Sonic Print: Timbre Classification with Live Training for Musical Applications

Jean-François Charles[1], Gil Dori[2], and Joseph Norman[3],

[1] University of Iowa, School of Music, USA
[2] Ben Gurion University, Israel
[3] Grinnell College, USA

Abstract. Sonic Print provides performers, composers and improvisers the ability to use automatic timbre recognition in live musical applications. Sonic Print can be trained quickly, during a performance, to discriminate between four classes of sounds. The musician can use the result of the live timbre classification to process sound with different audio effects, or to trigger musical events. Sonic Print is implemented as native Max code (widely used by artists), and as a Max for Live device (for integration in Ableton Live, a popular Digital Audio Workstation). We detail how Sonic Print contributed to two distinct creations in Israel and the United States.

Keywords: Interactive machine learning; Live timbre classification; Multivariate linear regression; Music composition; Music performance.

1 Introduction

Some musicians need the ability to classify sound in real time according to select audio features. For instance, a singer automatically applies a long reverberation to a voice sung in a high register, and distortion to spoken voice. Or a composer triggers specific sound files every time a performer transitions to a breathy timbre, and other files when a particular note is played. Furthermore, the nature of musical practice calls for a classifier that can be trained in an interactive way, during a performance or in rehearsal. Live training is necessary to automatically take into account the specificity of the current microphone and acoustic conditions. Interactive training also provides new creative possibilities. An improviser, for example, can not only perform, but also design different timbre-specific interaction patterns on the fly. Finally, such a live timbre classification tool must be seamlessly integrated in musicians' workflow and performance environment.

Sonic Print uniquely combines audio timbre classification and live training in an interface designed for intuitive use in musical applications. It is implemented in environments that are widely used by musicians. The users need no previous knowledge of machine learning algorithms.

2 Related Work

Our work extends existing literature at the crossroads of timbre classification, interactive machine learning, and the design of tools facilitating musicians' creativity.

2.1 Timbre Classification

Timbre *covers many parameters of perception that are not accounted for by pitch, loudness, spatial position, and duration. It is thus, by definition, multidimensional* (McAdams, 1999). Different musical contexts influence how humans perform timbre discrimination (Grey, 1978). Since timbre is defined by human perception, no algorithm can claim perfect accuracy in classification. Nevertheless, the quest for automatic timbre recognition has produced many useful results, allowing to classify musical instruments (Park & Cook, 2005), ethnomusicological recordings (Fourer et al., 2014), environmental sounds (Chachada & Kuo, 2014), acoustic activity (Laput et al., 2018).

With Sonic Print, we aim at a perceptually relevant classification of sounds: we operate in the domain of timbre classification. We apply the word timbre to classes going beyond the case of musical instruments; for instance, we could try to discriminate between the timbres of a fricative consonant, a dishwasher in rinsing mode, an alto flute in the low register, and chirping cicadas.

2.2 Machine Learning for Musical Performance

In the field of machine learning applied to music performance, a significant body of work has been published with the goal of mapping physical gesture to sound processing, with input vectors built from sensor values.

The MnM toolbox for gesture to sound mapping uses multivariate linear interpolation and is integrated in Max (Bevilacqua et al., 2005). Gillian developed a machine learning toolbox for musicians, which enables live training and classification of multivariate temporal signals (Gillian et al., 2011). Rebecca Fiebrink has produced a solid body of work on the creative applications of interactive machine learning. Her Wekinator toolbox has been used to *easily prototype complex relationships between performer gesture and performative outcome* (Schedel et al., 2011). It offers a choice of supervised learning algorithms working with low training and running time, compatible with the limited availability of training examples, and enabling on-the-fly machine learning (Fiebrink et al., 2009). They mention training times of a couple of seconds. The Wekinator is offered as a standalone application. It does not embed audio analysis, but may communicate with musical software performing spectral analysis.

