MULTIDIMENSIONAL CROSS-CORRELATED DATASETS EXTRACTION FROM REFERENCE AUDIO FILES

Marco Cernuto

Dipartimento di Musica Elettronica,
Conservatorio Vincenzo Bellini,
Catania, Italia
marcocernuto@conservatoriocatania.it

Renato Messina

Dipartimento di Musica Elettronica,
Conservatorio Vincenzo Bellini,
Catania, Italia
renatomessina@conservatoriocatania.it

ABSTRACT

The purpose of this article is to describe a procedure for obtaining, from a reference sound, a multidimensional dataset representing the time-invarying frequency envelope and the cross-correlation between the individual frequency components of its spectrum. By using FFT spectral analysis, the magnitude of the frequency bands is sampled in irregular time series (unevenly sampled), suitable as model-based control for dynamic systems in sound design and auditory display applications. The article explains how to classify these relationships and use the resulting datasets repository.

1. INTRODUCTION

Generally, the variations in time and space of each element of a complex object, from biology to linguistics, are strictly constrained and interdependent. For example, when one finger moves, the others follow or oppose it. Similarly, the frequency bands of a complex sound are articulated according to shared trajectories and trends. We can classify this correlation by observing the acoustic features of the sound,[1] or, in a predictive way, by attributing specific levels of covariance to the spectral bands, depending on the sound source, or the purpose of use for which it was produced[1].

In order to numerically represent these relationships within the spectrum, the method proposed here performs the FFT on the reference audio file, by decomposing it into a moderate number of components, up to 10, representing a multivariate time series. The dynamics of this time-dependent variable data provides a dataset of amplitude envelopes, formally organized, which can be applied simultaneously as control functions to multiparametric audio synthesis or processing systems.

A significant feature of the datasets is that the functions representing the amplitude envelope of each band are obtained through the adaptive data smoothing technique UID (Unevenly Invariant Downsampling) [2], not based on windowing functions, bandwidth or averaging methods [3] but on a sample-by-sample analysis of the envelope.[2] This creates a unique and invariant vector segmentation of the reference file, facilitating the recognition of the coherence/equivalence between the reference model and its temporal or frequency manipulations.

A prototype has been developed in Max and can be downloaded at https://zenodo.org/records/11355613.

2. METHODOLOGY

The following methods describe how to derive datasets of meaningful control data for sound synthesis and manipulation algorithms.

2.1. Data selection

The choice of the sounds (which we will refer to with the term *clips*) depends on the intended use of the data that we will extract. Since the analysis algorithm, at this initial stage of our research, produces a time-invarying frequency envelope for the whole clip, it is more reliable to use clips of short duration, or basically with pitch-invariant spectra (or with small and repetitive variations, for example the texture of the wind, or a spoken voice). However, interesting results can also be obtained by using music clips, consisting in a sequence of chords, or notes. In this way, e.g., a diatonic spectral mixture, rather than the specific spectra of each note, can be produced. Therefore, the duration of the reference file can typically vary from a few seconds to about a minute. An important aspect of the research is that the methodology adopted aims to define, classify and identify a dataset depending on the immediate recognition of a known sound content, e.g. a classical music theme, thus simplifying a type of selection based on numeric or non-self-explanatory tags.

Music performances, with acoustic or electronic instruments, can provide a wide collection of cross-correlated expressive functions that can be applied, as happens for *grooves* [4], to overcoming the inexpressiveness of quantized control data.[3] In the classic repertoire it is possible to identify famous tunes whose derived datasets, although applied to systems of resynthesis or timbre manipulation, can maintain their recognizability and therefore their iconic value basically unchanged, with different gradations (see Figure 1). The mimetic dimension and the semantic meaning of the material used (i.e. its referential relationships with a specific source or context) can assume different levels or can be

[1] A general classification of these relationships is discussed in the last paragraph dedicated to the application criteria.

[2] Some explanation on UID is given in section 2.2.2.

[3] The term "quantization" here refers to the automatic mapping, e.g. in a score editor, of a large set of values (as pitches or velocities) to a smaller or fixed set of elements.

completely omitted through non-linear transfer functions applied to the processing of the spectral parameters [5].

