TISMIR

Collins, N., et al. (2018). A New Curated Corpus of Historical Electronic Music: Collation, Data and Research Findings. *Transactions of the International Society for Music Information Retrieval,* 1(1), pp. 34–55. DOI: https://doi.org/10.5334/tismir.5

DATASET

A New Curated Corpus of Historical Electronic Music: Collation, Data and Research Findings

Nick Collins, Peter Manning and Simone Tarsitani

A corpus of 1878 recorded works of historic electronic music from 1950–1999 has been collated. This novel data set empowers chronological study of variation over time, and the answering of research questions based on associated annotated metadata, such as art music versus popular music or comparing female and male composers. We describe the challenges of building our new corpus, audio analysis over all the works in it carried out via the SuperCollider Music Information Retrieval code library, and results of tackling two example research questions. The article involves some discussion of the material, but also accompanies release of the data itself.

Keywords: electronic music; MIR data set; computational musicology

1. Introduction

Music information retrieval research often confines itself to popular music alone, but a wider perspective on music provides richer perspective on the challenges of music understanding by machine, and alternative repertoires are often those most of interest to musicologists. Although there have been multiple computational musicology studies using popular music databases analysing trends over time (Mauch et al. 2015; Percino et al. 2014; Zivic et al. 2013; Serrà et al. 2012), electronic music has not received much prior attention, especially towards its art music arm, but also concerning the experimental fringes of popular electronic music. Databases in MIR, such as the Million Song Dataset (Bertin-Mahieux et al. 2011) or AcousticBrainz (Porter et al. 2015), often contain works of popular electronic music, given the importance of such electronic production technology to current trends in music, but are not specifically curated for musicological purposes, nor generally inclusive of more experimental and art music work. Knees at al. (2015) released data sets of key and tempo annotated electronic dance music excerpts (2 minutes at a time) sourced from amateur producers on the BeatPort site, though the works themselves were not of historical prominence.

The UbuWeb art resource website, however, holds a corpus of historic electronic art music of 476 MP3 files of complete pieces, openly available online. This corpus was analysed by Collins (2015); it has many problems as a well rounded data set including strong gender imbalance, no representatives for certain historic years, and no attempt to bridge across the spectrum of electronic music from art to popular.

Durham University, GB

Corresponding author: Nick Collins (nick.collins@durham.ac.uk)

A new database of electronic music is presented in this article which is around four times the size of the UbuWeb corpus, and curated carefully by scholars of electronic music to enclose works of historical significance. The corpus and associated metadata is released to accompany this article, and whilst the original CD quality WAV files cannot be released for copyright reasons, we do supply feature extraction data for every piece, as per the stance of the Million Song Dataset (Bertin-Mahieux et al. 2011).

We proceed to consider the difficulties of preparing a corpus of works, the nature of the audio feature extraction, the contents of the data released including fields available within the metadata, and two example small scale studies across the dataset.

2. The Challenges Of Corpus Formation

A concept such as 'electronic music' is a site of ongoing definitional debate, motivated by new theories and historical discoveries, and where it pertains to continually released new music, a moving target. We restrict ourselves to the second half of the twentieth century to have at least a little objective distance, though many of the artists involved are still alive at the present moment. Whatever era we chose, we would still contend with selecting representative works and the central issue of identifying those most historically significant (London 2013). We draw upon electronic music textbooks on the topic (for instance, Manning 2013; Collins, Schedel and Wilson 2013) but still must acknowledge inevitable bias. The openness of the data release accompanying this article is one counterbalance to the infelicities of curation.

Representative electronic music would make central the musical concerns of new technology, especially the novel sound worlds opened up by electronic sound synthesis and

sound transformation. Nonetheless, no category is perfect, and acoustic sound sources necessarily appear within a host of historic recordings; it is inherently in the nature of musique concrète and related movements in sampling to feature such material, even if subsequently transformed. The human voice is a particular special case, and where electronic music techniques interface with popular song, a singer will often be recorded with fully electronic backing, or transitional or hybrid instrumentations appear. For example, within synth pop Gary Numan is an important UK populariser of the synthesizer, but his most famous works (from 1979) combine rock instrumentation with lead synthesizer parts.

Indeed, whether to include the vast territory of popular music alongside art music is a contention for some musicologists (Landy 2007), though we find a hard categorical boundary between art and popular problematic (and turn this into a research question below). We take an inclusive approach, bringing in many currents within electronic dance music and non-classical electronica.

The corpus contains 1878 works of uncompressed audio (around 100GB), totalling 582495 seconds or nearly 7 days worth of audio. A typical audio analysis run as detailed in the next section would take around 6 hours of compute time. The scope of the database means it took some time to compile from CD ripping to metadata entry, but is still insufficient for extensive coverage across many musical styles associated with a broad church approach to electronic music, and cannot hope for exhaustive coverage of any one artist, nor of all major movements (holdings of 1990s "bedroom studio" experimental electronica could certainly be expanded, for instance). Nonetheless, we have collated or are working on a number of additional side databases, and the database released here is also ripe for refactoring into modules. The release of the data itself makes transparent the current stable formation and makes the data set available to other researchers.

It is worth acknowledging certain issues with the selection of works. **Table 1** lists some key questions and our resolution of these.

Table 1: Corpus formation issues and their resolution.