Tools for live timbre classification integrated in a musician's work environment are not widely available. Moreover, while these tools are highly customizable and flexible, they present a steep learning curve to musicians with no knowledge of machine learning concepts and algorithms.

3 Live Timbre Classification

The purpose of Sonic Print is to perform automatic classification of sounds. We explain here our choice of audio features and machine learning algorithm.

3.1 Choice of Audio Features

Sonic Print performs both the learning and the analysis steps in real time, on a live stream of audio. To get the most responsive results, we restrict our observations to the

instantaneous repartition of energy in the frequency spectrum, rather than to the temporal succession of features.

The Mel-Frequency Cepstral Coefficients (MFCC) are widely used to characterize the instantaneous repartition of energy along the frequency spectrum, including in speech recognition, speaker identification (Chougule & Chavan, 2013), environmental sound classification (Chachada & Kuo, 2014), and musical instrument recognition (Sturm et al., 2010). We use a comparable approach, modified in two aspects. First, in order to give the real-time algorithm a greater computational efficiency and a shorter latency, we dispense with the conversion from spectral to cepstral domain: we work directly with the spectral energy distribution. Second, as we explain in the next paragraph, we use filters distributed on the Bark scale rather than the Mel scale.

Zwicker introduced the Bark scale after extensive experimental research on critical bands (Zwicker et al., 1957). The Bark and Mel scales have similarities, but differ in origin: the Mel scale was built after a perceptive study of melodic relations, whereas the Bark scale was built after the more general critical bands. Thus, by construction, the set of Bark filters is more attuned to the human perception of timbre. The Bark scale has been used successfully in many applications including the classification of percussion instruments (Brent, 2009), hand gestures on a tabletop (Jathal, 2017), music genre (Costa et al., 2012), vowels (Hillenbrand & Gayvert, 1993), spoken Marathi numerals (Ghule & Mukherji, 2016), stress and emotion in natural speech (He et al., 2011) and (Lugger & Yang, 2008), truthful vs. deceptive speech (Sanaullah & Gopalan, 2013), and noise (Eamdeelerd & Songwatana, 2008).

3.2 Choice of Learning Algorithm

We work with input vectors of dimension 24 (Bark values), which we assign to distinct classes. For this first version of the tool, we decided to fix the number of classes to four. Since the user knows which timbre they want to classify into which class, we are working in the context of supervised learning.

We could use a general algorithm such as k-nearest neighbors (k-NN), which is widely used and offers a straightforward implementation. Yet, its performance for classification of live data is slow compared to, for instance, that of neural networks. Thus, we look into the data properties to see if we could use a faster algorithm. Specifically, if it is likely that the classes we want to discriminate between are linearly separable, then we will be able to use algorithms that are more efficient than k-NN.

Following in the steps of Cover (Cover, 1965), Gardner showed that the probability that data in a multidimensional space be linearly separable increases with the correlation of the points in the input space, especially when similar input patterns have identical outputs (Gardner, 1988). It is exactly the situation for our application: the musician assigns a single output – sound class – to a set of correlated input vectors – Bark values representing correlated timbres. Given that we work with limited training data to allow for live learning, it is even more likely that the four training sets will be linearly separable. Thus, using a single-layer perceptron to perform the classification is appropriate.

Although the data sets are likely to be linearly separable, it is not guaranteed, especially in the case of live training during a performance. If the training data are not linearly separable, a perceptron with a sigmoid or tanh activation function cannot be trained – the learning algorithm does not converge. We thus decide to use no sigmoid activation function, reducing the perceptron to a multivariate linear regression. This approach presents several advantages. First, we get a direct learning step through linear algebra, which is faster than the iterative training required for a sigmoid-activated

perceptron. Second, this approach is guaranteed to produce a result in a short, finite time, even when the training data points are not linearly separable. Third, a mean squared error indicator can be computed extremely rapidly, which will enable interactive machine learning.

In Sonic Print, the dimension of the input space – the number of Bark frequency bands into which we divide the spectrum – is 24. To classify sounds into four classes, we work with a bidimensional output space. We label the examples corresponding to four sound classes with the respective output values (-1, -1), (-1, 1), (1, -1) and (1, 1).

