Section 2.1). The model gives a score to each element d_i of the database, and infers that the element with the highest score also presents the characteristic desired by the composer.

We evaluated several methods for obtaining the model. This was done because there is no particular prior reason to assume that a model is better than another, hence a detailed evaluation is necessary. Each method has its own rationale, which will be described below.

The first method, **Minimum Distance** (MD), gives an element a score which is the inverse of the minimum Euclidean distance between itself and any element from the query, as depicted in Expression 1. It is equivalent to performing a few queries-by-example (using single examples) in parallel, and then selecting the best results. Hence, it may not be considered as a valid method for Query-by-Multiple-Examples.

$$\mathrm{MD}(d_i) = 1/(\min_j \|d_i - e_j\|). \quad (1)$$

The second method, **Minimum Mutual Distance** (MMD), was inspired by work by Schnitzer et. al [18], which observes that an element that belong to a cluster must not only be close to the cluster but also distant from other clusters. Hence, it scores each sample from the database with the minimum distance between itself and an element from the query minus the minimum distance between itself and an element in the database, as described in Expression 2. Although this method is more complex than MD, it also does not perform a Query-by-Multiple-Examples, but multiple queries-by-single-example in parallel.

$$\mathrm{MMD}(d_i) = 1/(\min_j \|d_i - e_j\| - \min_k \|d_i - d_k\|). \quad (2)$$

Third, we consider the **Naive Bayes** (NB) approach. In this method, the elements of the query are used to estimate the mean μ_n and variance σ_n of a gaussian model for each dimension, which is assumed to be independent from the others. Thus, the score of an element of the database is given by:

$$\mathrm{NB}(d_i) = \prod_{n=1}^{N} g(d_{i,n}, \mu_n, \sigma_n), \quad (3)$$

where $g(x, \mu, \sigma) = (1/\sigma\sqrt{2\pi}) \exp(-(\mu - x)^2/2\sigma^2)$.

The NB approach assumes that dimensions spanned by features are orthogonal. There is no evidence that this condition is true, therefore we applied Principal Component Analysis (PCA) to obtain an orthogonal projection $\boldsymbol{B}$ of the query with a minimal approximation error. We expect that this projection will be a better representation for the composer's perception than the raw feature set itself.

The projection is made using $M - 1$ vectors, where M is the number of elements in the query, because this is the maximum rank of the projection provided by PCA. The projection is calculated using only elements from the query, and then applied over the whole database. Then, the naive Bayes approach is used normally as described above, hence the method is named **PCA-Bayes** (PB).

Hence, four different modelling methods were applied. Two of them are simple applications of simple query-by-example schemas, whereas the other two use the correlations within the query to build a different model. For comparison purposes, a random recommender (recommending a random element from the database) was also used in the evaluation set, which will be described in detail in the next section.

Table 1. Brief description of acoustic features

Feature	Description
Energy	Sum of the squared values of $X_q[k]$. Indicates how loud the frame is.
Spectral centroid	Centroid of $X_q[k]$. Correlates with the brightness of the frame.
Spectral roll-off	Frequency above which there is less than 15% of the energy of a frame.
Spectral flatness	Indicates how close the frame is to white noise.
Mel-Frequency Cepstral Coefficients	Vector representation of audio textures, inspired in cochlea models.

3. EVALUATION AND RESULTS

The evaluation process aimed at detecting whether the system is able to retrieve audio samples from the database as if the composer was searching for it.

To reproduce this scenario, we interviewed three composers, all of them graduate students from the Music Department. After a brief talk about their composition processes, we asked them to group pre-defined audio samples (from a set of 51 elements, around 10s long, extracted from a personal music collection) into subsets that made sense for them, and, if possible, explain what criteria was used for grouping. Their criteria was significantly different, as it is discussed below.

