PRECISE PITCH CONTROL IN REAL TIME CORPUS-BASED CONCATENATIVE SYNTHESIS

Aaron Einbond
Center for Research in New Music (CeReNeM)
University of Huddersfield
A.M.Einbond@hud.ac.uk

Christopher Trapani
Computer Music Center (CMC)
Columbia University, New York
cmt2150@columbia.edu

Diemo Schwarz
Ircam-CNRS-STMS
Centre Pompidou, Paris
Diemo.Schwarz@ircam.fr

ABSTRACT

The need for fine-tuned microtonal pitch combined with the timbral richness of corpus-based concatenative synthesis has led to the development of a new tool for corpus-based pitch and loudness control in real time with CataRT. Drawing on recent research in feature modulation synthesis (FMS) as well as the bach library for Max/MSP, we have implemented a set of new modules for CataRT that permit the user to define microtonal harmonies graphically and combine them with other audio descriptors to trigger concatenative synthesis in real or deferred time. Pitch information is generated from a pitch analysis or extracted from soundfile meta-data, and loudness may be controlled independently for different sound sets. Musical implementations already suggest promising results as well as future goals to generalize this approach to further timbral features for corpus-based FMS.

1. INTRODUCTION

Composers have long used synthesis techniques permitting a high degree of pitch control, such as sampling or additive synthesis, to create electroacoustic music with fine-tuned microtonal harmony. The possibility of combining this finesse with the blossoming field of audio feature analysis presents promising potential for music that is both harmonically controlled and timbrally rich.

The idea of corpus-based transposition in CataRT was born out of this desire to work with real-time *Corpus-based concatenative synthesis* (CBCS) focusing on the pitch domain, with a precise control over the resultant pitch content of grains selected for playback. It became quickly evident that this type of real-time adjustment had ramifications for other musical parameters as well, so that any descriptor, such as spectral centroid, loudness, or noisiness, could possibly be modulated to a target value.

This approach takes advantage of three unique features of CBCS: Its multi-dimensional descriptor-based selection allows one to find automatically the best compromise between a desired timbral sound character and target pitch. Because it can handle very large corpora of sound, pitches close to the target are likely to be found. And finally, the off-line pre-analysis of the corpus generates a wealth of information about the sound units it comprises, combining automatic descriptor analysis and manually attributed meta-data, and allowing for informed and targeted transformations of the units.

2. PREVIOUS AND RELATED WORK

This approach draws on existing techniques for corpus-based concatenative synthesis and feature modulation synthesis.

2.1. Corpus-Based Concatenative Synthesis

The recent technique of corpus-based concatenative sound synthesis [12] builds up a database of prerecorded or live-recorded sound by segmenting it into *units*, usually of the size of a note, grain, phoneme, or beat, and analysing them for a number of sound descriptors, which describe their sonic characteristics. These descriptors are typically pitch, loudness, brilliance, noisiness, roughness, spectral shape, etc., or meta-data, like instrument class, phoneme label, etc., that are attributed to the units. These sound units are then stored in a database (the *corpus*). For synthesis, units are selected from the database that are closest to given *target* values for some of the descriptors, usually in the sense of minimizing a weighted Euclidean distance. The selected units are then concatenated and played, after possibly some transformations.

Corpus-based concatenative synthesis and related approaches are summarised in a survey [11] that is constantly kept up-to-date on-line.[1]

[1] http://imtr.ircam.fr/imtr/Corpus-Based_Sound_Synthesis_Survey

2.2. Feature Modulation Synthesis

The generalized technique of *Feature Modulation Synthesis* (FMS) [5, 6, 8] can be applied to change timbral features beyond pitch and loudness, the features traditionally treated by a sampler. FMS in general is an analysis–synthesis approach that can be regarded as a meta-synthesis technique, borrowing from various other synthesis techniques in order to modulate a certain feature of a sound. It consists in finding the precise sound transformation, and its parameters, that have to be applied to a given sound, in order to change its descriptor values to match given target descriptors. The difficulty is here that a transformation usually modifies several descriptors at once, e.g. pitch shifting by resampling changes the pitch *and* the spectral centroid and other descriptors. Recent approaches [8, 9] therefore try to find transformation algorithms that only change one descriptor at a time.

A data-driven search-based approach to FMS using CBCS has been introduced in [13]. A related approach to "descriptor-driven transformation" in an audio mosaicing context is introduced in [1], but without the real-time capability that is one of the key goals of CBCS.