Short-duration clips extracted from jingles or earcons can be used to standardize auditory display systems based on a formally heterogeneous set of signals, through derived models reproduced by homogeneous proprietary synthesis processes. Speech clips provide a wide range of information that can be used in an extension of prosody modeling techniques to abstract sound materials.

2.2. Sampling

The algorithm of the software engine developed in Max for the automatic generation of datasets from reference audio files is illustrated here.

2.2.1. Spectral analysis

The data preprocessing involves the extraction via FFT of the subset of bins from which to generate the amplitude envelopes for the dataset. It is developed according to the following steps.
1) Select the highest magnitude bins in each FFT frame.
2) Count the recursion of the bins selected in all frame.
3) List the bins sorted by score (these will be the bands for which to extract the amplitude envelope).
Calculating a class set of unchanged bins/pitches for the entire length of the analyzed sound is a specific purpose for which the algorithm is being developed: to generate a constant environment, valid even in the presence of spectral variations. Thus allowing to intentionally act on the sonification interface, depending on specific expressive needs; e.g. to emphasize an event through an harmonic reorganization or distortion.

To speed up the analysis operations, the FFT is carried out in offline mode, using the Max *jit.fft* opcode, active on the two planes of a matrix compiled with the complex signal to be analyzed.

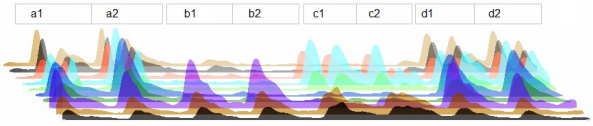

Figure 1: First 4 bars of Prelude N.1 in C major, Well Tempered Clavier, Vol.1, by J. S. Bach. We note the recursion of a binary form, characteristic of the compositional structure of the prelude, applicable to auditory display contexts where the reiteration of rhythmic patterns is required.

2.2.2. Dataset generation

The envelopes of the bands are isolated using a bank of bandpass filters with cutoff frequencies tuned to the frequency of the FFT bins and with a bandwidth equal to the fundamental frequency of the FFT, and then smoothed with a 16th order IIR lowpass filter.

The multidimensional dataset extracted from the spectral data is obtained through the non-uniform downsampling procedure called UID (Unevenly Invariant Downsampling), which we have already mentioned in the Introduction. It is based on the calculation of the dynamic variation in the envelope of a signal with a data smoothing method capable of detecting variations via an adaptive threshold comparator with hysteresis. Its main advantage is to convert an amplitude envelope into a series of intervals comparable to those of the traditional music notation (Fig. 2).

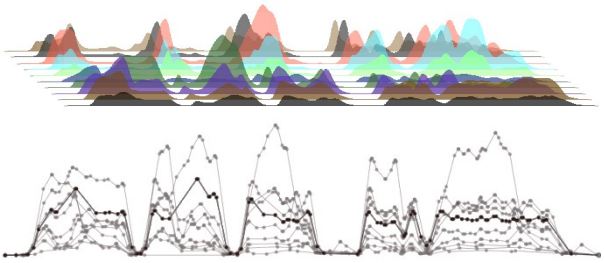

Figure 2: The superpositions of the 10 functions of the bands with the greatest magnitude contained from 172 Hz (highlighted function) to 2497 Hz of a spoken male voice sentence made up of two words and five distinct utterances.

2.3. Time series cross-correlation

The cross-correlation coefficient of the bands makes it possible to classify and index the datasets depending on a symmetry or covariance parameter, which can replace the simple taxonomic reference to its source in the selection of a sound clip.

For non-uniform time series (unevenly sampled), the correlation can be measured by different methods, e.g. with the Gaussian kernel correlation. Here we present the Pearson coefficients [6] of the series belonging to two consecutive bins, measured between the distribution of points over time and between variations in amplitude.