Issue	Resolution
Should only studio pieces be included, or live performances?	Studio recordings were the mainstay of the corpus, though allowing that some pieces are the result of live performance in a studio.
Art music or popular music?	The decision was made to include examples from many styles and not try to impose a hard category boundary.
Representation of female composers	An effort was made to include such pioneers as Daphne Oram and Else Marie Pade; See below for more data.
Representation outside of standard Western canon	We attempted to include some composers outside of the typical European and North American spheres, but are limited by historical imbalance in access and critical coverage.
How electronic is electronic?	In some cases, iconic pieces of 'electronic music' involve part acoustic or standard rock instrumentation.
How obscure is permissible?	We try to include works that are discussed in critical texts or representative of trends in electronic music, even if not a mass market release or unambiguously acknowledged in histories.
How minimal can works be?	Some drone pieces should be represented. Eliane Radigue's *Transamorem – Transmortem* (1973) is a very subtly and slowly shifting 67 minute work. Its inclusion potentially distorts feature averages taken across works, but to ignore it is to ignore a rich territory of electronic drone music. Nonetheless, Raymond Scott's repetitive *Tic Toc* (1963, from Soothing Sounds for Baby Vol. 1) was not included, though other pieces from the same album were.
What source is best?	Audio CDs were the primary source, but there are still issues of original releases versus remastered CDs, edited or mixed versions. Resolution here is often pragmatic because of the availability of particular releases to purchase.
Are humour and kitsch acceptable?	Some representation is necessary to capture the breadth of musical life. For instance, selected pieces from Perrey & Kingsley's *The Out Sound From Way In! The Complete Vanguard Recordings* (1966–7) were included.
Exact date to attribute	In metadata we initially separated Year of Composition from Year of Release/First Performance but in practice the two were so often close together as to provide no real gain to the database, and too difficult to get data on for every piece. Year of Composition is the standard one used in this study. If a range of years of composition was provided, only the latest year was taken.
Complete works versus movements	Occasionally, works would appear split into movements (one movement per audio file). We would allow this, except where it became unmanageable for many very small segments, when we joined the snippets back together as one file.
Exact name to attribute	A producer (such as Tim Simenon) may hide behind an alias (such as Bomb the Bass). Well known aliases as attributed on releases themselves are used rather than producer names (we use Aphex Twin rather than Richard James, for instance).

Representation is a difficult topic, bringing in issues of positive discrimination for fairer social coverage, set against the danger of distorting the historical situation as it existed at the time (e.g., a predominance of male composers in 1950s studios, or earlier electronic dance music, for instance). Recent discoveries may not represent influence at the time; few electronic musicians knew in earlier decades of Halim el-Dabh's 1944 wire recorder piece as a precedent to musique concrète, for instance. Some pieces are obscure, but of great interest, and the underground art work shouldn't necessarily be dropped in favour of a mass market representative. Even the latter can involve issues of reach, in that it may have an uneven world impact across markets; for example, for the core Western territories, the case of European versus US hits.

We tried to actively include female pioneers, and noted three levels of involvement; male led (e.g., Gary Numan), female led (e.g., Björk), and a mixed group (for example the Art of Noise including Anne Dudley). Of the 1878 tracks in the corpus, 1523 were male only, 222 female only, and 133 mixed; women therefore led 13% of works, and 23% of works include female musicians. This is nowhere near equality, and reflects historical imbalance and curation choices. A smaller equal corpus can be formed from the larger unbalanced set, matching male and female led artists (see below); the whole corpus is at least open and omissions and bias can be interrogated by other researchers. The mixed group was often tricky to judge since many tracks, especially in electronic dance music, involve a sampled female vocal part added on, sometimes attributed as 'featuring' the singer, and sometimes anonymous (notoriously, Black Box's *Ride on Time* (1989) failed to acknowledge the extensively sampled Loleatta Holloway on first release; the song is in the corpus, and marked as mixed to reflect its complex provenance). For both male and female led works, predominantly in the popular sphere, sometimes an artist is labelled as solo, but there may be other producers in the background (for example, Salt-N-Pepa's first producer Herby Azor; we note this rap trio as female only though, reflecting the strong public perception of the group) or backing vocals from the other gender (as in Dr Alban's *It's My Life* (1992)) or other ambiguities (Outlander's *Vamp* (1991) samples a Yazoo track with Alison Moyet's vocal). We note that Björk in particular has had many issues in the past with gaining sufficient credit for her active production work against male producer collaborators such as Mark Bell or Matmos (Collins, Schedel and Wilson 2013). The three levels used in the database are an inadequate tool if making any deeper attempt to assess relative contributions of multiple participants.

As already apparent in the previous paragraph, electronic music's overlap with recorded music, where sampling allows any sound imaginable, is ripe for difficult categorisation, and creditation of works is challenged. For example, *Snow* (1963) by Daphne Oram is her studio treatment of an existing cover version, gradually sped up through a train's journey. *Can I Kick It?* (1990) is an iconic hip hop track which samples Lou Reed, though the artist name recorded for the database is A Tribe Called Quest. Hip hop selection was problematic in general, given the genre's origins in funk and disco instrumental backings, and sampling fixation; more electronic backings were preferred. An especially awkward decision was for *California Love* (1995) by 2Pac + Dr. Dre, kept for the talkbox vocals but not otherwise overtly electronic in timbre. Run-DMC's *Rock Box* (1984) was rejected as too electric guitar led (though other tracks from their eponymous debut album were included), Though the electric guitar is an electrified instrument, its associations with rock music rather than electronic music per se are awkward; nonetheless, Public Enemy's *Brothers Gonna Work It Out* (1989), which samples a Prince guitar solo, was kept (it is only part of the texture), as was New Order's *The Village* (1983) which has guitar and drums, albeit with a prominent sequenced synth. The timbral bias, of scratching and drum machines rather than electric guitar, is a cultural bias; sampling makes clear that there is however little really pure electronic music and many ultimately unresolvable inconsistencies in corpus inclusion (why exclude piano music but welcome rave piano riffs? Why exclude acoustic instrument and electronics mixed music pieces, but allow the human voice and electronics within popular song?).

Figures 1–4 show the coverage in the corpus year by year based on number of pieces, total duration of material in minutes, log(duration) per piece, and counts for female artist involvement.

That coverage is uneven year by year can be addressed by the selection of subsets with a more even balance (for example, choose N randomly for each year), though it will mean less variety overall. It is also possible to work with windows of a range of years at a time (as in Serrà et al. 2012), which overcomes reduced or absent examples for any single year, and ambiguity on the exact date of composition of a piece.