The present pilot study does not implement a full FMS approach, but limits itself to parameters of pitch and intensity awaiting a more general FMS implementation in the future, or its integration with a hybrid concatenative synthesis approach, i.e. concatenation of segments in a parametric sound representation such as additive or source–filter models, as employed by SYNFUL [7].

3. ACCURATE TARGET PITCH AND LOUDNESS TRANSFORMATION

One advantage afforded by corpus-based transposition is that the larger the sample corpus, the more likely it is to contain a unit with a descriptor value within a given threshold of a target value. In order to modulate this unit to precisely match the target value, a relatively small alteration in the original unit can be made. Therefore other descriptors, for which modulation is not desired, remain relatively undisturbed.

Starting with the pitch for a given corpus unit f_u in Hz or m_u in MIDI note number that have already been estimated during the descriptor analysis phase, and given a desired target pitch of f_t Hz or note number m_t, we can determine the necessary transposition in semi-tones t, or directly the needed resampling factor $r = 2^{t/12}$ as

$$t = m_t - m_u = \frac{12}{\log 2}(\log f_t - \log f_u) \quad (1)$$

$$r = f_t / f_u = 2^{(m_t - m_u)/12} \quad (2)$$

since conversion between Hertz and MIDI, relative to a reference tuning of $f_r = 440$ for $m_r = 69$, are given by

$$m = m_r + \frac{12}{\log 2} \log \frac{f}{f_r} \quad (3)$$

$$f = f_r \cdot 2^{(m - m_r)/12} \quad (4)$$

Analogously, for a corpus unit of mean loudness level l_u in dB, and given a desired target level of l_t, we can calculate the necessary gain factor g in dB and the resulting amplitude multiplication factor $a = e^{g/20 \log e}$ as

$$g = l_t - l_u \quad (5)$$

$$a = e^{\frac{l_t - l_u}{20 \log e}} \quad (6)$$

Note that with *loudness*, we sloppily denote the mean logarithmic energy of a corpus unit, not the psychoacoustic percept of sound pressure. Neither do we take the *sonie* into account, for the moment, i.e. the time-dependent perceptual integration of loudness of the unit.

4. IMPLEMENTATION

A pitch- and loudness-modulation module has been implemented combining the latest full releases of CATART-1.2.2, FTM.2.5.0.BETA.22, and bach-v0.6.7 alpha for MAX/MSP5. The CATART software system for MAX/MSP realises corpus-based concatenative synthesis in real-time. It is a modular system based on the freely available FTM&CO extensions[2] [10], providing optimised data structures and operators in a real-time object system. CATART is released as free open source software at `http://imtr.ircam.fr`.

4.1. Bach

Using the bach library developed by Andrea Agostini and Daniele Ghisi for MAX/MSP,[3] a "target pitches" interface has been implemented to combine with existing CATART modules. The first major advantage of bach is its visual interface, capable of representing several precise gradients of microtonal pitch. A second feature is the sequencer playback of the bach.score (for metered music) or bach.roll object which can store the entire score or pitch content of a piece in musical notation. All of these features can be harnessed and interfaced with CATART.

4.2. Corpus-Based Transposition

Our current model of targeted transposition works best with a corpus whose grains are clearly segmented into units of definable and constant pitch, for example by using segmentation based on change of pitch on a harmonic sound, or by loading banks of samples. One or more target pitches are defined before playback. As grains are selected from the corpus by proximity to target descriptors, which may include pitch itself and/or other descriptors,

[2] `http://ftm.ircam.fr/`

[3] `http://www.bachproject.net`

their note number content is examined, and a transposition value equivalent to the difference between the estimated note number and the target pitch is sent to CataRT before playback as defined in equation 1. If more than one target pitch is defined, for example a harmonic field of possible pitches, the pitch of each sample can be either drawn at random (with or without replacement) or chosen based on the shortest distance to the original pitch of the unit. Using an interface combining modules from CataRT and the bach library, a dense field of microtonal pitches can be easily edited by the user (see Figure 1).

