WHAT NEXT? CONTINUATION IN REAL-TIME CORPUS-BASED CONCATENATIVE SYNTHESIS

Diemo Schwarz
Ircam – Centre Pompidou, Paris

Sylvain Cadars
Sound Engineer and Researcher

Norbert Schnell
Ircam – Centre Pompidou, Paris

ABSTRACT

We propose an extension to real-time corpus-based concatenative synthesis that predicts the best sound unit to follow an arbitrary sequence of units depending on context. This novel method is well suited to interactive applications because it does not need a preliminary analysis or training phase. We experiment different modes of interaction and present results with a quantitative evaluation of the influence of the new method on corpora of drum loops, voice, and environmental sounds.

1. INTRODUCTION

Real-time corpus-based concatenative synthesis (CBCS) [10] is a new technique for musical sound synthesis that has many artistic and commercial applications, and is more and more widely used in concert, performance, and production settings [12]. It permits to create music by selecting snippets of a large database of pre-recorded sound by navigating through a space where each snippet takes up a place according to its sonic character, such as pitch, loudness, brilliance. This allows to explore a corpus of sounds interactively, or by composing this path, and to create novel harmonic, melodic and timbral structures.

The database of source sounds is segmented into short *units*, and a *unit selection* algorithm finds the units that match best the sound to be synthesised, called the *target*. The selection is performed according to a match with the *descriptors* of the units, which are characteristics extracted from the source sounds, or higher level meta-data attributed to them. The selected units are then concatenated and played, after possibly some transformations.

In non-real-time CBCS, the stress was always on how to generate sequences that simultaneously satisfy the two criteria of finding a good match with the target and finding pairs of units to concatenate which present an appropriate transition (e.g. avoid abrupt spectral changes).

Real-time CBCS systems, on the other hand, were up to now almost exclusively concerned with the first criterion, and the choice of which unit to concatenate with the last selected one was unconstrained. However, finding the next unit to play, depending on interactive input and on the context of what has been played just before, is the natural next step in augmenting the expressive power of this approach, allowing to exploit the inherent behaviour of a corpus, defined by the many examples contained therein, and to guide the synthesis process.

2. PREVIOUS AND RELATED WORK

CBCS has its roots in concatenative speech synthesis based on unit selection, where continuity is a prime interest to limit glitches and other artefacts at the join points of the speech units to be concatenated. Blouin [1] reports that respecting the high-level context, defined by phonological and linguistic features, is the best way to assure a good concatenation quality.

For musical sound synthesis, a general context distance for non-real-time unit selection by dynamic programming has been proposed for the *Caterpillar* system [8], which favours the selection of units out of the same context in the database as in the target sequence. In real-time CBCS, some related approaches how to model the continuation of a sequence of units are explained in the following. See [9] for an extensive overview and classification of past and current work on CBCS.

Hidden Markov Models do usually and tractably only model a context of order one, i.e. only the last state is taken into account for the decision about the next state, so that longer term context would have to be encapsulated in the state variable. However, their extension *periodic N-grams* have been successfully applied to probabilistic rhythm transcription and generation [6] where they predict the probability of the next beat depending on the $N-1$ previous beats. Their drawback is that this model has to be precalculated on the corpus.

Input Driven Resynthesis [7] starts from a database of recorded FFT frames as units, analysed for loudness and 10 spectral bands as descriptors, which form a trajectory through the descriptor space. Phase vocoder overlap–add resynthesis is controlled in real-time by audio input that is analysed for the same descriptors, and the selection algorithm tries to follow a part of the database's trajectory, limiting jumps by distance constraints.

The first commercial application using ideas of CBCS is the *Synful* software synthesiser [5], which aims at the reconstitution of expressive solo instrument performances from a corpus of real instrument recordings. Here, the right type of transition and sustain units is predicted from real-time MIDI input by a set of rules on a pitch- and loudness-based distance.