Applying a multivariate regression model to our data means looking for β such that ε is minimized in $Y = X\beta + \varepsilon$. Y is the output vector holding the classification labels, of dimension $n \times 2$, with n the number of samples in our training set. X is the input vector of dimension $n \times (24 + 1)$. The dimension of β is 25×2. The dimension of the error matrix ε is $n \times 2$. When we use the ordinary least square estimator, the training algorithm is reduced to the simple matrix calculation: $\beta = (X'X)^{-1}X'Y$.

The mean squared error gives an indication of the quality of learning: $\varepsilon = Y - X\beta$. After training, the regression estimate is available with $\hat{Y} = \beta\hat{X}$, where $\hat{X}$ is the data to classify, i.e. one vector of Bark values. The classification result is given by the point closest to $\hat{Y}$ among the four training outputs from (-1, -1) to (1, 1).

3.3 Implementation

It is important to make Sonic Print an integral part of musicians' creative environment. Hence, we implemented the tool within Max, a multimedia programming environment widely used by sound artists and well suited to both live spectral sound processing and matrix processing (Charles, 2008). In addition, to make the tool available to musicians who do not program new media interactions, we created a Sonic Print Max for Live device. Artists can integrate this version in the popular Digital Audio Workstation Ableton Suite, which requires no programming knowledge.

We take advantage of Max built-in libraries to compute the Bark values: for each window of samples, we first apply a Hanning tapering window, then compute the Fast Fourier Transform (FFT). With both FFT and hop sizes equal to 512 samples, and a sampling rate of 44.1 kHz, each second of audio is represented by 86 vectors of data. Each vector holds the amplitude for 256 frequency bins linearly spaced from 0 to 22.05 kHz. We process these 256 values through a set of 24 filters to get the Bark values.

3.4 User Interaction

To train the Sonic Print models, the user needs to play a sound with a particular timbre while pressing a *Learn* button (see Figure 1a). As soon as the button is pressed, a one-second audio excerpt is recorded, processed into a set of vectors of Bark values, and labelled with a desired output vector (depending on the class choice from 1 to 4). The data are immediately concatenated with the data from the other classes, then processed to update the regression hyperplane defined by β.

As soon as training is over, the device resumes its classification task: live audio is converted to Bark descriptors, which are fed to the regressor. A classification is deduced, and the live sound is routed to one of four possible audio outputs, or tracks in the Max for Live version. These different outputs may be populated with four different sound processing algorithms (see below Trio: *Fracture/Morphosis*). The Max version

may be used to trigger specific events upon classification (see below Timbre Centric Composition: *Siete Dolores*).

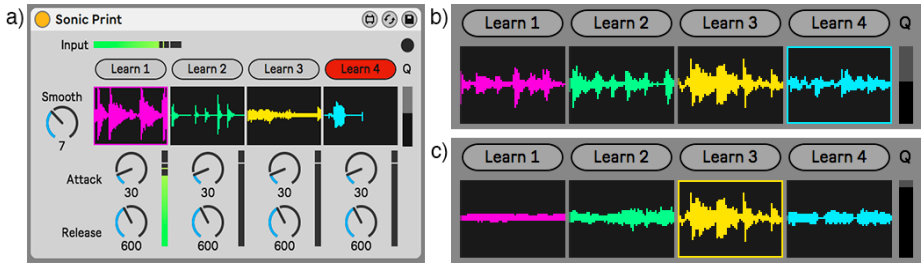

Fig. 1. Screenshots of the Sonic Print Max for Live device. **a)** The interface features a Learn button for each sound class. Visual feedback includes recorded waveform, learning quality Q, and a frame around the live classification result. The Smooth parameter controls a low pass filter to slow down variations in the live classification output. In **b)**, the quality of learning Q is low: very similar samples have been assigned to different classes; the discrimination is impossible and the squared error is high. In **c)**, different samples have been assigned to the different classes, they belong to different sub-spaces; the low squared error is displayed as a high quality of learning Q.