Composer **C1** stated that processed samples were used as material in the piece composition process in order to reach a particular sound characteristic. The presented sets were predominantly grouped using characteristics linked to auditory aspects of each sample. Typical grouping criteria were labeled *dry/dark timbre*, *static harmonic sound*, *brightness and compression* and *glissando*.

Composer **C2**'s composition process uses the semantic values of the audio samples, in addition to their sound. The grouping process considered semantic-valued characteristics, that is, the context in which each sample was obtained. In this case, grouping criteria were of higher level, such as *celtic*, *drums* and *prepared piano*.

Composer **C3** preferred to use sound samples as source for granular synthesis processes. Grouping criteria was based on auditory characteristics of samples, as well as general semantics. Among the grouping criteria, it was possible to find *vocal*, *regular*, *synthesizer*, *orchestral* and *attack modes*.

The subsets presented by each composer were assumed as ground-truth, that is, a query containing some elements of each subset should find the remaining elements. Several queries were made from each group, considering different numbers of elements. The queries made from the groups of each composer were considered separately, so that their different reasoning towards the search process could be analyzed.

For each query, the system was asked to retrieve three samples from the database. A retrieved sample is considered correct if it belongs to the subset from which the query was made. Then, we calculated two accuracy measures, *Acc1* and *Acc2*, defined as:

$$\text{Acc1} = \frac{\text{\# retrievals with at least 1 correct sample}}{\text{\# total queries}}. \quad (4)$$

$$\text{Acc2} = \frac{\text{\# correctly retrieved samples}}{\text{\# total retrieved samples}} \quad (5)$$

Acc1 measures the probability that a query will retrieve at least one useful sample, which is desirable because it means that the search space has been narrowed. *Acc2* measures the probability that a retrieved sample is useful, which is also a desirable characteristic of the system. It is important to note that *Acc1* is higher when the system avoids false negatives, whereas *Acc2* is higher when the system presents fewer false positives.

We tested the system using all query modelling methods discussed in Section 2.2. The results for the datasets related to composers C1, C2 and C3 were considered separately. *Acc1* and *Acc2* for each test case are shown, respectively, in Figures 2 and 3.

Figure 2. Results using Acc1.

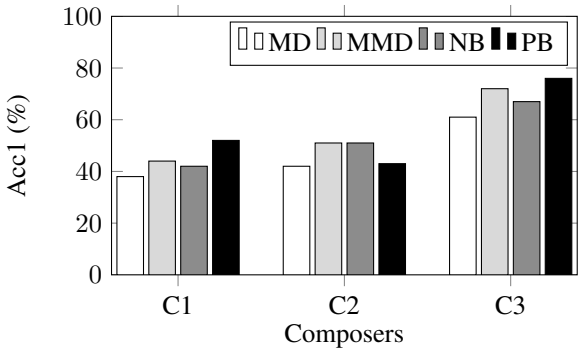
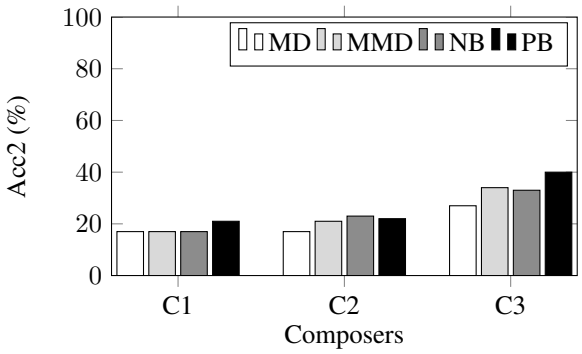


Figure 3. Results using Acc2.



As it can be seen, the MD method performed worse for all of the test cases. Nevertheless, its results are comparable to those yielded by the other methods, i.e., it is not possible to claim that the difference is huge. Hence, MD may be considered a baseline method for further discussion.

Another key result is that MMD always performs better than MD. This shows that its assumption – that an element

that belongs to a set must not only be close to the set, but also distant from the other sets [18] – is useful for our purposes. More interesting results, however, derived from the application of the NB and PB methods.

