Musical Acoustics, Timbre, and Computer-Aided Orchestration Challenges

Aurélien Antoine[†] *and Eduardo R. Miranda*

Interdisciplinary Centre for Computer Music Research (ICCMR), Plymouth University, UK
[†]aurelien.antoine@postgrad.plymouth.ac.uk

ABSTRACT

Research on timbre has produced a lot of work over the last 50 years, whether on analysis of single tone, instrument timbre, synthetic timbre, or perception and emotion. However, there is still a lot to investigate for polyphonic timbre. This aspect is perhaps one of the biggest challenges in the field of computer-aided orchestration. This paper reports on the development of a system capable of automatically identify perceptual qualities of timbre within orchestral sounds. Work in this area could enrich current and future computing systems designed to aid musical orchestration.

1. INTRODUCTION

Timbre is a complex and multidimensional attribute of sound, whose definition has been largely discussed among the research community, see [1] or [2] for examples. However, a standard definition often cited in papers related to timbre is the one proposed by the American National Standards Institute [3]: *"Timbre is that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar"*. In summary, timbre is defined as all the sound attributes, except pitch, loudness, and duration, which allow us to distinguish and recognize that two sounds are dissimilar [4]. However, some research challenges this definition and suggests that timbre, pitch, and loudness can interfere in the auditory sensation [5, 6, 7]. Furthermore, several works have demonstrated the importance of acoustic features in defining musical timbre [8, 9, 10].

We are particularly interested in the perception of musical timbre. Research into this field has identified various important audio features to represent timbral perception [11, 12, 13, 14, 15]. However, timbre can also be described using verbal descriptors [16, 4, 17]. For example, terms like brightness or roughness are words from everyday language used to describe perceived musical timbres. These terms can be more intuitive for musicians, composers, and audio engineers, than their acoustic correlates (e.g. spectral centroid, critical bands, etc.).

The system presented in this paper is built upon the existing research about the correlation between acoustic features and verbal description of timbre qualities. A similar approach has been used in [18] and [19], but they utilized a visual representation approach of timbre properties. We believe that using verbal descriptors of timbral attributes can aid in making the tool accessible to a broad audience by alleviating the need to have expertise in acoustics or psychoacoustics.

The aim of this paper and the system presented within is to propose an effective and direct application of the work done on semantic description of timbre perception. The development of such a system could aid in standardizing the metrics for perceived responses of timbral qualities, and also enrich computer-aided orchestration systems.

2. AUTOMATIC TIMBRE CLASSIFICATION SYSTEM

This section introduces the technical aspects of the automatic timbre classification system. First, we discuss the verbal descriptors currently implemented in the system, with details of their corresponding acoustic features. Then, we report on the design of our classification algorithms developed to identify the dominant timbral content within an audio file.

2.1. Verbal Descriptors and Acoustic Features

For our timbre classification system, we chose to represent timbral qualities by using verbal descriptors. There are currently five timbral attributes implemented in the presented system: *breathiness*, *brightness*, *dullness*, *roughness* and *warmth*. Corresponding acoustic features for each attribute are detailed below.

2.1.1. Breathiness

To identify the level of *breathiness* within an audio file, we need to calculate the amplitude of the fundamental frequency against the noise content, and also the spectral slope [20, 21]. The bigger the ratio between amplitude of the fundamental and the noise content, the breathier the sound.

2.1.2. Brightness

The acoustic correlates for the attribute *brightness* are the spectral centroid and the fundamental frequency [22, 23]. The higher the spectral centroid, the brighter the sound.

2.1.3. Dullness

Similar to brightness, to identify the *dullness* of a sound we need to calculate its spectral centroid. However, in this case, a low spectral centroid value suggests that the sound is dull [24].

2.1.4. Roughness

To determine the *roughness* index of a sound, we need to calculate the distance between adjacent partials in critical bandwidths and also the energy above the 6th harmonic [25, 26, 8].

2.1.5. Warmth

To calculate the *warmth* of a sound, we need to calculate its spectral centroid and retrieve the energy in its first three harmonics [4, 27]. A low spectral centroid and a high energy in the first three harmonics suggest that the sound is warm.

2.1.6. Acoustic Features Analysis

We have developed our acoustic features analysis algorithm within the `Matlab`[1] environment. We also utilized some functions from the `MIRtoolbox`[2] [28]. This toolbox proposes several `Matlab` functions designed specifically for retrieving and extracting various musical features from audio files.