The quality of learning is presented to the user as a graphical representation of the mean squared error. This enables interactive machine learning. When the user finds the indicator to be too low, they can easily re-train the system (see Figure 1, b & c).

4 Applications

Sonic Print was made available to several artists with different backgrounds and aesthetic approaches. The artists could use the tool after a short demonstration. We present here two new works using Sonic Print: a piece for a trio of electro-acoustic musicians, and a piece for solo saxophone and audio-reactive multimedia.

4.1 Trio: Fracture/Morphosis

Fracture/Morphosis is a work for trombone, violin, and double bass with live electronics, by Joseph Norman. It is featured on his album *Dysphonia* (Norman, 2018).

Each musician performs with their instrument and a laptop processing their own sound. Sonic Print acts as a gate to route the musician's sound to one of four processing lines, depending on the resemblance of the timbre to one of four learned classes. The composer designed the live audio processes to give each of the four individual sounds a unique processed identity. The audio effects include variable delay units, four-channel frequency shifters, and a grain delay (see Figure 2).

The output of the processing lines for each instrument is routed to a surround spatialization system common to all three instruments. The delay units, the grain delay, and the spatialization make use of unique chaotic controls akin to low frequency oscillators. They provide constantly morphing delay time, grain size, grain density, pitch range, rate of change, and position within the speaker-delimited space.

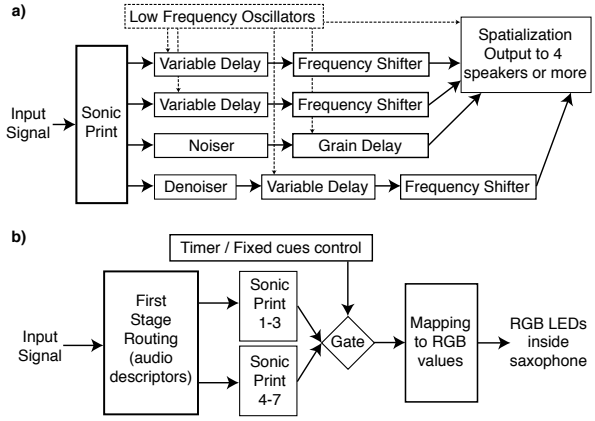

Fig. 2. Signal path in **a)** *Fracture/Morphosis* (shown for one of the three instruments), and **b)** *Siete Dolores*. Each composer designed their system independently.

In *Fracture/Morphosis*, each instrumentalist is provided with four distinct initial musical gestures that are later diffused in a modular and probabilistic performance format. The learning step of Sonic Print is built into the piece as an introduction during which each instrument's gesture is performed without audio processing (see the double bass gestures in Figure 3).

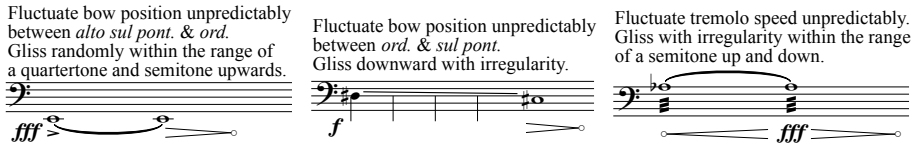

Fig. 3. *Fracture/Morphosis*: an excerpt from the double bass part showing three of the four sonic textures used to train Sonic Print. These sounds are played during the first part of the work.

After the introduction, the musical gestures are altered over the course of the piece, so that the distinctions between them become more and more ambiguous. The sonic identity of the pairing between musical gesture and specific effect gradually dissolves. The variability of the instrumentalists' individual choices allows for Sonic Print to react heuristically to incoming audio signals. While the piece is formally stable in a statistical sense, the individual sound materials, both acoustic and processed, are unique to each performance experience.