In the case of C1 and C3, it is clear that PB outperforms all other methods, considering either *Acc1* or *Acc2*. However, in the case of C2, PB is outperformed by both NB and MMD for *Acc1* and has a similar performance considering *Acc2*. This may be due to the grouping process performed by composer C2.

Composer C2 applied a predominantly semantic grouping of elements, that is, elements that do not sound alike, but are culturally related (for example: the sounds of Celtic fiddles and of Irish tap dancers) were grouped together. This means that the feature space – which describes auditory characteristics – was divided in several clusters with useful data. This caused the MMD method to present a better performance.

Also, NB performed better than PB for this case, which shows that the original feature set defined better local maxima to the score function than the reduced-dimension orthogonal set. Although orthogonality is a desirable trait, it is important to note that a high-dimensional space has a greater chance to have at least one dimension in which any points, chosen at random, are positioned in a linearly separable convex hull. However, if more dimensions were used in the PCA reduction, they would be linearly dependent of the previous ones, which means another method for dimension reduction would be required.

The next section conducts further discussion on these results.

4. DISCUSSION

One interesting point shown by the results is that they are highly dependent on the criteria used by the composer for classification. In our tests, Acc1 varied from 50% to 75%, and Acc2 from 20% to 40%. This is a great relative step, which has to be considered when performing future evaluations.

Although the system was evaluated using objective measures, it is important to note that it is a recommendation system, which will interact with users. Hence, it is necessary to conduct further tests to detect whether the results provided by the system are useful for the composer, despite of not being expected *a priori*. These results may show if the system is able to recommend useful samples (maybe some sample that may be used, but the composer would not have thought of about alone), thus allowing generalization towards a bigger database.

The results obtained above show that each mindset for sample grouping – auditory or semantic – can be better modelled by a different algorithm: auditory-based criteria are well suited for the PB method, whereas semantic-based are better modelled using MMD. This happens because the dimensions spanned by acoustic describe auditory characteristics, which means semantic information is only present as an underlying function of the acoustic features. Possible ways to deal with this situation may involve other dimensionality-reduction techniques, such as Independent Component Analysis (ICA) or non-linear PCA.

Although MMD – which corresponds to multiple queries by single-example – outperforms NB and PB for the queries corresponding to composer C2, it is important to note that the system was built aiming at detecting sonic characteristics, rather than semantics. For the queries corresponding to composers C1 and C3, which follow the idea of describing sounds without considering semantics, PB – a method that clearly takes advantage of the correlations within the multiple example query – outperforms all others. Therefore, the results show that the proposed system, using PB, provided a meaningful contribution towards the problem of searching for sound characteristics within a database.

Next section presents conclusive remarks.

5. CONCLUSION

We presented a search method, namely Query-By-Multiple-Examples. It receives as input a few audio samples that examples of a particular sound characteristic yields other samples, from a database, that also present that characteristic. The method is aimed at extending the search possibilities of composers in the context of sound-based music, that is, music based on the evolution of sonic characteristics.

The method is based on mapping all audio samples from a database into a vector space using low-level acoustic features. Queries are received from the user and modelled, yielding a score for each element in the database. We tested four different methods for modelling the query: minimum distance and minimum mutual distance (corresponding to several queries-by-single-example executed in parallel), and naive Bayes and PCA-Bayes (corresponding to query-by-multiple-example).

We evaluated all variations using ground-truth queries, defined by three different musicians. They provided very different proposals for the grouping of similar audio samples, according to their typical composition process. It has been shown that considering the correlations within the provided inputs improves the search accuracy for auditory-inspired queries.

The obtained results, however, do not account for the interaction between composers and the computer, which is an important part of the composition process. Thus, it is necessary to evaluate whether the wrong results yielded by the system are useful suggestions (despite of being unexpected) or if they are just plain wrong, and, more than that, how the system would behave in an unknown database. This points a clear direction for future work.