Our system starts by computing the spectrum of the audio file input by the user, using the `mirspectrum` function from the `MIRtoolbox`. Then, the system calculates the acoustic features for each attribute as described previously. Finally, the analysis returns a value for each attribute, which are all stored in a singular vector that will be used by the classification algorithms, described in the following section.

2.2. Classification Algorithms

There is a significant amount of work about the perception and description of musical timbre, upon which this work is based. However, there are no universally agreed metrics for classifying audio samples according to perceived responses of timbre quality. Therefore, before the development of a classification algorithm, a comparative scale for each timbral attribute must be established.

2.2.1. Comparative Scale

We began with gathering data in order to establish the scale of each timbral attribute. Therefore, we collected over 250 audio recordings of various orchestral pieces and split each audio recording into 1, 2, 3 and 4 seconds long audio files. We chose to go up to 4 seconds as it could represent a bar's worth of music, or it can correspond with the length of a computer-aided orchestration system's output. The acoustic features analyzed are time-based and therefore longer durations could omit important data and produce inaccurate indexes. As a result, we analyzed over 236 000 audio files and collected values for each timbral attribute.

We analyzed the dataset in order to retrieve various statical values, such as minimum value, maximum value, standard deviation, and the distribution of the values for each attribute. This allowed us to establish a scale for each attribute and therefore develop a normalization algorithm in order to be able to compare the timbral values. The system continually calibrates the statistical values as new audio files are analyzed in order to adjust and improve the comparative scale. The dataset is then used by the classification algorithms we developed in `Python 3.5`.

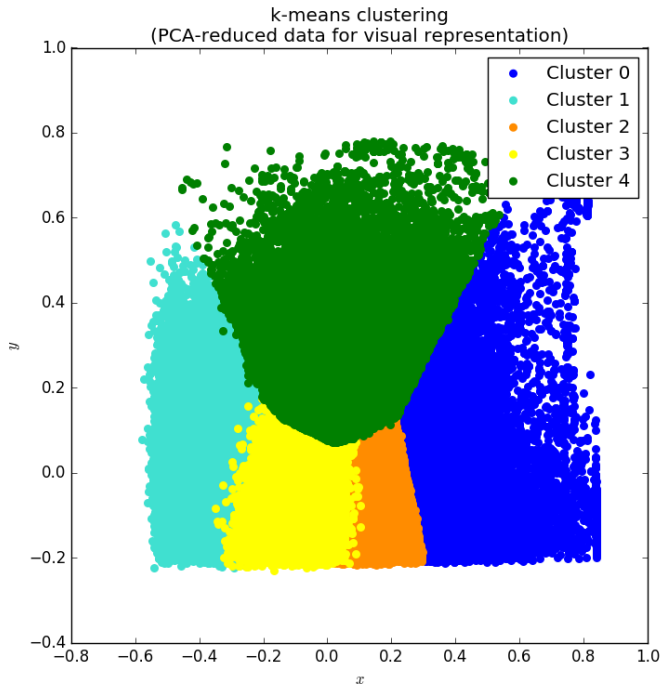

Figure 1. Graph showing the results for the *k*-means clustering performed on the 236 000 samples dataset.

2.2.2. K-Means Clustering

We first experimented with an unsupervised learning algorithm to identify a classification model. We wanted to divide the dataset into 5 parts—to represent the 5 timbral attributes. Therefore, we decided to perform a *k*-means clustering using Lloyd's algorithm [29] on our dataset.

We used the `KMeans` function from the `Scikit-Learn` package[3], with a `k-means++` initialization (which speeds up convergence [30]). Each entry of the dataset was input as a singular vector into the *k*-means algorithm. Results of the clustering is shown in Figure 1.

2.2.3. Support Vector Machine

Following the experiment with *k*-means clustering, we opted to implement a machine learning model usually utilized for data classification and regression analysis: Support Vector Machine (SVM) [31]. First, we needed to create a corpus training for the SVM algorithm. Therefore, for each timbral attribute we took and labeled the 250 highest index samples from the normalized dataset. Thus, the corpus training contained 1250 labeled samples.