Sonic Print provided structural and material inspiration to the conception of *Fracture/Morphosis*. Norman chose to build on the constraint of four distinct initial timbres, using this limited reservoir as a framework for multifaceted output. Although the performers received no training in machine learning concepts, they could use the system very quickly in rehearsal. Thanks to the intuitive interface, they could focus their energy on the quality of their musical interpretation.

4.2 Timbre Centric Composition: Siete Dolores

Siete Dolores (2019), for tenor saxophone and electronics, is the first composition from an ongoing collaborative project between composer Gil Dori and saxophonist Jonathan Chazan. Dori and Chazan explore elements of a musical composition (pitch, dynamics, material development, formal unity, etc.) through the prism of timbre, with the prospective goal of developing a method for composers to indicate specific timbres, and for performers to interpret them precisely.

Chazan, an experienced performer of both early and contemporary music, observed that different musical styles embed in their performance practice different sound production modes, each resulting in a slightly different timbre. Dori explores in *Siete Dolores* the transition between seven such states. He separates timbre production from other elements of sound, to reconstruct it in ways that are not necessarily natural or intuitive, but can result in intriguing and unpredicted sonic outcomes. Throughout the piece, motives morph through changes in timbral quality, and slowly reveal more of themselves (see Figure 4). This unfolding of musical material invites the listener to direct their attention to the manner in which timbres drive the composition.

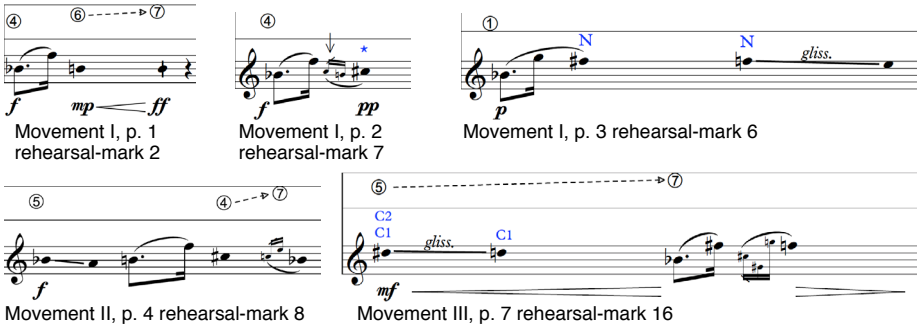

Fig. 4. Excerpts from the *Siete Dolores* score. The circled digits indicate which sound production state the performer should use, from 1 to 7. The successive excerpts show how a motive starts morphing over time.

The work uses Sonic Print to identify saxophone timbres, and map them to LED lights inserted into a second saxophone placed on stage. The LED colors provide visual signals that aim to support guiding the audience through the timbral experience. Since there are seven timbres to identify, the patch combines two Sonic Print units. To route the sound to the first or second unit, Dori uses *zsa.spread~*, which computes the spread of spectral data around the spectral centroid (Malt & Jourdan, 2009). When testing Sonic Print as an offline tool with pre-recorded sound files, the results were predictable and accurate. However, in a live setting, the system could not accurately identify subtle changes in timbre. Thus, fixed cues were used in some parts (see Figure 2).

Siete Dolores constitutes an important milestone towards a timbre-centric composition method. It has been successful in displaying this notion through several live performances since April 2019 (Dori & Chazan, 2019). However, more work is required to refine this method. It is important to find a more conducive way to incorporate Sonic Print as a compositional tool, and to improve the technology of identifying small-scale variations in timbre in real-time. Such progress will help to quantify timbres better, and to increase the quality and accuracy of communicating timbral ideas and executing them.

5 Discussion

5.1 Key Findings

The Sonic Print software was made available to artists, who used it successfully for the creation of new musical compositions. The experimentation confirmed that multivariate linear regression offers advantages including the certainty of a short, finite training time. The learning process took typically less than 15 ms on a 2015 laptop.