3. Feature Extraction Over The Corpus

Table 2 lists the audio features extracted across all works in the corpus; exact open source code definitions for all features are available via the project download page at composerprogrammer.com/emcorpus.html. Features were extracted at just over 43 frames per second (44100 Hz sampling rate, hop size 1024 samples, frame size 2048 samples). Windowed means, maximums, minimums and standard deviations were recorded (window size two seconds, hop size one second), as well as summary averages over the whole track. The features coincide with or derive from core machine listening features within SuperCollider, and many can be related to those detected in previous studies (compare for example **Table 1** in Collins 2015 which also uses SCMIR, though note feature order and type does differ). A more recent development is the use of source separation following FitzGerald's median separation algorithm for separating tonal and percussive components of an audio file (FitzGerald 2010).

We introduce two new features relating to spatiality of an audio file (all sources are stereo). Whilst the absolute loudness difference between the two ears is clear, a 'stereo spatial ebb' is defined as a joint spectral flux, as follows:

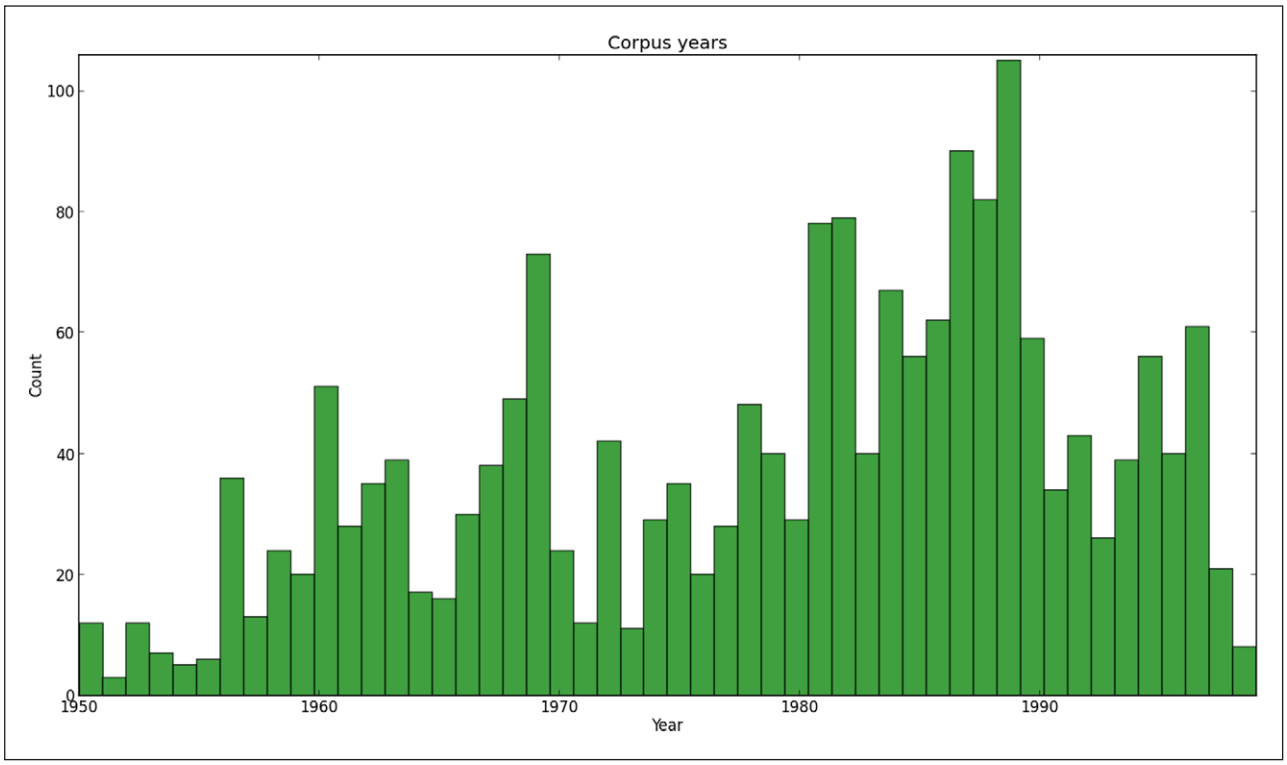

Figure 1: Number of pieces per year in the corpus, 1950 to 1999.

