Identifying Metric Types with Optimized DFT and Autocorrelation Models

Matthew Chiu[1], Jason Yust[2]

[1]Eastman School of Music, Rochester NY, USA mchiu9@u.rochester.edu
[2]Boston University, Boston MA, USA jyust@bu.edu

Abstract. This paper explores the classification of metric types using different feature representations. Using weighted timepoint, DFT, and autocorrelation, we train feedforward neural networks to distinguish allemandes, courantes, sarabandes, and gavottes in the Yale-Classical Archives Corpus. Autocorrelation and DFT models perform better than a baseline, with DFT consistently better by a small amount.

Keywords: Discrete Fourier transform · Autocorrelation · Meter classification · Metric types · Neural networks.

Music theorists typically define meter as an abstract hierarchy, either as a hierarchical accent pattern on an underlying pulse stream [9] or a containment hierarchy of timespans [17]; see [5]. This is sufficient to represent musical time signatures, but musical practice also recognizes metrical types with the same time signatures and/or metrical hierarchies, for which these kinds of theories are therefore too abstract. Traditional dance meters of eighteenth-century Western Europe are a convenient example of this: Metric types which sometimes share metrical hierarchies are nonetheless distinguishable in practice. In this paper we explore the classification of allemandes, courantes, sarabandes, and gavottes using machine learning methods and three feature representations, a baseline weighted timepoint representation, autocorrelation, and discrete Fourier transform (DFT). Autocorrelation may be understood as an interval-based representation, while DFT is a periodicity-based representation.

After a review of both techniques and a discussion on the corpus preparation, we report on three feedforward neural networks models trained on data from the Yale Classical Archives Corpus [16] using the three representations and evaluated the models based on their ability to classify the four different baroque dance types.

1 Procedure

1.1 DFT

The DFT transfers a signal from the time domain to the frequency domain. With a discrete time-domain signal represented as a vector $X = (x_1, x_2, \dots)$ of length N, the DFT is a complex-valued vector, $F(X)$, defined by

$$F_k(X) = \sum_{n=0}^{N-1} a_j e^{-i2\pi kn/N} = \sum_{j=0}^{N-1} x_j(\cos(2\pi kj/N) + i\sin(2\pi kj/N)) \quad (1)$$

Each place in the DFT vector, k, represents a periodic function of period N/k. We are only interested in the size of each of these, so we take the norm of each component, $|F_k(X)|$ (eliminating the phase) and divide by $|F_0(X)|$ so that all values range from 0 to 1. The DFT is an orthogonal transform; each of the $F_k(X)$ for $1 \leq k \leq N/2$ is independent of the others.[1] Therefore, it partitions the total weight of the time-domain signal into all the possible frequencies dividing the fixed period N. When k divides N of these will coincide with traditional metrical periodicities. The DFT has been used for meter detection in audio signals [11, 7] and it has been used in music analysis to relate meter to form [4] and to describe rhythmic canons in Steve Reich's music [18, 19].

1.2 Autocorrelation

Autocorrelation is a correlation of a signal with itself at every possible lag value. It acts like a rhythmic interval vector, listing the weighted number of occurrences of each rhythmic interval (temporal distance between onsets). More precisely the autocorrelation is a vector $R(X)$ defined by

$$R_k(X) = \frac{1}{\sigma_X^2} \sum_{i=0}^{N-1} (x_i - \bar{x})(x_{i+k} - \bar{x}) \quad (2)$$

A number of studies demonstrate the use of autocorrelation to identify meter in symbolic (score or MIDI) data [3, 14, 15, 12] and audio [6].

Autocorrelation is closely related to the DFT. Specifically, it can be understood as *squaring the signal in the frequency domain* by appealing to the convolution theorem; see [2]. To make this precise, accounting for the normalization in Equation 2, define X' as the zero-mean version of X, i.e. $X' = X - \bar{x}$. This affects only the zeroth DFT coefficient. Then

$$R(X) = \frac{1}{\sigma_X^2 N} F(|F(X')|^2) \quad (3)$$

The function $(1/N)F(X)$ is the inverse Fourier transform, so this means that after removing the phase information autocorrelation returns the data to the time domain. The autocorrelation is therefore a vector of time intervals, whereas the DFT is a vector of frequencies, but otherwise contain essentially the same information.