The *Guidage* algorithm [3] is based on the audio version of the *Factor Oracle*, which is a model of the temporal morphology of audio content. In fact, this amounts to precalculating the concatenation distance on the whole corpus of vector quantised FFT-frames. The applica-

tions of *Guidage* are retrieval of bits of audio from large databases for similarity search and concatenative synthesis. In an unfortunate twisting of the established terminology in CBCS, in [3] the source database is called the target, while the synthesis target is called the query.

Hazan et al. [4] work on interactive continuation of drum sequences by expectation modeling by a Factor Oracle model. Here, the database is restricted to a window in the live input.

The drawback of these last two approaches is the lack of control possibilities and restricted interaction: while *Guidage* is not interactive at all, Hazan's system reacts to the incoming signal, but for both the only access to the retrieval is by examples of bits of sound and the evident higher-level symbolic meaning of sound descriptors is neglected. Our CataRT system [11] explained in section 4 does exploit the musically meaningful information in the descriptors. In the following sections we will see how context-aware continuation is added to it.

3. CONTEXT-AWARE UNIT SELECTION

CBCS selects units out of a database of N segments of source sounds, analysed for K descriptors, in order to match target units, specified in the same descriptors, minimising the distances defined in the following.

3.1. Descriptors

The sound descriptors that are analysed in the existing system are the fundamental frequency, aperiodicity, loudness, the spectral descriptors centroid, sharpness, flatness, high frequency energy, mid frequency energy, high frequency content, and first order autocorrelation coefficient (that expresses spectral tilt). To this list, we added zero crossing rate, log attack time, spectral spread, skewness, kurtosis, and rolloff, and energy in 12 Bark-scale critical frequency bands, in order to model the human perception of the frequency spectrum [2]. The instantaneous descriptors at a frame-rate of about 50Hz are condensed to a scalar values by taking the mean over the duration of each segment.

3.2. Distance Functions

The quality of the match between a database and a target unit is determined by two distance functions between their descriptor vectors. First, the *target distance* C^t is a weighted Euclidean distance function that expresses the match between the target unit t_τ and a database unit u_i

$$C^t(u_i, t_\tau) = \sum_{k=1}^{K} w_k^t \, C_k^t(u_i, t_\tau) \tag{1}$$

based on the individual squared distance functions C_k^t for descriptor k between target descriptor value $t_\tau(k)$ and database descriptor value $u_i(k)$, normalised by the standard deviation of this descriptor over the corpus σ_k:

$$C_k^t(u_i, t_\tau) = \left(\frac{t_\tau(k) - u_i(k)}{\sigma_k} \right)^2 \tag{2}$$

Now, using the target distance alone, the choice of the best match would be done completely oblivious of the previous states of the system and the surroundings of the candidate units in the database. Thus, in order to capture the context of the database units and to match it to the context of the target units, and therefore to predict the best continuation of this context, we define the second distance function *continuation distance* C^c as

$$C^c(u_i, t_\tau) = \sum_{j=0}^{r} w_j^c \, C^t(u_{i-j}, t_{\tau-j}) \tag{3}$$

based on the target distances between the r preceding units in the database u_{i-j} and in the target $t_{\tau-j}$. This is a causal version of the *symmetric context distance* C^x for non-real-time CBCS, proposed in [8], favouring the selection of units out of the same context in the database as in the target (known entirely in advance) by considering a sliding context in a range of r units around the current unit with weights w_j^x decreasing with distance j:

$$C^x(u_i, t_\tau) = \sum_{j=-r}^{r} w_j^x \, C^t(u_{i+j}, t_{\tau+j}) \tag{4}$$

The target distance terms in equation (3) are balanced by the weight w_j^c, which for $j = 0$ gives the influence of the direct target match, and for $j > 0$ the influence of the context. Usually, we keep $w_j^c = \frac{1-w_0^c}{r-1}$ for $j > 0$ so that the size of the context r does not influence the balance of target to context. Other distributions of w_j^c are possible.