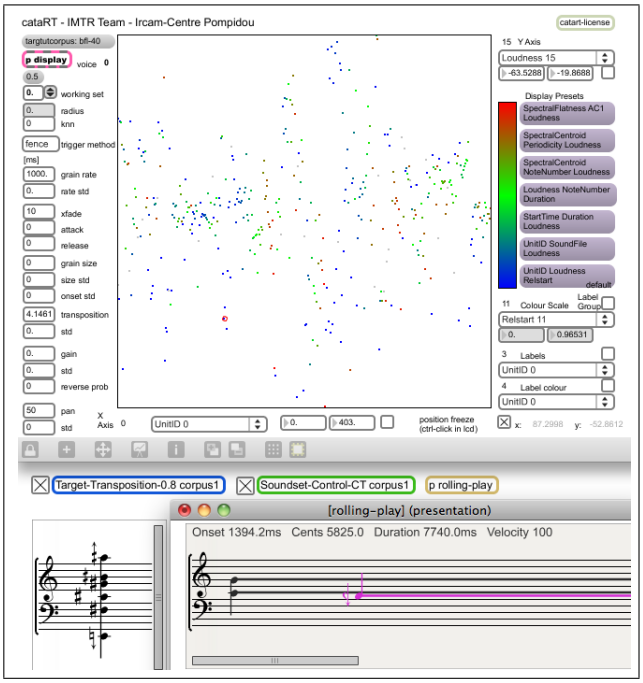

Figure 1. Screenshot of the targeted transposition interface combining CataRT and bach modules.

4.2.1. Pitch descriptors

For units with relatively low noise content and stable pitch, CataRT's NoteNumber descriptor, based on the *yin* pitch analysis algorithm [2], yields a useable pitch estimate.

4.2.2. Sample meta-data

However for percussive, unstable, or noisy units, CataRT's pitch estimate may not correspond to the perceived pitch. For soundfiles annotated with pitch information in their file names, for example in the case of sample banks, the ideal note number value is gleaned through a regexp operation on the filename, and if CataRT's note number estimation falls outside an acceptable percentage of this value, the ideal note number from the filename is used instead. A threshold value permits a maximum transposition distance in semitones, rejecting grains whose note number value falls outside this acceptable range. The pitch extracted from the filename on import is stored in the corpus as new descriptor FilenameNoteNumber, which serves also for selection according to pitch during composition.

4.3. Loudness Modulation

The current version of CataRT includes a loudness descriptor in units of decibels, therefore a simple subtraction is sufficient to adjust the gain on playback to a target value, as given in equation 5.

4.3.1. Sound sets

A finer control of loudness can be constrained by other descriptor values, for example SoundSet, a user-specified index associated with each unit. This is particularly useful when using different directories of samples, for example, associated with different instruments or playing techniques. For this purpose a SoundSet-control abstraction was created, which allows various subsets of the corpus grouped by SoundSet (indexed by directory by default) to be enabled and disabled in real time. Integrated into this module is a prototype mixer for loudness modulation, allowing a specific decibel level to be sent according to SoundSet classification before playback (see Figure 2).

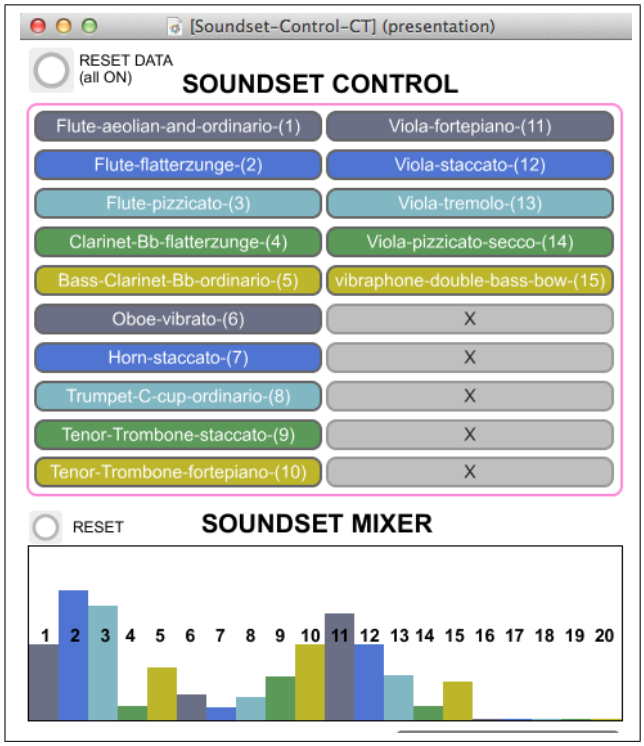

Figure 2. SoundSet control module, allowing target loudness to be chosen independently for each SoundSet.

4.4. Corpus-Based Transcription

In addition to its capabilities for real-time synthesis, CataRT has been used effectively for real- and deferred-time audio mosaicing and computer-assisted composition [4]. In both cases, a live or recorded audio input target

is analyzed and compared to a preloaded corpus according to descriptors chosen and weighted by the user. This process may be termed "Corpus-Based Transcription" and the goal is to create a mosaic of samples form the corpus that best approximates one or more audio features of the target.