[1] `http://www.mathworks.com/products/matlab/`
[2] MIRtoolbox is available at `https://goo.gl/d61EO0`
[3] `http://scikit-learn.org/stable/index.html`

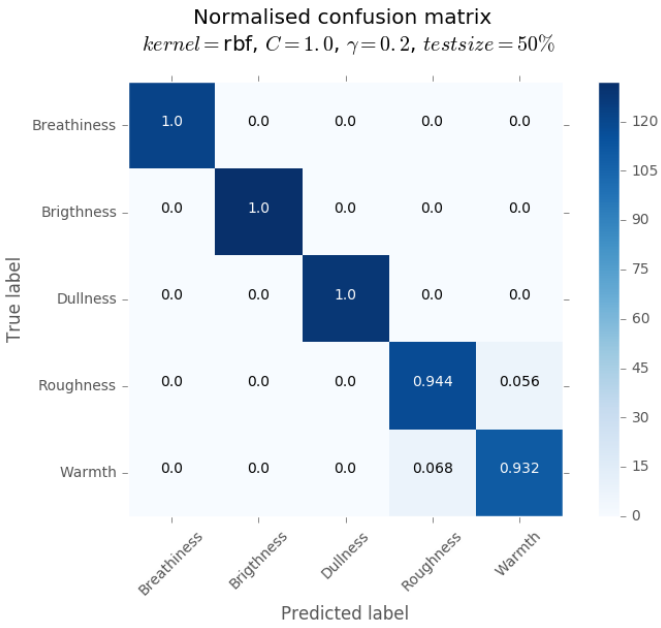

Figure 2. Representation of the normalized confusion matrix for the SVM algorithm, with test size = 50% (625 training samples, 625 testing samples).

We used the `svm.SVC` function from the `Scikit-Learn` package. After performing the `Scikit-Learn`'s functions for parameters' tuning, we selected a `rbf` kernel for our SVM algorithm, which performed a score of 0.976 (with 1.0 being the maximum). The resulted confusion matrix is shown in Figure 2.

3. RESULTS AND DISCUSSION

In [32], we reported on a pilot system that used the acoustic analysis described in section 2.1. We also reported on an experiment conducted with 20 participants to evaluate the accuracy of the analysis. Results indicated a correlation between the system's timbre analysis and human perception, which led us to develop the presented automatic classification system.

Timbral values being on different scales, the data gathering, and audio analysis helped us to evaluate the distribution of the values for each verbal descriptor, and therefore be able to normalize the data in the range (0.0–1.0), which then can be used in classification algorithms.

We performed a *k*-means algorithm on the 236 000 normalized samples in order to divide the dataset in 5 clusters, corresponding to the 5 timbral attributes. The result is shown in Figure 1, using a Principal Component Analysis (PCA) algorithm for visual representation. The *k*-means algorithm has been able to produce 5 distinct clusters, however due to the nature of the classification method the correspondent timbral attributes remains unknown until evaluated and labeled by the user.

To test the performances of the SVM algorithm, we divided the corpus training set into different batches of learning samples and testing samples. The SVM estimator with a `rbf` kernel performed a success score of 0.976. Figure 2 shows the normalized confusion matrix with a test size of 50%, in this case 625 samples for learning and 625 samples for testing. Although the SVM presents successful learning scores, it is dependent on a previously created corpus training. While this required action can be seen as a negative additional task, it could enable the user to create their own training data, which would represent their perceptual preferences.

4. CONCLUSION

In this paper, we have introduced the development of a computing system capable of automatically identifying the timbral qualities contained in orchestral audio samples. This system is built upon the existing research about the relation between acoustic features and verbal description of timbre qualities. It currently integrates five verbal descriptors: *breathiness*, *brightness*, *dullness*, *roughness* and *warmth*, introduced in section 2.1. Then, we detailed the creation of a comparative scale, based on audio recordings analysis. This scale enabled us to normalize data across the 5 timbral attributes, which is used for the classification algorithms presented in section 2.2.

Both *k*-means and SVM algorithms performed successful samples classification. However, a user action is still required, whether afterward for clusters labeling for *k*-means, or beforehand for corpus training's creation for the SVM. Nevertheless, these additional tasks can be used to calibrate the classification algorithms to the user's own musical perception, which could offer a solution to the challenging variation in music perception between individuals. Such developments could enrich computer-aided orchestration systems by harnessing perceptual aspects of polyphonic timbre within orchestral sound.

REFERENCES

[1] C. L. Krumhansl, "Why is musical timbre so hard to understand," *Structure and perception of electroacoustic sound and music*, vol. 9, pp. 43–53, 1989.

[2] D. Smalley, "Defining timbre—refining timbre," *Contemporary Music Review*, vol. 10, no. 2, pp. 35–48, 1994.

[3] American National Standards Institute, *Psychoacoustic terminology S3:20*. New York, NY: American National Standards Institute, 1973.

[4] R. Pratt and P. Doak, "A subjective rating scale for timbre," *Journal of Sound and Vibration*, vol. 45, no. 3, pp. 317–328, 1976.