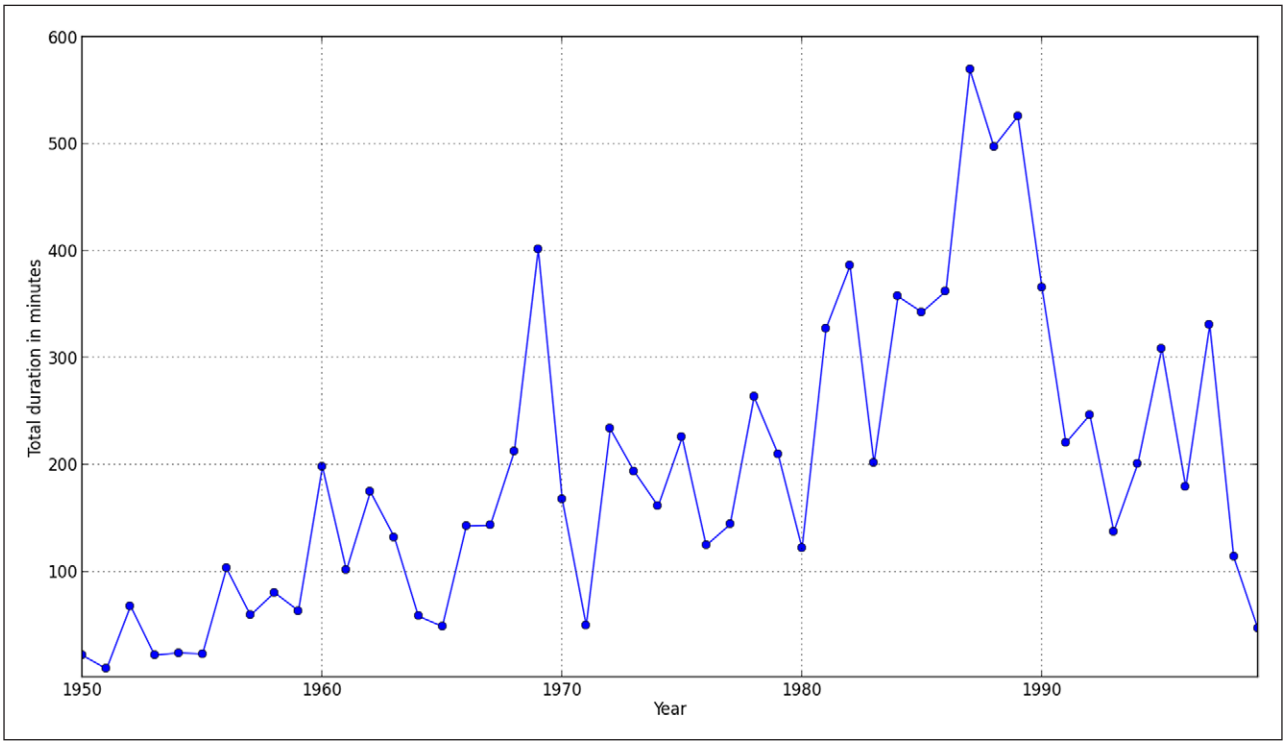

Figure 2: Total duration in minutes summing across all pieces from each year in the corpus.

$$SSE(\mathrm{t}) = \sum_{\text{over spectral bins } i} \left| \left(L(i,t+1) - L(i,t)\right) - \left(R(i,t+1) - R(i,t)\right) \right|$$

where L(i, t) is the i^{th} spectral bin at time t in the left ear, and R for the right. We carry out the calculation using ERB band bins corrected for equal loudness contours, rather than original FFT bins. The feature will have a high value only if a big change on the left has an opposite change on the right, or vice versa, that is, the condition of energy moving between the two channels, and hence a sense of spatial ebb.

Note that these features explore timbral and rhythmic aspects of the sound in the main, with no attempt to track and transcribe melody or harmony. As is expected for

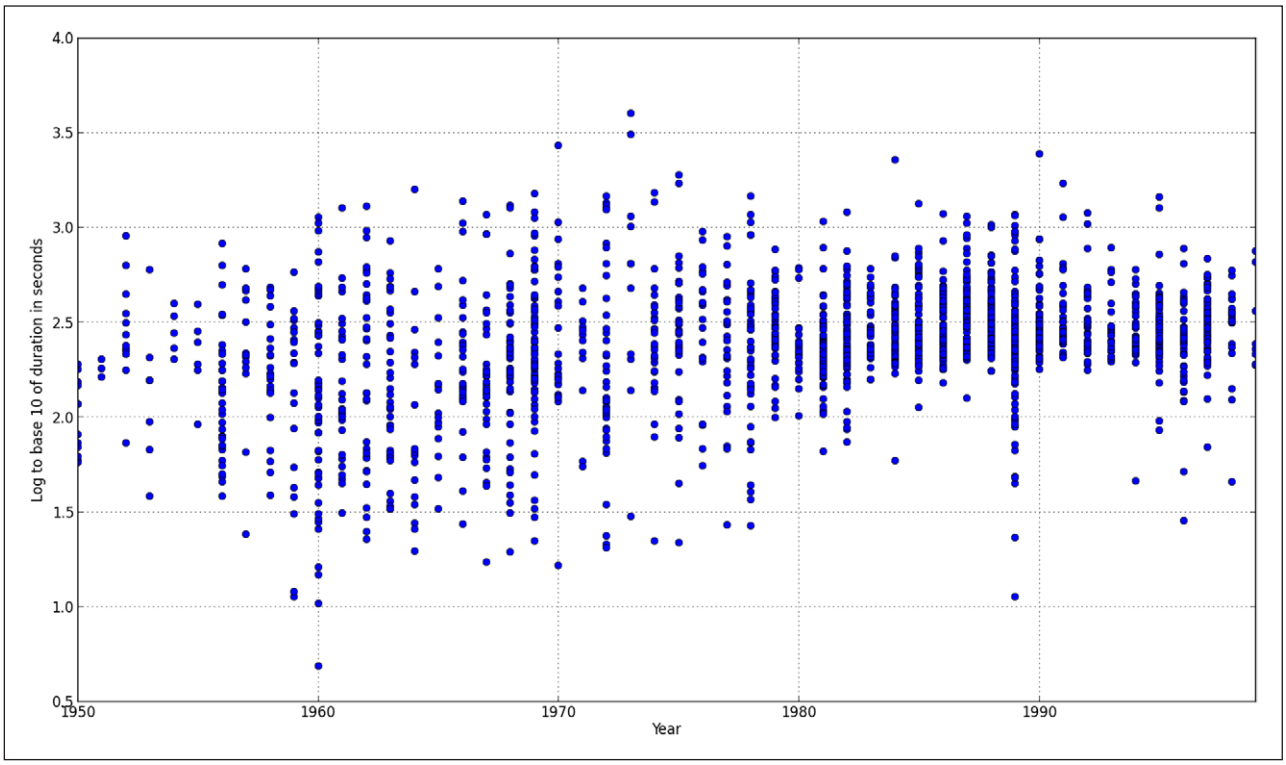

Figure 3: Log to base 10 of duration of each piece, plotted for its year of composition.