Setting $w_0^c = 0$ means that the target unit is ignored and the selection is solely based on the best continuation of the current target context, in the sense of finding the unit that follows the sequence of $r-1$ units from the database with the least distance to the target units $t_{\tau-r} \dots t_{\tau-1}$.

Note that the continuation distance does not correspond to the concatenation distance known from speech synthesis, which predicts continuity of sound at the join points of selected units, but to the continuity of a trajectory through the database, predicting the joint similarity of sequences of length r, and thus assessing the appropriateness of a candidate unit to continue the trajectory according to its context in the database.

3.3. Selection and Update

Given a target context $t_{\tau-r} \dots t_\tau$ where we want to select the current unit at time τ, the algorithm performs a linear search for the database unit u_i with the least continuation distance:

$$i = \operatorname*{argmin}_{1 \le j \le N} C^c(u_j, t_\tau) \tag{5}$$

Note that for $r = 0$ this falls back to the context-free selection of the current unit as being the closest to the current target unit t_τ, regardless of the previous states.

The selected unit is then played and the target context is updated by advancing τ (in practice, shifting the target units left, and setting t_τ from input of new target values).

This triggers a new round of selection and can happen at the end of the currently playing unit for continuous synthesis, or at any rate for layered or event-based synthesis.

4. IMPLEMENTATION AND INTERACTION

We implemented the model as an extension to the modular CataRT real-time interactive CBCS systemfor *Max/MSP*, allowing descriptor-based navigation through a two- or higher-dimensional descriptor space [12]. That space is constituted by segmenting and analysing any number of sound files or live audio input. Using CataRT as a testbed allowed us to easily experiment different modes of interaction and combinations and weightings of descriptors.

Interaction with CBCS can here be approached in two ways: First, using a target sequence, possibly cyclic as with a drum loop, one can interact by choosing the target from an existing recording, changing it, changing the weights of the descriptors, setting the blur radius of the selection (that permits units close to the target to be selected randomly), and changing the size and influence of the context (parameters r and w_j^c).

The second approach to interaction, called *autocontext* is by replacing the target sequence by live user input, combined with the intrinsic behaviour of the corpus, as defined by the continuation distance. Here, the target sequence could be recorded from user input of units and descriptors, and the balance between strictly following the target and taking into account the context of the last r units is controlled in the same way than above. This is feasible if a steady stream of target descriptors is available, e.g. from analysing live audio. However, in symbolic interaction such as via a 2D interface as that of CataRT, the user input is sparse, but the system should output units at a faster and more constant rate than the input is available (repeating the last target is not acceptable). Therefore, we propose a mode where, as long as there is no user input, $w_0^t = 0$, i.e. no target is taken into account, and where the target context is constituted from the last r *selected units*. When input is available, it is used as the current target unit t_τ and $w_0^t > 0$. This mode will follow the inherent behaviour of the corpus when left alone, but can be nudged into a specific direction by the user, possibly taking the system onto a new track in the sonic landscape of the corpus.

5. RESULTS AND EVALUATION

In this section, we will try to evaluate our approach to sequence prediction and assess the influence of the parameters of our method with different corpora. Sound examples and the CataRT software system can be found at http://imtr.ircam.fr under Projects, Corpus Based Synthesis.

First, we constituted a corpus of 78 drum loops for commercial music production [2]. All loops are 2 bars long, binary, in a tempo of 120bpm, and don't contain extensive microtiming (a.k.a. groove) such as shuffle or swing. This ensures that the loops can be reliably segmented into 16th note units by splitting them into 32 equal-sized segments. The choice of drum loops also limits the range of sounds which permits the validation of the descriptors, and provides us with a well-defined aim for evaluation of the target and continuation distances, namely the match with a similar sounding drum loop and a choice of rhythmic sub-sequences that reproduce the style or the feel of the target, even if the individual drum sounds are different.