Figure 3 shows the Pearson correlation coefficient between the first and second functions (172 Hz and 258 Hz) represented in Figure 2. The analyzed clip (spoken voice lasting approximately 3 seconds) shows a very high data correlation on the time axis, R = 0.99; while on the amplitude axis the correlation is weak, R = 0.33.

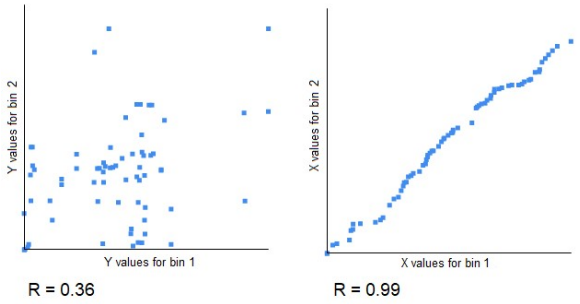

Figure 3: Pearson correlation indices for the distribution of data points on the x (time) and y (amplitude) axes. The high temporal correlation highlights the synchronicity of the bands in speech; the low correlation on the amplitudes demonstrates the dynamic variations between the functions.

The correlation data demonstrate that using these functions in sonification contexts can allow for an excellent level of synchronization (phase correlation) and a high dynamic/expressive diversification between the functions (amplitude correlation).

2.4. Replication

The pattern extraction from prerecorded audio material allows sonification systems to replicate the expressive structure of the reference file, inheriting, in addition to the formal structure, the dynamic and expressive variations. The time series obtained from the band analysis can share the same trend or present partially divergent characteristics. In the use of spoken vocal samples, prosody, although capable of expressively diversifying the spectrum, keeps the phases of speech articulation and segmentation into utterance unchanged. Conversely, the analysis of polyphonic clips can present interesting divergent patterns and complex metric and polyrhythmic characteristics.

3. APPLICATION CRITERIA

The application criteria of the exposed method depend on the selection of the reference audio samples and their mapping in the analysis and resynthesis stages. Both depend, in turn, on the need to keep the starting material recognizable, from a frequency or dynamic point of view.

In audification context (direct translation of data into sounds [7]) oriented to the faithful replica of a spectrum using external sources, it is not fruitful to use the amplitude data to control other parameters, e.g. pitch or density, as this breaks the expressive balance of the spectrum. Instead, in a free creative process, the relationships inherited from the analyzed file do not appear to be binding, but rather the interdependence between the functions, in terms of formal correlation or divergence. The referential relationships are fundamentally oriented towards a sonification based on physical models and on the possibility of recognizing these models, even inadvertently, as belonging to the same expressive category. Relationships in which a referentiality is not required are oriented towards serial data manipulations, typical of spectral composition [8].

Although not systematic, some application methodologies are suggested.

3.1. Primitive

This is the most basic transcription system of a spectrum, in which the frequency bands magnitudes control the corresponding frequency amplitudes of the sounding algorithm (this is a well-known vocoder effect). We are not concerned here with the match of the frequencies of the controller and the controlled sound environment, but rather with the reproduction of the balance of the frequency bands, understood as a pitch class. So, we can even use a VST virtual instrument with its own complex spectrum superimposed on the analyzed one.

The higher the temporal correlation, the easier it becomes to identify the processing product as derived from a recognizable source. Indeed, the phase shift of the bands (i.e. the temporal decorrelation of the functions) makes the onsets unclear and decreases the recognizability of the sound object content. The amplitude decorrelation appears less important for an assessment of referentiality.

In this direct approach, the application criteria can be carried out directly by manipulating the spectral parameters, within the same analysis/resynthesis system. Some examples follow.

a) The frequency bins magnitudes control the pitches of the sounding algorithm. The pitch increases proportionally with the magnitude of the spectral band. It may be considered an audification case. Rescaling is required and music serialism techniques are applicable, such as transposition or inversion.[9] In polyphonic scenarios, the movement of each voice follows a pattern inherited from the bands of the analyzed sound, and interesting phrases and imitation musical effects are generated.

b) The frequency bins magnitudes control the time parameters (duration, pause or delay) of the sounding algorithm. All the parameter controlled by the same dataset are synced by a common class of spectral bands.

c) The frequency bins magnitudes control heterogeneous parameters; e.g. the fundamental band magnitude controls the instrument articulation, and the 2th harmonic controls the pitch.