Acknowledgments

The authors thank FAPESP (proc. 2013/17329-5) for funding this research.

6. REFERENCES

[1] M. Solomos, *De La Musique Au Son – L'emergence du Son Dans la Musique des XXe-XXIe Siecles*. Presses Universitaires de Rennes, 2013.

[2] D. Schwarz, "Current research in concatenative sound synthesis," in *Proceedings of the ICMC 2005*, 2005.

[3] T. T. Opie, "Creation of a real-time granular synthesis instrument for live performance," Master's thesis, Queensland University of Technology, 2003.

[4] G. Nouno, A. Cont, G. Carpentier, and J. Harvey, "Making an orchestra speak," in *Proceedings of the SMC 2009*, 2009.

[5] J. C. Platt, C. J. Burges, S. Swenson, C. Weare, and A. Zheng, "Learning a gaussian process prior for automatically generating music playlists," in *Proc. Advances in Neural Information Processing Systems*, vol. 14, 2002, pp. 1425–1432.

[6] S. Pauws, "Pats: Realization and user evaluation of an automatic playlist generator," in *Proceedings of the ISMIR*, 2002, pp. 222–230.

[7] Y. Kodama, S. Gayama, Y. Suzuki, S. Odagawa, T. Shioda, F. Matsushita, and T. Tabata, "A music recommendation system," in *Consumer Electronics, 2005. ICCE. 2005 Digest of Technical Papers. International Conference on*, Jan 2005, pp. 219–220.

[8] B. Shao, D. Wang, T. Li, and M. Ogihara, "Music recommendation based on acoustic features and user access patterns," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 8, pp. 1602–1611, Nov 2009.

[9] B. Jensen, J. Saez Gallego, and J. Larsen, "A predictive model of music preference using pairwise comparisons," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, March 2012, pp. 1977–1980.

[10] D. Bogdanov, M. Haro, F. Fuhrmann, A. Xamb, E. Gmez, and P. Herrera, "Semantic audio content-based music recommendation and visualization based on user preference examples," *Information Processing & Management*, vol. 49, no. 1, pp. 13 – 33, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0306457312000763

[11] P. Knees and M. Schedl, "A survey of music similarity and recommendation from music context data," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 10, no. 1, pp. 2:1–2:21, Dec. 2013. [Online]. Available: http://doi.acm.org/10.1145/2542205.2542206

[12] O. Moravec and J. Stepanek, "Verbal descriptions of musical sound timbre and musician's opinion of their usage," in *FORTSCHRITTE DER AKUSTIK*, 2005.

[13] M. Sarkar, B. Vercoe, and Y. Yang, "Words that describe timbre: a study of auditory perception through language," in *Language and Music as Cognitive Systems Conference (LMCS-2007)*, 2007.

[14] G. Li and A. Khokhar, "Content-based indexing and retrieval of audio data using wavelets," in *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, vol. 2, 2000, pp. 885–888 vol.2.

[15] L. Lancieri, M. Manguin, and S. Mangon, "Evaluation of a recommendation system for musical contents," in *Multimedia and Expo, 2008 IEEE International Conference on*, June 2008, pp. 1213–1216.

[16] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *Speech and Audio Processing, IEEE Transactions on*, vol. 10, no. 5, pp. 293–302, Jul 2002.

[17] M. Helén and T. Virtanen, "Audio query by example using similarity measures between probability density functions of features," *EURASIP J. Audio Speech Music Process.*, vol. 2010, pp. 1:1–1:9, Jan. 2010. [Online]. Available: http://dx.doi.org/10.1155/2010/179303

[18] D. Schnitzer, A. Flexer, M. Schedl, and G. Widmer, "Using mutual proximity to improve content-based audio similarity," in *Proceedings of the ISMIR*, 2011.