For these experiments, the sequences of drum sounds were generated with a target given by one chosen drum loop, that was then removed from the corpus [2], and the descriptor weights were tuned by hand. Subjective evaluation shows that the structure and style of the rhythm is clearly preserved, the stressed beats and basic bass- and snare-drum pattern being reproduced by the selection. However, for the context-free selection ($r = 0$), we observe the tendency to produce rather jagged rhythms, mainly due to cymbal sounds, that have a decay much longer than a segment, and that would be cut off by the next segment. Here, augmenting the context size and weight clearly improves the result and the perceived steadiness of the rhythm. We can see this in the spectrograms in figure 1 where the context makes the selection result resemble more closely to the target as a whole.

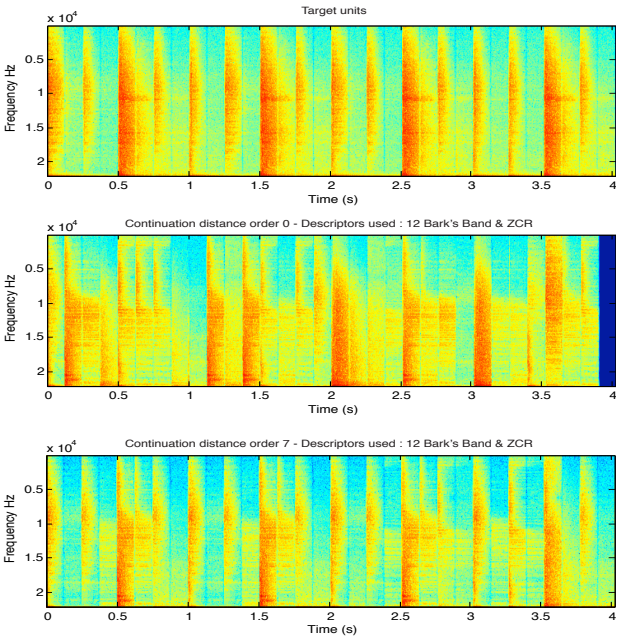

Figure 1. Influence of the context on selection of a drum loop. Upper spectrogram: target, middle: selection without context, lower: selection with context.

Further, to evaluate the influence of the new context distance to unit selection quantitavely, we examine the length and number of contiguous sequences of units selected from the corpus under different parameters of the continuation distances (figure 2). This is based on the assumption that, given a target sequence that does not occur in the corpus, the resulting selection will jump around

wildly in the corpus, satisfying the target match wherever the distance is lowest. Now, adding a context, the number and length of contiguous sequences increase because the continuation distance is lowest for the unit following a sequence, wherever a too large target distance does not make us break out of the sequence.

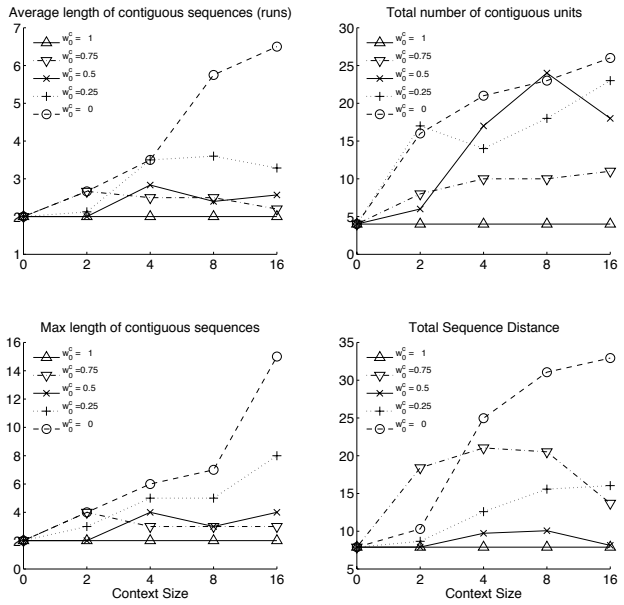

Figure 2. Influence of the context size and weights w_0^c on contiguous sequences and total distance.