Taking a feature modulation approach, corpus units can be altered to match better the descriptor values of the target. In the simplest case, the feature used to match target to corpus is the same one modulated. For example when the catart.analyzer~ module is used to retrieve NoteNumber estimates from the audio input and these values are treated as targets for pitch modulation, the resulting transpositions are relatively small resulting in a re-synthesis relatively faithful to the original corpus samples. However as other descriptors are added, values for transposition become higher, resulting in more significant changes in sample playback speed.

5. RESULTS AND DISCUSSION

The musical results of this approach can already be heard in new compositions by the authors, and in sound examples accessible online[4]. While these examples answer the musical motivations presented in the introduction, they point the way for several future directions.

5.1. Ensemble Musiques Interactives

Targeted transposition with CataRT made its public performance debut in a new piece *Five Out of Six* for the Ensemble Musiques Interactives by Christopher Trapani at the Festival of Interactive Music at Columbia University in March 2012. A large bank of instrumental samples is distributed over two CataRT modules, each carrying up to 27 SoundSets per corpus. An interface permitting a high degree of control over all playback parameters (grain length, envelope, gain) creates a constantly evolving web of textures in real time. A second component of the work involves live video, in collaboration with the Madrid-based duo Things Happen. Using MIDI controllers to manipulate up to three layers of live images, musical and visual data are freely exchanged and interact. For instance, the degree of luminosity of an image corresponds to a given descriptor continuum of a selected grain, or the movement of a projected image across a screen is broken down into x- and y-coordinates that correspond to two descriptors on the axes of the catart.lcd, so that the position of the image triggers grain selection according to a predetermined descriptor space.

5.2. Voice and Electronics

In a new work *Without Words* by Aaron Einbond for voice, ensemble, and live electronics commissioned by the Fromm Music Foundation for Ensemble Dal Niente and premiered in June 2012, the approach of Corpus-Based Transcription is used to create a mosaic of vocal samples based on targets from live input and pre-recorded field recordings. Due to listeners' perceptual sensitivity to playback of recorded voice, only small alterations in the corpus of vocal samples could be tolerated. Therefore NoteNumber alone was used as a probe-feature from the target to search the corpus for matching units, and the chosen unit descriptor values differ from those of the corpus by a maximum of two semitones, as summarized below. In this case corpus-based transposition results in relatively small changes in the unit playback speed. Nevertheless, the added nuance and variability in unit playback produces a noticeably more rich and dynamic synthesized texture.

5.3. Statistical Evaluation

These two recent compositions present case studies with which to quantify and evaluate the effectiveness of the Corpus-based transposition approach. For *Five Out of Six* two CataRT modules are used, each with its own preloaded corpus containing respectively 1907 units in 682 sound files of 76.2 min. using 769.2 MB and 1518 units in 636 sound files of 80.5 min. using 812.8 MB. The former is divided in to 27 SoundSets ranging in size from 6 to 338 units and corresponding to a collection of standard orchestral solo instruments, one instrument per set, including those shown in Figure 2. The SoundSet-control abstraction provides a fast and flexible way to navigate the instrumental timbres of such a large corpus.

Without Words also employs two corpora containing, respectively, vocal samples and instrumental samples. The former contains 3457 units in 293 sound files of 42.0 min. using 424.1 MB and is divided into 10 SoundSets sorted by vocal performance technique. The latter contains 8443 units in 1343 sound files of 180.2 min. using 1819.0 MB, divided into 15 SoundSets corresponding to the live instruments of the ensemble as recorded by the composer to generate the source-material for the score.

For a typical Corpus-based transcription task in which a target field recording is analyzed and the single descriptor MIDI NoteNumber is used to probe the corpus of vocal samples, the median transposition (absolute value) required during 3039 unit selections is 0.02 semitones and the maximum is 1.81 semitones. This is consistent with the assumption that transposition values for a large corpus are small enough to preserve sound quality and minimally affect other descriptors. However if a second descriptor Periodicity is introduced into the selection along with NoteNumber, the median transposition rises to 0.45 semitones and the maximum to 6.49. Adding more than two descriptors to the selection raises these values further.

6. CONCLUSION AND FUTURE WORK

These preliminary results immediately suggest several promising directions for further research and creative application. These include improvements in estimation of pitch and other features, further exploitation of meta-

[4] http://vimeo.com/user10514686

data, and implementation of a more comprehensive FMS framework.