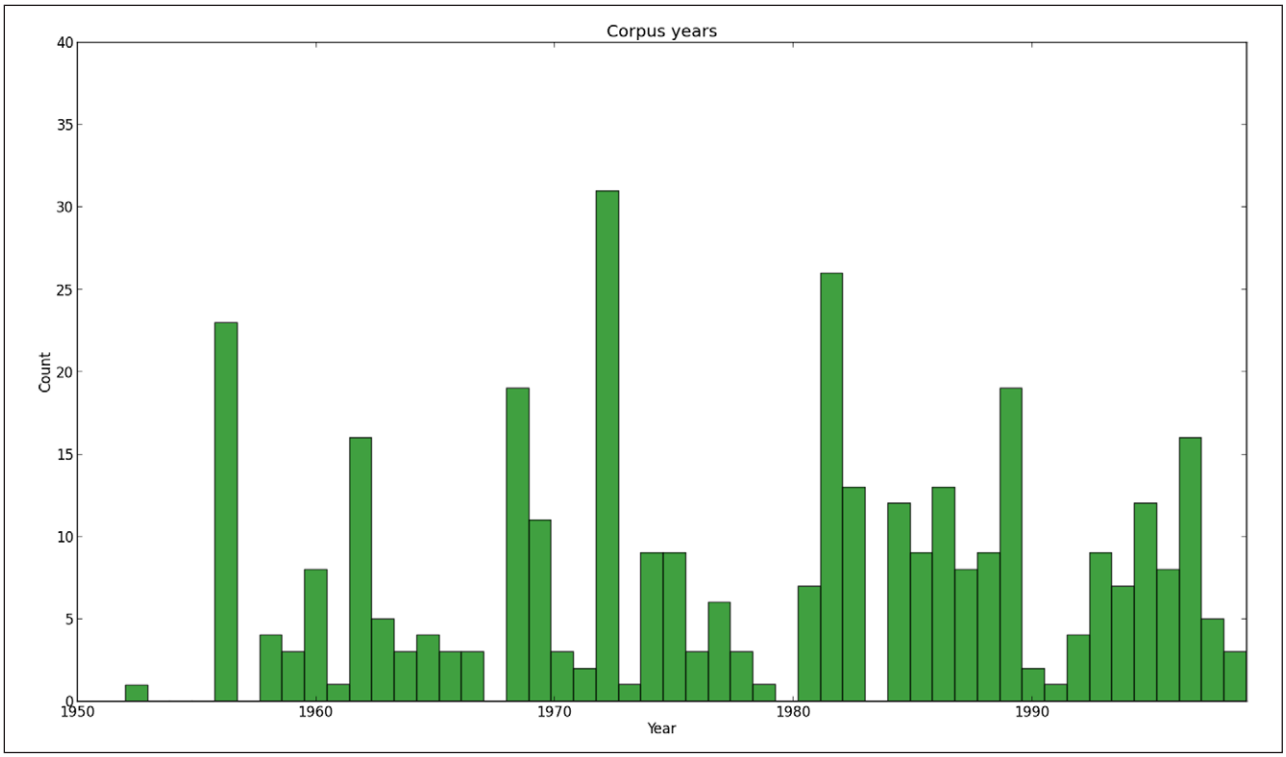

Figure 4: Incidence counts for music from female artists (either solo or within a group).

electronic art music in particular, for much of the music in the corpus, pitch is not the foremost parameter, though many pieces do contain conventional pitch material, especially in the popular music side of the works.

To illustrate the trends in feature values over time, **Table 3** lists linear regression (line fitting) results for average feature means against year of composition. The vast majority are significant at the $p < 0.01$ level in rejecting the null hypothesis of a zero gradient with no trend to the line (the values would also easily allow for Bonferroni correction, if applied). So there is definite evidence of changing features within electronic music over the five decades of the corpus. It is likely that the results are skewed by the lower quality recordings, and less popular music

Table 2: The 22 features extracted.

Feature number	Feature	Description
0	Loudness	Psychoacoustic model of loudness
1	Sensory dissonance	Psychoacoustic sensory dissonance model after Sethares (2005)
2	Spectral centroid	Measure of brightness
3	Attack slope	Average of the last ten attack slopes in the signal (with detection of attacks via an energy based onset detector)
4	Jensen-Shannon divergence	Compare the spectral distributions over ERB bands within the last two seconds; acts as a spectral change detector (so, similar spectral frames mean little divergence)
5	Transientness	Measure of transient energy in the signal, based on a wavelet transform
6–8	Onset statistics	In the last two seconds, the density (raw count) of attacks, and the mean and standard deviation of inter-onset intervals
9–12	Beat statistics	Beat histogram statistics; the entropy of the beat histogram, the ratio of the largest to the second largest entries in the beat histogram, the diversity (Simpson's D measure) of beat histogram, and metricity (consistency of high energy histogram entries to integer multiples or divisors of strongest entry)
13	Harmonicity	Root mean square amplitude (over 1024 sample windows) of tonal (harmonic) component of signal after median source separation
14	Percussiveness	Root mean square amplitude (over 1024 sample windows) of percussive component of signal after median source separation
15	Key clarity	Acting on the tonal part of the signal, the degree of presence of a clearcut major or minor key mode (note this does not assume the work has to be in 12TET, just that 'clarity' is an interesting attribute varying between pieces)
16	Spectral entropy	Spectral entropy of spectral distribution of tonal component of the signal.
17–19	3 Energy bands	Energy for low (400 Hz cutoff), mid (centred 3000 Hz), and high frequency (cutoff 6000 Hz) regions
20	Stereo spatial ebb	Spectral movement measure comparing left and right channels (see text)
21	Two channel loudness difference	Absolute difference in perceptual loudness between the left and right channels

materials, of the earlier decades. All slopes are relatively shallow, but this is due to normalisation (between 0.0 and 1.0) and subsequent averaging of feature values leading to relatively small ranges of feature variation.

The five with the largest absolute regression coefficient (and associated smallest p-values) are plotted, with y axis offset for ease of reading, in **Figure 5**. These lines demonstrate a slight drop in the complexity of rhythms in later years (due perhaps to the increased presence of electronic dance music's generally more predictable rhythms), and increased energy of signals in both tonal and percussive component, and in low and high frequency bands. Beat histogram entropy is only marginally significant at the 0.01 threshold and would fail to be significant after Bonferroni correction (0.01/22 = 0.0004545…); spectral entropy of tonals, and the two channel loudness difference, are not significant, showing two features with no consistent change over the decades of the corpus.