Artists have used the tool with no training in machine learning algorithms. The seamless integration of the software in the Max development environment proved crucial for the artists to actually experiment with its possibilities. Moreover, since the code is native to Max, without need for an external library, it worked through several generations from Max 6 to Max 8, on both Windows and Mac OSX systems.

5.2 Limitations & Future Work

Given the low number of learning examples, the model learnt by Sonic Print is necessarily overfitted. This is counter-balanced by the fact that the system always chooses an output class, even when the smallest distance between a live sample and a learnt class is relatively high. To mitigate overfitting, we could work on the size of the learning set: on the one hand, we could increase the length of the sampled sounds; on the other hand, we could design an iterative learning scheme, allowing new samples to be added to an existing set. A "no match" result could also be useful; alternative designs could include thresholding to confirm a class match, or the output of a continuous value representing the distance from the analyzed vector to the closest class.

In this version of the tool, the number of classes is fixed to four. To fit applications such as Dori's work, it is necessary to allow for more flexibility and a greater number of classes. It is also important to work further on the feature selection, starting with the exploration of two variations on the current system. First, we could use more than 24 filters – for instance, (Ishibashi et al., 2020) and (Laput et al., 2018) use 64 Mel-spaced values; second, we could use coefficients spaced according to the Equivalent Rectangular Bandwidth scale, a more recent variation on the Mel and Bark scales.

These options need to be evaluated in musical situations, to see how musicians use the tool in their creative practice.

6 Conclusion

Sonic Print successfully allowed several composers and performers to use automatic timbre recognition and live learning in new musical creations.

The unique contributions to the field include the seamless integration of a live timbre training and recognition algorithm in Max and Max for Live, environments commonly used by creative musicians. Moreover, the linear model applied to multidimensional spectral features enables both a streamlined interface and a very short, finite, learning time. With these features, the tool can be used intuitively by musicians with no previous knowledge about machine learning algorithms.

References

Bevilacqua, F., Müller, R., Schnell N. (2005). MnM: a Max/MSP mapping toolbox. *Proceedings of the International Conference on New Interfaces for Musical Expression*.

Brent, W. (2009). Perceptually based pitch scales in cepstral techniques for percussive timbre identification. *Proceedings of the International Computer Music Conference*.

Chachada, S., Kuo, C.-C. J. (2014). Environmental sound recognition: a survey. *APSIPA Transactions on Signal and Information Processing*, 3. http://dx.doi.org/10.1017/ATSIP.2014.12

Charles, J.-F. (2008). A tutorial on spectral sound processing using Max/MSP and Jitter. *Computer Music Journal, 32*(3), 87-102. http://dx.doi.org/10.1162/comj.2008.32.3.87

Chougule, S. V., Chavan, M. S. (2013). Comparison of frequency-warped filter banks in relation to robust features for speaker identification. *Recent Advances in Electrical Engineering*.

Costa, Y. M., Oliveira, L. S., Koerich, A. L., Gouyon, F., Martins, J. G. (2012). Music genre classification using LBP textural features. *Signal Processing, 92*(11), 2723-2737. http://dx.doi.org/10.1016/j.sigpro.2012.04.023

Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE transactions on electronic computers, 3*, 326-334. http://dx.doi.org/10.1109/PGEC.1965.264137

Dori, G., & Chazan, J. (2019). Siete Dolores excerpts [Video]. https://www.youtube.com/watch?v=sPBxBiqDGY0

Eamdeelerd, C. & Songwatana, K. (2008). Audio noise classification using bark scale features and k-nn technique. *2008 International Symposium on Communications and Information Technologies*, 131-134. http://dx.doi.org/10.1109/ISCIT.2008.4700168

Fiebrink, R., Trueman, D., Cook., P. R. (2009). A meta-instrument for interactive, on-the-fly machine learning. *Proceedings of the International Conference on New Interfaces for Musical Expression*.

Fourer, D., Rouas, J.-L., Hanna, P., Robine, M. (2014). Automatic timbre classification of ethnomusicological audio recordings. *Proceedings of the International Society for Music Information Retrieval Conference*.