The second corpus is made up of speaking voice, segmented into phonemes by automatic alignment with the known text. This corpus allows to evaluate how the context reproduces the morphology of a language on syllable and word level. While browsing the corpus according to sonic characteristics, the context helps to keep a speech-like morphology of syllables and short words, without the output being intelligible speech, of course. Artefacts by discontinuities at segment (phoneme) level are greatly reduced with rising influence of the context.

Third, we generalised the corpus to environmental sounds to evaluate the short- and medium-term structure-keeping capability of the context. Here, the texture of the source sounds is kept, and with rising context influence, the transitions typical for the original corpus start to shine through in the selection output. Future work will complete the objective evaluation statistics with experiments on these corpora of voice and environmental sounds.

So far, we did not yet address the problem of the choice of the descriptor weights w_k^t for the target distance in equation (1). One possibility for this comes from speech synthesis and performs a search for the combination of weights that minimise an acoustic distance measure between the target and selected sequences.

6. CONCLUSION

We presented a new method of taking into account the inherent behaviour of a sound corpus for synthesis by the definition of a continuation distance capturing the context of a unit in the database and permitting to predict the most probable continuation unit, depending on that context.

One advantage of our method is that it does not need a preliminary analysis or training phase, so that it is applicable to corpora constituted by live recording of audio. Still, the computational complexity is only multiplied by the context size r from $O(KN)$ distance calculations to $O(rKN)$ in our straightforward linear implementation. This can be reduced to negligible $O(rK\log(N))$ by a binary search using a kD-tree index.

Our evaluation showed a mesurable effect on continuity for a test corpus of drum loops, and distinctly noticeable effects for voice and environmental sound corpora.

This improves the musicality of the synthesised sound and augments expressivity since interaction is enriched by a notion of context and the constraints coming from the sound source.

7. REFERENCES

[1] Ch. Blouin. *Sélection des unités pour la synthèse vocale par concaténation*. PhD thesis, France Télécom R&D Lannion, LIMSI, 2003.

[2] S. Cadars. Modélisation temporelle et synthèse concaténative de boucles rythmiques. DEA ATIAM, Univ. Paris VI, 2007.

[3] A. Cont, S. Dubnov, and G. Assayag. Guidage: A fast audio query guided assemblage. In *ICMC*, 2007.

[4] A. Hazan, P. Brossier, P. Holonowicz, P. Herrera, and H. Purwins. Expectation along the beat: A use case for music expectation models. In *ICMC*, 2007.

[5] E. Lindemann. Music synthesis with reconstructive phrase modeling. *IEEE Sig. Proc. Mag.*, 24(1), March 2007.

[6] J.K. Paulus and A.P. Klapuri. Conventional and periodic N-grams in the transcription of drum sequences. In *Intl. Conf. on Multimedia and Expo*, 2003.

[7] M. Puckette. Low-dimensional parameter mapping using spectral envelopes. In *ICMC*, Miami, 2004.

[8] D. Schwarz. *Data-Driven Concatenative Sound Synthesis*. PhD thesis, Université Paris 6, 2004.

[9] D. Schwarz. Concatenative sound synthesis: The early years. *JNMR*, 35(1), March 2006.

[10] D. Schwarz. Corpus-based concatenative synthesis. *IEEE Sig. Proc. Mag.*, 24(2), March 2007.

[11] D. Schwarz, G. Beller, B. Verbrugghe, and S. Britton. Real-Time Corpus-Based Concatenative Synthesis with CataRT. In *DAFx*, Montreal, 2006.

[12] D. Schwarz, S. Britton, R. Cahen, and T. Goepfer. Musical applications of real-time corpus-based concatenative synthesis. In *ICMC*, Copenhagen, 2007.