4. The Data Set Released

The original audio takes up 102.75GB, but cannot be distributed for copyright reasons; nonetheless, following the precedent of other MIR projects such as the Million Song Dataset (Bertin-Mahieux et al. 2011), extracted feature data can be released. Feature data totals 254MB in binary archive format, and 883MB as ASCII text files.

The data is downloadable from composerprogrammer.com/emcorpus.html with metadata within a tab-separated text file, one line per track, and per audio file feature data in both SuperCollider archive format (with code provided for reading the data into SuperCollider) and in ASCII text file format. The SuperCollider archive files open directly as FloatArrays in SuperCollider with 1878 entries, one per audio file. Each array entry is an array of 22 values (for summary values across a whole file), or 22* (ceiling integer of tracklength in seconds) values (for running windows of feature values over a piece, window size two seconds and hop size one second). The ASCII files are arranged similarly except that SuperCollider array entries per line are just tab separated columns with one string float value per column. So each ASCII file has 1878 lines, and either 22 columns or 22* (ceiling integer of tracklength in seconds) columns.

Following anonymous reviewer request, we also make additionally available 12 MFCCs and 12 per octave chroma features frame by frame (hop size 1024 audio samples at 44100 Hz sampling rate, around 43 Hz frame rate), accessible as a separate download at the web address

Table 3: Linear regression results for average feature means against year of composition (no Bonferroni correction on p-values).

Feature number	Feature	Slope (gradient)	Intercept	Regression coefficient	p-value
0	Loudness	0.00278	0.123	0.867	3.821e-16
1	Sensory dissonance	0.00049	–0.0002	0.841	2.172e-14
2	Spectral centroid	0.00146	0.105	0.719	4.092e-09
3	Attack slope	0.00046	0.009	0.689	3.240e-08
4	Jensen-Shannon divergence	–0.00038	0.036	–0.557	2.609e-05
5	Transientness	0.00104	0.013	0.857	1.904e-15
6	Attack density	0.00386	0.297	0.826	1.566e-13
7	Mean of IOIs	–0.00064	0.095	–0.795	5.658e-12
8	Standard deviation of IOIs	–0.00127	0.102	–0.921	2.716e-21
9	Beat histogram entropy	5.1e-05	0.993	0.394	0.005
10	Beat histogram ratio first to second largest entries	0.00051	0.356	0.885	1.603e-17
11	Beat histogram diversity	–1.3e-05	0.982	–0.633	8.283e-07
12	Metricity	0.00066	0.299	0.681	5.397e-08
13	Strength of tonal part	0.00267	0.047	0.895	1.888e-18
14	Percussiveness	0.00157	0.013	0.924	1.175e-21
15	Key clarity of tonal part	0.00118	0.404	0.744	5.983e-10
16	Spectral entropy of tonal part	2.3e-05	0.015	0.125	0.386
17	Low frequency energy	0.0022	0.034	0.901	4.818e-19
18	Mid frequency energy	0.00085	0.017	0.839	2.854e-14
19	High frequency energy	0.00111	–0.003	0.912	3.333e-20
20	Stereo spatial ebb	0.00055	0.017	0.668	1.186e-07
21	Two channel loudness difference	–4e-06	0.036	–0.004	0.978

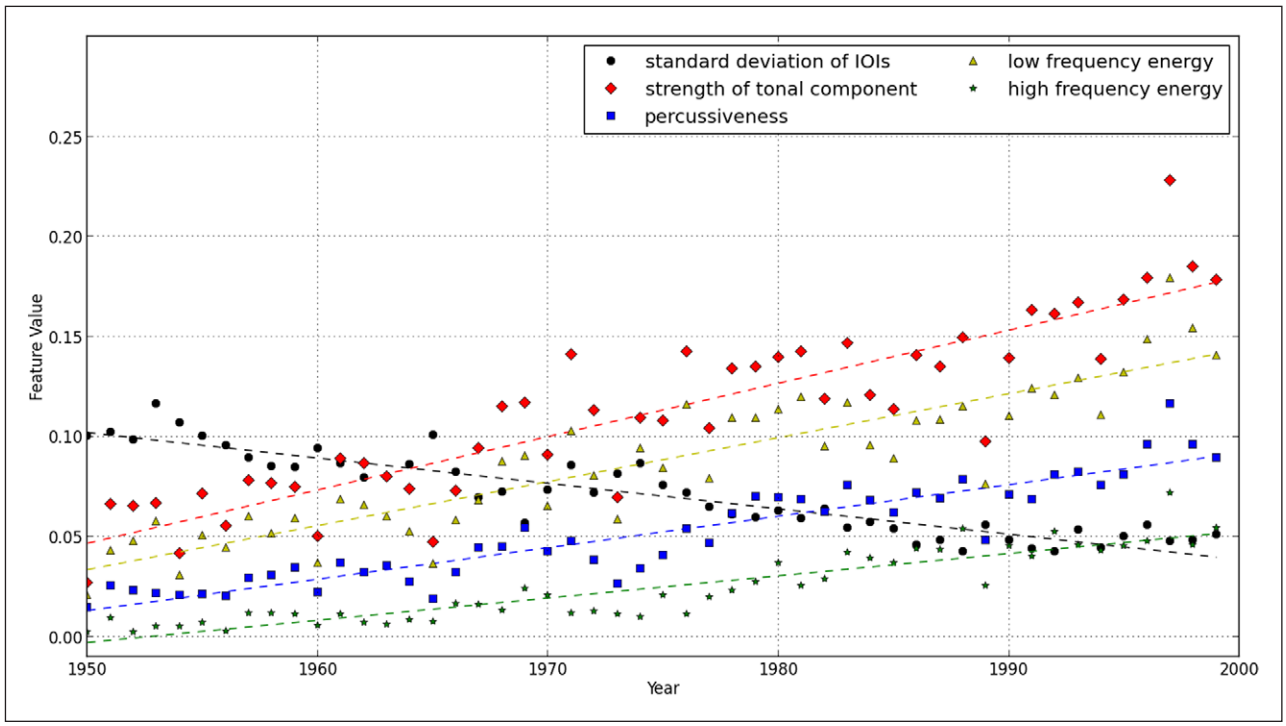

Figure 5: The five most significant feature mean trails against years of composition. The features are: 1) standard deviation of IOIs 2) strength of tonal component 3) percussiveness 4) low frequency energy 5) high frequency energy.

above (4.7GB, unzips to 10.5GB of data, one tab-separated 24 features per line ASCII text file per audio file).