[5] J. R. Platt and R. J. Racine, "Effect of frequency, timbre, experience, and feedback on musical tuning skills," *Perception & Psychophysics*, vol. 38, no. 6, pp. 543–553, 1985.

[6] R. D. Melara and L. E. Marks, "Interaction among auditory dimensions: Timbre, pitch, and loudness," *Percep-*

tion & psychophysics, vol. 48, no. 2, pp. 169–178, 1990.

[7] V. C. Caruso and E. Balaban, "Pitch and timbre interfere when both are parametrically varied," *PloS one*, vol. 9, no. 1, p. e87065, 2014.

[8] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and models*. Springer Science & Business Media, 2007, vol. 22.

[9] J. F. Schouten, "The perception of timbre," in *Reports of the 6th International Congress on Acoustics*, vol. 76, 1968, p. 10.

[10] J. M. Grey, "An exploration of musical timbre," Ph.D. dissertation, Stanford University, 1975.

[11] J. M. Grey and J. W. Gordon, "Perceptual effects of spectral modifications on musical timbres," *The Journal of the Acoustical Society of America*, vol. 63, no. 5, pp. 1493–1500, 1978.

[12] S. McAdams, S. Winsberg, S. Donnadieu, G. De Soete, and J. Krimphoff, "Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes," *Psychological research*, vol. 58, no. 3, pp. 177–192, 1995.

[13] R. A. Kendall, E. C. Carterette, and J. M. Hajda, "Perceptual and acoustical features of natural and synthetic orchestral instrument tones," *Music Perception: An Interdisciplinary Journal*, vol. 16, no. 3, pp. 327–363, 1999.

[14] V. Alluri and P. Toiviainen, "Exploring perceptual and acoustical correlates of polyphonic timbre," *Music Perception: An Interdisciplinary Journal*, vol. 27, no. 3, pp. 223–242, 2010.

[15] S. McAdams, *The Psychology of Music*, 3rd ed. Elsevier, 2013, ch. Musical Timbre Perception, pp. 35–67.

[16] G. von Bismarck, "Timbre of steady sounds: A factorial investigation of its verbal attributes," *Acta Acustica united with Acustica*, vol. 30, no. 3, pp. 146–159, 1974.

[17] 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.

[18] A. Tsiros and G. Leplâtre, "Animorph: animation driven audio mosaicing," in *Proceedings of the Conference on Electronic Visualisation and the Arts*. British Computer Society, 2015, pp. 115–116.

[19] S. Soraghan, A. Renaud, and B. Supper, "A perceptually motivated visualisation paradigm for musical timbre," in *Proceedings of the conference on Electronic Visualisation and the Arts*. British Computer Society, 2016, pp. 53–60.

[20] J. Hillenbrand and R. A. Houde, "Acoustic correlates of breathy vocal qualitydysphonic voices and continuous speech," *Journal of Speech, Language, and Hearing Research*, vol. 39, no. 2, pp. 311–321, 1996.

[21] M. A. Landera and R. Shrivastav, "Effects of spectral slope on perceived breathiness in vowels," *The Journal of the Acoustical Society of America*, vol. 119, no. 5, pp. 3339–3340, 2006.

[22] E. Schubert, J. Wolfe, and A. Tarnopolsky, "Spectral centroid and timbre in complex, multiple instrumental textures," in *Proceedings of the international conference on music perception and cognition, North Western University, Illinois*, 2004, pp. 112–116.

[23] A. C. Disley, D. M. Howard, and A. D. Hunt, "Timbral description of musical instruments," in *International Conference on Music Perception and Cognition*, 2006, pp. 61–68.

[24] D. B. Bolger, "Computational models of musical timbre and the analysis of its structure in melody," Ph.D. dissertation, University of Limerick, 2004.

[25] E. Terhardt, "On the perception of periodic sound fluctuations (roughness)," *Acta Acustica united with Acustica*, vol. 30, no. 4, pp. 201–213, 1974.

[26] v. W. Aures, "A procedure for calculating auditory roughness," *Acustica*, vol. 58, no. 5, pp. 268–281, 1985.

[27] R. Ethington and B. Punch, "Seawave: A system for musical timbre description," *Computer Music Journal*, pp. 30–39, 1994.

[28] O. Lartillot, P. Toiviainen, and T. Eerola, "A matlab toolbox for music information retrieval," in *Data analysis, machine learning and applications*. Springer, 2008, pp. 261–268.

[29] S. Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[30] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[31] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[32] A. Antoine, D. Williams, and E. Miranda, "Towards a timbral classification system for musical excerpts," in *Proceedings of the 2nd AES Workshop on Intelligent Music Production*, London, 2016.