[1] For $k > N/2$, $F_k(X)$ and $F_{N-k}(X)$ have equal magnitude and opposite phase for a real-valued signal, X, by the aliasing principle.

1.3 Corpus and Data Preparation

The Yale Classical Archives Corpus (YCAC) is comprised of "salami slices" of MIDI performance data [16]. A slice occurs everywhere that a new note is introduced, or a note ceases to sound. We isolated pieces in the YCAC by Bach with "allemande," "courante," "sarabande," or "gavotte" in the title. This procedure found 76 pieces, consisting of 90,688 pitch slices altogether (Table 1.3). In an attempt to emphasize newly introduced notes [13], notes that were contained in the immediately prior slice were removed. Sarabandes and courantes are in triple meters, usually 3/4, and allemande and gavottes are in duple meter, usually 4/4 and 2/2. Each dance style also has a corresponding rhythmic character: sarabandes accent beat 2, gavottes have long pick ups (starting halfway through a measure), allemandes have quick, sixteenth note pick-ups, and courantes frequently contain metric ambiguities.

Table 1. Corpus

	Allemande	Sarabande	Gavotte	Courante	Total
Piece count	24	22	8	22	76
Total slices	16,833	9,249	4,014	15,068	45,344
Ave. slices per piece	701	429	502	685	2,317
Length in ♩	4,228	2,924	1,366	4,981	13,468

1.4 Weighting, windowing, and training

Perceptual theories of meter finding (e.g. [9, 10]) often claim that a variety of musical features influence meter induction by imparting "phenomenal accent" to time points. We devised a relatively simple weighting scheme on slices based on three factors: the number of notes, duration, and bass notes. Recall that notes are removed if they occur in a preceding slice, so the first factor counts the number of new pitches introduced at a given time point. The duration factor represents the assumption that slices of longer duration will tend to have more metrical weight ("agogic accent"). We include a weighting parameter, δ, which we multiply by the duration of each slice in quarter notes. The register factor reflects the assumption that new bass notes will tend to be metrically weighted. We define a bass note as the lowest note within γ quarter notes before or after the given onset time. We add a constant, τ, for any slice where a bass note occurs. There are thus has three adjustable parameters, δ, γ, and τ. Figure 1 shows a sample score fragment with calculated weights for selected slices with (δ, γ, τ) set to $(1, 2, 3)$. During the training process we tuned the parameters to assess their impact.

We transformed the extracted chord slices to time-series data, encoding each score as an array dividing the full duration of the piece into 32nd notes. We placed the rhythmic weight for each slice in the time-series array at its corresponding

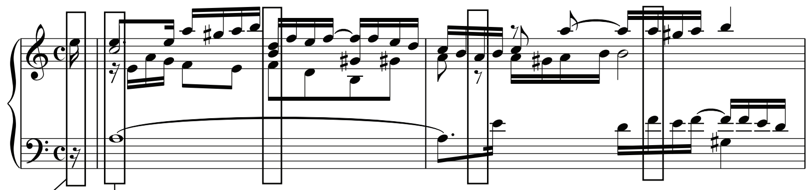

Fig. 1. Sample score with selected slices, showing the data structure and time-point weighting procedure with $(\delta, \gamma, \tau) = (1, 2, 3)$.

onset position, and put zeros elsewhere. Table 1.3 shows the number of quarter notes in each of the four metric types. To prevent wraparound for the DFT, windows were zero-padded with 96 additional 0s at the end of each vector.