The database fields supplied for each audio track are in **Table 4**.

To link the database to other resources, we hunted for MusicBrainz recording IDs using artist name and recording name for a slightly larger database of 2170 tracks (the 1878 track final database was a subset of this, having excluded certain pieces that were repetitions, or of inappropriate sound world as electronic music). A strict (exact) match was used wherever possible, available for 1655 pieces; we otherwise used the first search result from a non-strict match (in some cases, pieces had particular remix names, which were not picked up by MusicBrainz).

We checked for dates of release on MusicBrainz versus our hand-annotated data. Whilst MusicBrainz does have recording IDs for the majority of the database works, matching years were only found for 465 pieces if not strict, or 331 if strict only. Year of composition otherwise deviated on MusicBrainz, often due to a re-release or an alternative compilation CD where the piece could be found.

We also polled the multimillion track community database of analysed audio AcousticBrainz (Porter et al. 2015) for track matches from our database (this check requires a MusicBrainz recording ID as discovered above). Our corpus contained much material not available on AcousticBrainz, since only 812 out of 2170 works could be discovered. AcousticBrainz is likely to have more matches than this, due to the indirect search via MusicBrainz IDs, but there is support anyway for the assertion that there are corners of electronic music history that our corpus serves that are not being met by crowd-sourced work.

In all, 235 distinct CDs were ripped to create the corpus, with both collections of historic electronic music and sources from single artists. One anonymous reviewer pointed to the potential bias of including multiple tracks from one artist and from one album by a single artist. Given time resources for collection, and the lower count of available recordings for earlier decades of the chronology, the current corpus is not founded upon one CD source per track, nor on no more than one piece per artist. There are 576 distinct artists, and 74 out of the 235 CDs have no more than one track per artist name (e.g. they are collections across different artists rather than collecting multiple pieces by a single artist). Subsampling may be used to refine those pieces used by a researcher for a given project, the corpus metadata is entirely open to make criticism transparent, and future revised and expanded content may improve the disparity of source materials further.

5. Testing With The Corpus

The corpus had been annotated with some additional flags of direct relevance to particular research questions. We explore here the question of art music versus popular music, and go on to consider gender difference on a reduced matched subset of the corpus.

To assess the separability of art music and popular music, we attempted to discriminate the two using machine learning; if an algorithm can easily manage this task, it is strong evidence for a real distinction in the sound worlds, even if we know that a gross binary split has musicological and psychological issues. Corpus works were marked with a flag for popular (827), art (817), or borderline (234 works, e.g., Laurie Anderson). We ignored borderline

Table 4: Database fields.

Field	Detail
Popular vs art	Integer, 0 for popular music, 1 for art music, 2 for a borderline case
Gender	0 for male composed, 1 for female composed, 2 for a mixed group
Path name	Relative to a base directory for the project, with subfolders helping to group associated pieces
Artist	Artist name
Title of work	Title of work
Year of composition	Year first noted as completed as a composition (for a range, last year taken)
Year of release/first performance (if different to year of composition)	For some works in the corpus, the year of release or public performance is slightly different to the year of composition. Nonetheless, the year of composition is taken as the standard, and this data is not supplied for all works in the corpus
Source title	Title of audio CD from which music was ripped
Source record label	Record label of source audio CD
Source year of release	Source CD release; for example, some works were only available through remasters or re-releases with a more recent date than might otherwise be expected; historic works can only have appeared on CD from 1982
Source record label release code	Record label's own internal catalogue code for a release to help identify an exact source recording if necessary
Additional notes	Indicates, for instance, if there is slight mixing between works at the beginning or ending of a track (as appears in some electronic dance music releases in particular), or other matters
MusicBrainz recording ID	As found by an automated search based on strict match by artist name and track title, or where that failed, by closest match according to a non-strict inquiry to the MusicBrainz API
MusicBrainz strict match found	Flag to indicate those tracks where the recording ID was found through an exact artist and title match, and thus, most likely to be accurate

cases of art-popular crossover and concentrated on discriminating the remainder; selection by chance would perform at 50% here already. Test and training sets were formed by randomising the order of the combined art and popular tracks and splitting into two 822-track sets. Using a Naïve Bayes algorithm as a baseline learning algorithm and average feature vectors (22 features) per piece, we immediately achieved 728 out of 822 on the training set, and 727 out of 822 on the test set, rather similar results indicating strong generalisation. Whilst performance was quite good, examples of mis-labelled tracks included:

Bruce Haack, *Super Nova* (1969)
Fad Gadget, *Collapsing New People* (Berlin Mix) (1983)
Eliane Radigue, *Triptypch Part 3* (1978)

The first should perhaps have been labelled as a borderline piece in the first place, and incorporates elements of popular and art music. The second is an edgy synth pop work, with perhaps a few experimental timbral facets but more overlap with mainstream synth pop than electroacoustic art music, and the third a piece of electronic drone music whose slowing evolving beating sinusoids must have caused some confusion given machine listening assumptions, especially under averaging of feature values over the whole work.

We followed up this result by assessing all 22 features alone for Naïve Bayes discrimination. The top performing feature was the high frequency energy (689 out of 822 of the test set correct) and the second best percussiveness (677/822); the presence of drum parts are the likely main aural discriminating factor. The worst performing feature was the beat histogram entropy (430/822), operating around chance, which also performed poorly at spotting any trend over time in the corpus, and potentially indicates that the feature itself is somewhat divorced from human listening.