3.2. Derivative

Since one of the main aims of the research was to overcome the simplicity/obviousness of a direct application of spectral analysis to the sonification system, several attempts have been made in the search for novel techniques capable of giving the sonification system a sort of interpretative criterion, to create the feel of an interaction, free but at the same time with a perceivable correlation with the sonified sound object. [10]

From this perspective, a system in which the sonification environment is based on initially independent variables which are progressively modeled on the parameters acquired through spectral analysis, provided much more convincing results. In the audio example 5, we try to create an accompanying sound environment for a speaking voice.[4] We list below (table 1) the parameters used in the algorithm and their assignment to the corresponding spectral parameters of the dataset.

Sound accompaniment parameters	*Dataset parameters*
Interval of the envelope follower of the voice in the time domain	FFT window size
Pitch set	Frequency envelope
Delay time of the voices	Average interval between amplitude pitches
Velocity of the voices	Sequence of the amplitude points
Parameters jitter	Average distance between amplitude points

Table 1. Audio example 5. Sound accompaniment parameters linked to the voice spectral dataset.

[4] Examples of audio processing are available at: https://www.musicaelettronicabellini.it/icad2024.

Through the application of the spectral dataset, we try to convert a stochastic system into a model-controlled one, tuning its parameters depending on the spectral information.

3.3. Spectra classification

As mentioned in the Introduction, another interesting aspect is the search for a cataloging method of the spectra in relation to their cross-correlation. A simple classification can be traced back to a gradation of differences between phase/temporal correlation, from regular to irregular (so understood in terms of synchronicity) and amplitude correlation, in which the overlapping patterns, extrapolated from their original domain, draw reusable formal models for a constraints-free serial articulation. Their recursive structure suggests interesting applications, for example, in the design of warning signals and audible alarms.
Distinguishing the phase correlation from the amplitude correlation, we can also exemplify the following categories and items.
a) Low phase and amplitude correlation: chaotic sounds, earth nature and environmental sounds, polyphonic or polyrhythmic events, asynchronous synthesis techniques, nonlinear modulations.
b) High phase and amplitude correlation: spoken voice or other monophonic and monorhythmic sounds, periodic sounds produced by mechanical instruments, audio signals for encoded information.

4. CONCLUSIONS

The datasets obtained using the illustrated method will flow into a relational database, providing a useful and easy reference in the search for formal models already endowed with expressive characteristics culturally shared by a wide range of users. The figure 4 shows a portion of the dataset of a greeting phrase in Chinese by a female voice. The original audio file and a reworking with a piano VST are available at the audio example page (example 3).

An interesting future development of research aims at an inferential analysis through which to determine what are the levels of preservation of the semiotic structure and initial categorization of a sound, even in the presence of its formal alteration. Knowing the recognition relationships of a sound within a processing system can provide interesting insights into understanding the parametric gradients that contribute to the perception/identification/classification of a sound.

| female chinese spoken | 2.8 sec | 2048 fft size (48kHz sr) | sens 1. | x = time (normalized), y = amp (normalized)