For each metric type, we extracted 1,000 random windows, each 12 quarter notes in length. Our corpus thus consisted of 4,000 time-series windows and 4,000 corresponding labels identifying the correct metric type. We separated 950 samples of each type for training data, leaving 200 windows, 50 windows of each metric type, for evaluating the models. We fed the three inputs – baseline weights, autocorrelations, and DFTs – into the same neural network architecture: an input layer of 192, and 2 hidden layers of 30 and 10 neurons respectively each using relu activation [1]. We also used the Adam optimization algorithm [8]. The models were trained with 10 epochs on the training corpus (of 3,800 windows), repeated for each different tuning of weighting parameters (δ, γ, τ).

2 Results

Table 2 reports the evaluation scores as categorical accuracy, the percent of correct predictions based on the input. We found, as in [12], that adjusting the weighting parameters (δ, γ, τ) only alters predictions modestly and with no obvious trends. Even eliminating the duration weightings (δ) and bass note weightings (τ) entirely does not reduce performance, except in one case (eliminating both features in the autocorrelation condition). Therefore, in the DFT condition, where identification was the best, it appears to be based entirely on the basic rhythm and number of new pitches.

Excluding trials with 0-weighted features, the autocorrelation models ranged from 66%–77% with an average of 72% accuracy, and the DFT models ranged from 70%–80% with an average of 75% accuracy. Both models performed consistently around the same level with the DFT model modestly better. The control averaged 46% accuracy. The confusion matrix in Table 2 shows that gavottes were better classified than all other types, probably because there were fewer

Table 2. Categorical accuracy predictions

δ	1							0	1	2	3	4	5	6	0	1	2	3	4	5	6
τ	0	1	2	3	4	5	6	1							0	1	2	3	4	5	6
$\gamma = 2$																					
Control	.51	.47	.47	.47	.41	.40	.40	.49	.47	.48	.48	.51	.49	.48	.47	.47	.40	.40	.47	.44	.46
DFT	.80	.78	.80	.72	.68	.70	.73	.77	.78	.76	.79	.74	.76	.77	.79	.78	.75	.74	.75	.69	.73
Autocorr.	.75	.73	.71	.71	.70	.74	.75	.73	.73	.71	.73	.72	.72	.69	.44	.73	.66	.73	.69	.69	.70
$\gamma = 3$																					
Control	–	.49	.48	.44	.50	.47	.44	.50	.49	.56	.45	.50	.49	.44	–	.49	.52	.44	.47	.43	.40
DFT	–	.79	.75	.75	.76	.76	.74	.80	.79	.79	.75	.75	.80	.81	–	.79	.75	.74	.75	.74	.71
Autocorr.	–	.72	.73	.76	.76	.77	.76	.73	.72	.77	.73	.70	.71	.74	–	.72	.73	.72	.75	.74	.72

gavottes in the data set, so the classifier was more likely to be trained on excerpts from the same piece used in the test, and although these would not have been exactly the same window, they may have had similar traits.

Table 3. Confusion matrix for DFT/autocorrelation, all with $(\delta, \gamma, \tau) = (6, 3, 1)$.

	Allemande	Courante	Gavotte	Sarabande	Accuracy ($n = 50$)
Allemande	38/34	6/7	0/2	6/7	76%/68%
Courante	4/9	32/35	3/2	11/4	64%/70%
Gavotte	0/0	0/5	50/44	0/1	92%/92%
Sarabande	2/5	7/7	0/4	41/34	82%/68%

Figure 2 shows the average DFTs for $(\delta, \gamma, \tau) = (6, 3, 1)$ from which we can infer some of the differences that the classifier may have relied on to distinguish metric types.[2] The main differences are that allemandes are generally flat down to the sixteenth-note level, meaning that higher metrical levels were not very salient in these pieces. Higher metrical levels were better detected in sarabandes and gavottes, but differences relating to distinctions between duple and triple meter (in coefficients 6, 8, and 12) are weak at best. Therefore, the classifier is likely relying more on the salience of different metrical levels rather than differences between the duple and triple metrical hierarchies that would be emphasized in traditional metric theory.