Gardner, E. (1988). The space of interactions in neural network models. *Journal of physics A: Mathematical and general, 21*(1), 257-270. http://dx.doi.org/10.1088/0305-4470/21/1/030

Ghule, G., & Mukherji, P. (2016). A novel approach for Marathi numeral recognition using Bark scale and discrete sine transform method. *Conference on Advances in Signal Processing*, 191-195. http://dx.doi.org/10.1109/CASP.2016.7746163

Gillian, N., Knapp, R. B., O'Modhrain, S. (2011). A machine learning toolbox for musician computer interaction. *Proceedings of the International Conference on New Interfaces for Musical Expression*.

Grey, J. M. (1978). Timbre discrimination in musical patterns. *Journal of the Acoustical Society of America, 64*, 467-472. http://dx.doi.org/10.1121/1.382018

He, L., Lech, M., Maddage, N. C., Allen, N. B. (2011). Study of empirical mode decomposition and spectral analysis for stress and emotion classification in natural speech. *Biomedical Signal Processing and Control, 6*(2), 139-146. http://dx.doi.org/10.1016/j.bspc.2010.11.001

Hillenbrand, J. & Gayvert, R. T. (1993). Vowel classification based on fundamental frequency and formant frequencies. *Journal of Speech, Language, and Hearing Research, 36*(4), 694-700. http://dx.doi.org/10.1044/jshr.3604.694

Ishibashi, T., Nakao, Y., Sugano, Y. (2020). Investigating audio data visualization for interactive sound recognition. *Proceedings of the 25th International Conference on Intelligent User Interfaces*, 67-77. http://dx.doi.org/10.1145/3377325.3377483

Jathal, K. (2017). Real-time timbre classification for tabletop hand drumming. *Computer Music Journal, 41*(2), 38-51. http://dx.doi.org/10.1162/COMJ_a_00419

Laput, G., Ahuja, K., Goel, M., Harrison, C. (2018). Ubicoustics: plug-and-play acoustic activity recognition. *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, 213-224. http://dx.doi.org/10.1145/3242587.3242609

Lugger, M., & Yang, B. (2008). Cascaded emotion classification via psychological emotion dimensions using a large set of voice quality parameters. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 4945-4948. http://dx.doi.org/10.1109/ICASSP.2008.4518767

McAdams, S. (1999). Perspectives on the contribution of timbre to musical structure. *Computer Music Journal,23*(3), 85-102. http://dx.doi.org/10.1162/014892699559797

Malt, M., Jourdan, E. (2009). Real-time uses of low-level sound descriptors as event detection functions using the max/msp zsa. descriptors library. *Proceedings of the 12th Brazilian Symposium on Computer Music*.

Norman, J. (2018). Fracture/Morphosis. On *Dysphonia* [Digital album]. https://josephnorman.bandcamp.com/album/dysphonia

Park, T. H., & Cook, P. (2005). Radial/elliptical basis function neural networks for timbre classification. *Proceedings of the Journées d'Informatique Musicale*, Paris.

Sanaullah, M. & Gopalan, K. (2013). Deception detection in speech using bark band and perceptually significant energy features. *IEEE 56th International Midwest Symposium on Circuits and Systems*, 1212-1215. http://dx.doi.org/10.1109/MWSCAS.2013.6674872

Schedel, M., Perry, P., Fiebrink, R. (2011). Wekinating 000000Swan: using machine learning to create and control complex artistic systems. *Proceedings of the International Conference on New Interfaces for Musical Expression*, 453-456.

Sturm, B. L., Morvidone, M., Daudet, L. (2010). Musical instrument identification using multiscale mel-frequency cepstral coefficients. *Proceedings of the18th European Signal Processing Conference*, 477-481.

Zwicker, E., Flottorp, G., Stevens, S. (1957). Critical bandwidth in loudness summation. *Journal of the Acoustical Society of America, 29*, 548–557.