6.1. Pitch Estimate

The averaged gbr.yin~ pitch detection employed by CataRT to calculate unit NoteNumber descriptors is problematic especially for percussive sounds with a noisy attack transient, for example *pizzicato* strings or vibraphone played with with knitting needles. The detection could be improved upon, for example, by removing attacks or other noisy frames before averaging the pitch. Alternatively, one could calculate the median pitch on the whole segment, which should be undisturbed by the attack. That is possible to implement in CataRT-1.5's modular descriptor analysis architecture [14]. However the use of pitch meta-data from the filename will always be another effective method for difficult-to-detect pitches. Not merely an *ad hoc* solution, this alternative is necessary to accommodate users' subjective judgements of the pitch of instrumental samples of noisy or extended playing techniques that may leave only a faintly-perceptible pitch.

6.2. Meta-Data

Beyond pitch meta-data, other meta-data could be useful for generalized corpus-based feature modulation synthesis to include features not easily calculated on import. For example, "spatial location" descriptors could be defined in Cartesian or spherical coordinates and associated with each soundfile. These could then be manipulated and interpolated like other existing descriptors, with potential uses for spatialized CBCS as described in [3].

6.3. Feature Modulation Synthesis

Finally existing literature on feature modulation synthesis will be adapted for a more comprehensive corpus-based feature modulation synthesis framework. The expanded and expandable descriptor list in CataRT-1.5 will be advantageous in developing a list of modulatable timbral features, for instance spectral centroid, spectral flatness, and further features that exist in current or upcoming versions of CataRT may be processed based on a comparison of pre-calculated descriptor values and target descriptor values.

7. ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for their helpful comments. The work presented here is partially funded by the *Agence Nationale de la Recherche* within the project *Topophonie*, ANR-09-CORD-022, see `http://topophonie.fr`.

8. REFERENCES

[1] G. Coleman, E. Maestre, and J. Bonada, "Augmenting sound mosaicing with descriptor-driven transformation," in *Digital Audio Effects (DAFx)*, Graz, Austria, 2010.

[2] A. de Cheveigné and H. Kawahara, "YIN, a Fundamental Frequency Estimator for Speech and Music," *Journal of the Acoustical Society of America (JASA)*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.

[3] A. Einbond and D. Schwarz, "Spatializing timbre with corpus-based concatenative synthesis," in *Proc. ICMC*, New York, 2010, pp. 72–75.

[4] A. Einbond, D. Schwarz, and J. Bresson, "Corpus-based transcription as an approach to the compositional control of timbre," in *Proc. ICMC*, Montréal, Canada, 2009.

[5] M. Hoffman and P. Cook, "Feature-based synthesis: mapping acoustic and perceptual features onto synthesis parameters," in *Proc. ICMC*, Copenhagen, Denmark, 2006.

[6] ——, "Real-time feature-based synthesis for live musical performance," in *Proc. NIME*, 2007.

[7] E. Lindemann, "Music synthesis with reconstructive phrase modeling," *IEEE Signal Processing Magazine*, vol. 24, no. 2, pp. 80–91, Mar. 2007.

[8] T. H. Park *et al.*, "Feature modulation synthesis (FMS)," in *Proc. ICMC*, Copenhagen, 2007.

[9] T. H. Park, Z. Li, and J. Biguenet, "Not just more FMS: Taking it to the next level," in *Proc. ICMC*, Belfast, 2008.

[10] N. Schnell, R. Borghesi, D. Schwarz, F. Bevilacqua, and R. Müller, "FTM—Complex Data Structures for Max," in *Proc. ICMC*, Barcelona, 2005.

[11] D. Schwarz, "Concatenative sound synthesis: The early years," *Journal of New Music Research*, vol. 35, no. 1, pp. 3–22, Mar. 2006, special Issue on Audio Mosaicing.

[12] ——, "Corpus-based concatenative synthesis," *IEEE Signal Processing Magazine*, vol. 24, no. 2, pp. 92–104, Mar. 2007, special Section: Signal Processing for Sound Synthesis.

[13] D. Schwarz and N. Schnell, "Descriptor-based sound texture sampling," in *Sound and Music Computing (SMC)*, Barcelona, Spain, Juillet 2010, pp. 510–515.

[14] ——, "A modular sound descriptor analysis framework for relaxed-real-time applications," in *Proc. ICMC*, New York, NY, 2010.