References

1. Agarap, A.F.: Deep learning using rectified linear units (relu). ArXiv Preprint ArXiv:1803.08375 (2018).
2. Amiot, E.: Music in Fourier Space. Heidelberg: Springer (2017).
3. Brown, J.C. The determination of meter of musical scores by autocorrelation. Journal of the Acoustical Society of America **94**(4), 1953–1957.

[2] We include only even values here, because the zero padding produces distracting artifacts in the odd coefficients.

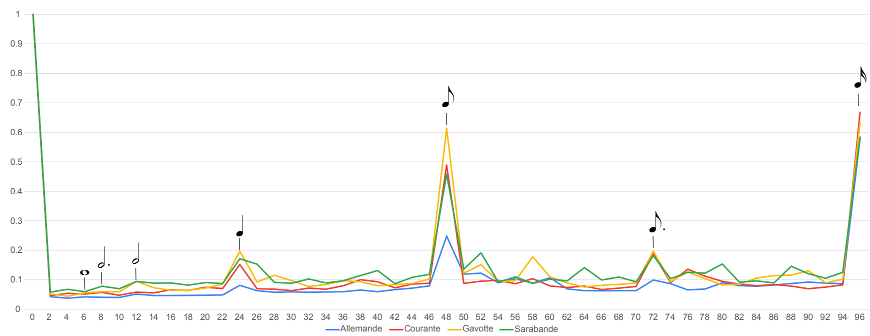

Fig. 2. Average DFT values for each metric type at $(\delta, \gamma, \tau) = (6, 3, 1)$.

4. Chiu, M.: Form as meter: Metric forms through Fourier space. Master's thesis, Boston University (2018).
5. Cohn, R.: Preferatory note: How music theorists model time. Journal of Music Theory **65**(1), 11–16 (2021).
6. Eck, D.: Identifying metrical and temporal structure with an autocorrelation phase matrix. Music Perception **24**(2), 167–176 (2006).
7. Flannick, J.E., Hall, R.W., Kelly, R.: Detecting meter in recorded music. In: Sarhangi, R., and Moody, R.V. (eds.) Renaissance Banff: Mathematics, Music, Art, Culture, pp. 195–202. Bridges Proceedings (2005).
8. Kingma, D.P., and Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).
9. Lerdahl, F. and Jackendoff, R.: A Generative Theory of Tonal Music. MIT Press (1983).
10. Povel, D.J., and Essens, P.: Perception of temporal patterns. Music Perception **2**(4), 411–440 (1985).
11. Sethares, W.A., and Staley, T.W.: Meter and periodicity in musical performance. Journal of New Music Research **22**(5), 149–158 (2001).
12. Eerola, T., and Toiviainen, P.: MIDI Toolbox: MATLAB Tools for Music Research. Finland: University of Jyväskylä (2004).
13. White, C.W.: Meter's influence on theoretical and corpus-derived harmonic grammars. Indiana Theory **35**(1–2), 93–116 (2018).
14. White, C.W.: Autocorrelation of pitch-event vectors in meter finding. In M. Montiel, F. Goméz-Martin, O.A. Agustín-Aquino (eds.) Mathematics and Computation in Music, Proceedings of the 7th International Conference, 287–296. Heidelberg, Springer (2019).
15. White, C.W.: Some observations on autocorrelated patterns within computational meter identification. Journal of Mathematics and Music **15**(2), 181–193 (2021).
16. White, C.W., and Quinn, I.: The Yale-classical archives corpus. Empirical Musicology Review **11**(1), 50–58 (2016).
17. Yust, J,: Organized Time: Rhythm, Tonality, and Form. Oxford University Press (2018).
18. Yust, J.: Steve Reich's Signature Rhythm and an Introduction to Rhythmic Qualities. Music Theory Spectrum **43**(1): 74–90 (2021).
19. Yust, J.: Periodicity-based descriptions of rhythms and Steve Reich's rhythmic style. Journal of Music Theory **65**(2) (2021).