score 14	Bin 8	210.94 Hz	x	0.000015	0.000378	0.007384	0.019287	0.074215	0.402332	0.434642
			y	0.000859	0.001001	0.016358	0.152459	0.082847	0.020347	0.156023
score 13	Bin 9	234.38 Hz	x	0.000015	0.000742	0.006686	0.020281	0.028585	0.041556	0.073428
			y	0.000692	0.001	0.014919	0.21569	0.208298	0.163446	0.097317
score 12	Bin 10	257.81 Hz	x	0.000015	0.001009	0.0061	0.020993	0.025134	0.034388	0.040265
			y	0.000605	0.001	0.012946	0.365799	0.363806	0.252561	0.234882
score 21	Bin 11	281.25 Hz	x	0.000015	0.00144	0.005937	0.026062	0.027464	0.05271	0.071685
			y	0.000537	0.001002	0.011387	0.803434	0.803421	0.195145	0.147553
score 14	Bin 12	304.69 Hz	x	0.000015	0.002041	0.006026	0.040235	0.041208	0.05553	0.070957
			y	0.000438	0.001002	0.010714	0.874939	0.875117	0.212622	0.181891
score 19	Bin 13	328.13 Hz	x	0.000015	0.002248	0.006315	0.02578	0.046751	0.047931	0.062112
			y	0.000406	0.001	0.011271	0.28974	0.663982	0.66423	0.287493
score 27	Bin 15	375 Hz	x	0.000015	0.001981	0.008252	0.03933	0.065933	0.06721	0.096722
			y	0.000454	0.001001	0.018736	0.169113	0.997904	0.997798	0.063919
score 30	Bin 16	398.44 Hz	x	0.000015	0.002182	0.045994	0.071974	0.073035	0.094548	0.181512
			y	0.000416	0.001003	0.149618	0.887989	0.888047	0.037588	0.149561
score 28	Bin 20	492.19 Hz	x	0.000015	0.003799	0.072434	0.130784	0.425529	0.445231	0.448808
			y	0.000201	0.001002	0.11005	0.031277	0.08208	0.338559	0.337718
score 21	Bin 21	515.63 Hz	x	0.000015	0.004185	0.073681	0.343329	0.347655	0.422049	0.443843
			y	0.000183	0.001002	0.079727	0.001	0.001	0.101202	0.512881

Figure 4: Portion of dataset obtained from a spoken voice sample and its hierarchical organization obtained through the recursion score of the spectral bands.

5. REFERENCES

[1] F. Eyben, “Acoustic features and modelling” in *Real-time Speech and Music Classification by Large Audio Feature Space Extraction*, Cham: Springer, pp. 9-122, 2016. https://doi.org/10.1007/978-3-319-27299-3.

[2] M. Cernuto, R. Messina, *Audio Envelope Music Notation by Unevenly Invariant Downsamplig* [Manuscript submitted for publication]. Department of Electronic Music, Music Conservatory of Catania, 2024.

[3] M.G. Christensen, “Dynamic Range Control”, in *Introduction to Audio Processing*, Cham: Springer, 2019. https://doi.org/10.1007/978-3-030-11781-8_11.

[4] Ableton Reference Manual, Chap. 13, “Using Grooves”, https://www.ableton.com/en/manual/using-grooves/.

[5] D. Smalley, “Spectromorphology: explaining sound-shapes”, in *Organised Sound* , vol. 2, no. 2, 1997, pp. 107-126. doi:10.1017/S1355771897009059.

[6] W. Kirch, “Pearson’s Correlation Coefficient” in W. Kirch (eds.): *Encyclopedia of Public Health*. Dordrecht: Springer, 2008, pp. 1090-1091. https://doi.org/10.1007/978-1-4020-5614-7_2569.

[7] F. Dombois, and G. Eckel: “Audification” in T. Hermann, A. Hunt, and J. Neuhoff (eds.): *The Sonification Handbook*, Berlin: Logos Verlag, 2011, pp. 302-326.
https://sonification.de/handbook/download/TheSonificationHandbook-chapter12.pdf.

[8] A. Cornicello, “Timbre and Structure in Tristan Murail's Désintégrations” in *Timbral Organization in Tristan Murail’s Désintégrations and Rituals by Anthony Cornicello*, PhD Thesis, Faculty of the Graduate School of Arts and Sciences, Brandeis University, Waltham, 2000.
https://www.anthonycornicello.com/dissertation/Chapter2.pdf.

[9] E. Krenek, “Extents and Limits of Serial Techniques”, in *The Musical Quarterly*, vol. 46, no. 2, 1960, pp. 210-232. https://doi.org/10.1093/mq/XLVI.2.210.

[10] N. Barrett, “Interactive Spatial Sonification of Multidimensional Data for Composition and Auditory Display”, *Computer Music Journal*, vol. 40, no. 2, 2016, pp. 47-69. https://doi.org/10.1162/COMJ_a_00358.