A greedy feature selection run (adding the best feature at each round) achieved a top discrimination score on the test set of 780/822, using the feature subset, in order of appearance, [19, 21, 4, 10, 20, 16, 15, 11, 8, 9, 12, 2, 6]. Note how the beat histogram entropy is utilised here as an earlier pick; it does seem to assist categorisation in combination with other features.

To explore gender, we created a reduced but matched corpus of male and female composer pieces, balanced by year and art/popular divide. For each year, we took as many art and popular pieces (not crossover art/popular or joint male and female group works) as were available separately from both male and female artists; the final corpus size was 185 tracks for each of male and female artists. We then attempted to discriminate the tracks on the basis of feature values; test and training sets were constructed by random selection of half of the corpus each. Using a greedy search for the best performing feature set, and a Naïve Bayes algorithm, the top performing classifier achieved 132/185 on the test set (71% success). Test and training attainment was similar; on average, rerunning random allocation of test and training set and greedy feature selection 100 times, the mean performance was 65%. This is statistically significantly different to chance; we also created a population of 100 random choice classifiers, averaging at 51% success. Both the samples of random and feature selection Naïve Bayes classifiers were normally distributed according to a Shapiro-Wilk test, and a t-test could reject the hypothesis that they arose from the same distribution (p value 1.5941e-59, t statistic 36.7668, degrees of freedom 99). So there is evidence of some difference between male and female composition work, at least as represented in the current corpus.

6. Conclusions

A new dataset of historical electronic music pieces has been released, through extracted feature data and some associated metadata. Whilst two research questions on the corpus were explored herein, many potential research questions remain. They may require in some cases additional annotation through the tracks of the corpus, and the corpus itself may well require further extension or auxiliary corpora (for instance, consolidating electronic dance music holdings and introducing more experimental electronica). Nonetheless, there is a baseline which we claim is much superior to the existing UbuWeb electronic music data set (Collins 2015).

Research questions for future investigation might include:

- How do works fall under unsupervised clustering not assuming any pre-existing genre stereotypes? Do similar years of composition cluster together?
- What ontologies of electronic music are supported by the corpus?
- If we hypothesise that the opening gesture of a piece is critical, at least for electroacoustic art music, can we categorise openings, and predict the importance of opening material with respect to whole pieces?
- To what extent does a particular piece X fit into the narrative of electronic music history? How well can you differentiate pieces X and Y against the backdrop of EM history? Is one more important an influence than another?

The corpus is also pliable for creative applications in new electronic music generation which respond to the past, extrapolate from past trends, or receive a baseline of training. The audio feature extraction can be effected live within SuperCollider to match up to the released feature data.

Acknowledgements

This research was funded by the Arts and Humanities Research Council under grant AH/L006820/1. Associated research data is released to accompany this article. With thanks also to three anonymous reviewers for their comments.

References

Bertin-Mahieux, T., Ellis, D. P. W., Whitman, B., & **Lamere, P.** (2011). The Million Song Dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)*. Miami, FL.

Collins, N. (2015). The UbuWeb Electronic Music Corpus: An MIR investigation of a historical database. *Organised Sound, 20*(1), 122–134. DOI: https://doi.org/10.1017/S1355771814000533

Collins, N., Schedel, M., & **Wilson, S.** (2013). *Electronic Music*. (Introductions to Music series) Cambridge: Cambridge University Press.

FitzGerald, D. (2010). "Harmonic/Percussive Separation using Median Filtering." *International Conference on Digital Audio Effects (DAFx)*. Graz, Austria.

Knees, P., Faraldo, Á., Herrera, P., Vogl, R., Böck, S., Hörschläger, F., & **Le Goff, M.** (2015). Two Data Sets for Tempo Estimation and Key Detection in Electronic Dance Music Annotated from User Corrections. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*. Málaga, Spain.

Landy, L. (2007). *Understanding the Art of Sound Organization*. Cambridge, MA: MIT Press.

London, J. (2013). Building a Representative Corpus of Classical Music. *Music Perception, 31*(1), 68–90. DOI: https://doi.org/10.1525/mp.2013.31.1.68

Manning, P. (2013). *Electronic and Computer Music* (4th edition). New York: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199746392.001.0001

Mauch, M., MacCallum, R. M., Levy, M., & **Leroi, A. M.** (2015). The Evolution of Popular Music: USA 1960–2010. http://arxiv.org/abs/1502.05417. DOI: https://doi.org/10.1098/rsos.150081

Percino, G., Klimek, P., & **Thurner, S.** (2014). Instrumentational Complexity of Music Genres and Why Simplicity Sells. *PLoS ONE, 9*(12), e115255. DOI: https://doi.org/10.1371/journal.pone.0115255

Porter, A., Bogdanov, D., Kaye, R., Tsukanov, R., & **Serra, X.** (2015). Acousticbrainz: a community platform for gathering music information obtained from audio. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*. Málaga, Spain.

Serrà, J., Corral, Á., Boguñá, M., Haro, M., & **Arcos, J. L.** (2012). "Measuring the evolution of contemporary western popular music." *Scientific reports 2* (Article number 521). DOI: https://doi.org/10.1038/srep00521

Sethares, W. A. (2005). *Tuning Timbre Spectrum Scale* (2nd Ed.). Berlin: Springer Verlag.

Zivic, P. H., Rodriguez, F. S., & **Cecchi, G. A.** (2013). "Perceptual basis of evolving Western musical styles." *Proceedings of the National Academy of Sciences, 110*(24), 10034–10038. DOI: https://doi.org/10.1073/pnas.1222336110

How to cite this article: Collins, N., Manning, P., & Tarsitani, S. (2018). A New Curated Corpus of Historical Electronic Music: Collation, Data and Research Findings. *Transactions of the International Society for Music Information Retrieval,* 1(1), pp. 34–55. DOI: https://doi.org/10.5334/tismir.5

Submitted: 20 September 2017 **Accepted:** 22 February 2018 **Published:** 04 September 2018

]u[*Transactions of the International Society for Music Information Retrieval* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS