USING CONCATENATIVE SYNTHESIS FOR EXPRESSIVE PERFORMANCE IN JAZZ SAXOPHONE

Esteban Maestre, Amaury Hazan, Rafael Ramirez, and Alfonso Perez
Music Technology Group, Universitat Pompeu Fabra
Ocata 1, 08003 Barcelona, Spain

Abstract

We present here a concatenative sample-based saxophone synthesizer using an induced performance model intended for expressive synthesis. The system consists on three main parts. The first part provides the analysis of saxophone expressive performance recordings and the extraction of descriptors related to different temporal levels. With the obtained descriptors and the analyzed samples, we construct an annotated sample database extracted directly from the performances. For the second part, we use the annotations to induce a performance model capable of predicting some features related to expressivity. In the third part, the predictions of the performance model are used to retrieve the most suitable note samples for each situation, and transform and concatenate them following the input score and the induced model.

1 Introduction

Modern concatenative synthesizers have reached nowadays high sound quality synthesis and offer a wide range of synthesis parameters, including some related to expression, normally at note and transition levels. However, these parameters must be tuned manually depending on user wishes, leading to a high effort to represent the expression and/or naturalness that musicians introduce when performing a piece.

In this paper we describe an approach to investigate the synthesis of jazz saxophone expressive performances based on concatenating note samples. The aim is to generate an expressive audio sequence out of an input score by means of a previously induced performance model and making use of an annotated saxophone note sample database extracted from real performances. The architecture of our system can be summarized as follows. First, given a set of expressive performance recordings, we get a description of the audio by carrying out segmentation and characterization at different temporal levels (note, intra-note, transition) and build an annotated database of pre-analyzed note segments for being used later in the synthesis stage. A performance model is trained using inductive logic programming techniques for matching the score with the description of the performances that we got when constructing the database. For synthesizing an audio given a score, the performance model predicts an expressive performance description, from which the most suitable note samples from the database are retrieved, transformed and concatenated. This work conforms an audio analysis/synthesis preliminary application of the studies on expressive performance already started in (Ramirez and Hazan 2005).

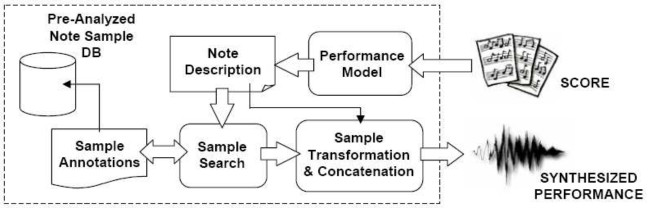

Figure 1: Overview of the system architecture

In terms of concatenative synthesis, the author in (Schwarz 2000) presented a general purpose system based on data-driven unit selection which allows different applications ranging from high quality sound synthesis to free synthesis. Authors in (Bonada and Loscos 2003) construct a singing voice synthesizer based on spectral concatenation resulting in a high quality system with a broad variety of synthesis parameters. Although the synthesis techniques proposed in these works have set up the basis of the present work, these systems lack of expressivity knowledge. Previous research in expressive performance synthesis includes a broad spectrum of approaches and techniques. In (J. Arcos and Serra 1997), it is presented a case-based reasoning system for generating expressive performances of melodies based on human examples. In (de Poli et al. 2004), the authors present an approach to modify the expressive content of a performance in a gradual way using a linear model to carry out the alterations based on previous semiautomatic segmentation and modifying melodies both at symbolic and audio signal levels. Trumpet performance is studied in (Danneberg and Derenyi 1998) by computing amplitude descriptors, and statistical analysis techniques used for analyzing trumpet envelopes led to find significant envelopes groupings, and to extend the work to a system that combined instrument and performance models. Authors in (Dubnov and Rodet 1998) have followed a similar line. We propose here to use the same database for carrying out learning and synthesis, having already obtained some preliminary results on intra-note amplitude features prediction using first order logic decision trees (Ramirez et al. 2005).

The rest of the paper is organized as follows. Section 2 briefly describes how expressive audio recordings have been analyzed and annotated. Then, we explain in Section 3 the procedure for building the performance model. Section 4 outlines how samples are retrieved from the database, transformed, and concatenated. Section 5 concludes and points out further work.

2 Audio analysis

Expressive performance recordings are automatically analyzed, and a set of low-level descriptors are computed for each frame. Then, we perform a note segmentation using low-level descriptor values and fundamental frequency. Using note boundaries and low-level descriptors, we carry out energy-based intra-note segmentation, and a posterior intra-note segment amplitude envelope characterization, as well as a transition description. This information will be used for (1) modeling expressive performance and (2) annotating the sample database used later in the synthesis stage. We have used an audio database consisting of 4 jazz standards played at 11 different tempi around the nominal one, played by a professional musician. Most of the phrases were repeated to test consistency between performances. We used a total of 3200 notes. The jazz standards recorded were *Body and Soul*, *Once I Loved*, *Like Someone In Love* and *Up Jumped Spring*.

Description scheme. In order to define a structured set of audio descriptors able to provide information about the expressivity introduced in the performance, we proposed in (Maestre and Gómez 2004) a broader version of the description scheme selected for this application (see Figure 2). Here, we define descriptors related to different temporal scales. Some features are defined as 'instantaneous' or related to an analysis frame, such as energy, fundamental frequency and spectral centroid. Some other are attached to a certain intra-note/transition segment (attack, sustain, release or transition), while descriptors attached to a certain note are also extracted. We consider that the proposed features can set up a simple but concise description of dynamics and articulation, adapted to our application context. However, it could be extended to provide a richer representation.

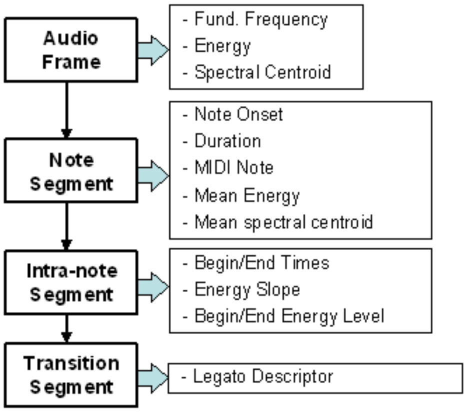

Figure 2: Audio description scheme

Audio segmentation and description. First, we get a melodic description of the audio phrases consisting on the exact onset and duration of notes, and the corresponding MIDI equivalent pitch. Notes are segmented using energy computation in different frequency bands and fundamental frequency. Energy onsets are first detected following a band-wise algorithm that uses some psycho-acoustical knowledge. In a second step, fundamental frequency transitions are also detected. Finally, both results are merged to find the note boundaries. We compute note descriptors using the note boundaries and the low-level descriptors values. The low-level descriptors associated to a note segment are computed by averaging the frame values within this note segment. Pitch histograms have been used to compute the pitch note of each note segment. A detailed explanation of the methods used for melodic description can be found in (Gómez et al. 2003). Then, notes extracted from the recordings are automatically segmented into attack, sustain, release or transition segments by studying energy envelope derivatives at different scales. After that, we approximate linearly the energy contour of each one of the extracted segments. In order to obtain a descriptor representing brightness, we extract a frame-averaged spectral centroid of the steady-state segment of each note. We also extract a legato numerical descriptor for transition segments. More explicit information about the procedure for carrying this intra-note and transition segmentation and characterization can be found in (Maestre and Gómez 2004). These descriptors, attached to each note sample note present in the database, will conform, together with other musical context descriptors used in the performance modeling stage, the annotations used for both, modeling expressive performance, and selecting, transforming and concatenating samples.

Database annotation. The recorded musical phrases are automatically segmented into notes using the information obtained from the melodic description outlined above. Each note sample from the performance is then indexed and annotated with the data coming from our note and intra-note segmentation and characterization. Note samples are stored as sequences of analyzed frames that will be used afterwards in the audio synthesis stage. We also classify notes into articulation groups depending on their context, dividing them in (1) coming from silence and going to silence, (2) coming from silence and going to transition, (3) coming from transition and going to silence, and (4) coming from and going to transition. This information will be used as a constraint during the sample retrieval stage in order to match the original articulation context of the notes used for synthesizing the output performance.

3 Expressive performance modeling

In this section, we briefly describe our approach to expressive performance modeling. Here we are concerned with note-level (in particular note duration, note onset, note energy, and note brightness), transition level (legato), and intra-note-level (in particular amplitude shape) expressive performance features. Each note in the training data is annotated with its corresponding extracted audio descriptors (see Section 2) and a number of score attributes representing both properties of the note itself and some aspects of the local context in which the note appears. Information about intrinsic properties of the note includes note duration, note metrical position, and note envelope information, while information about its context include the note Narmour group(s) (Narmour 1990), duration of previous and following notes, and extension and direction of the intervals between the note and the previous and following notes. With this information, we build two different models using machine learning techniques.

Note and transition level prediction. In order to induce predictive models for note duration ratio, onset deviation, note energy and legato, we have applied inductive logic programming techniques. We used a multi-variate regression tree model (Ramirez et al. 2005), i.e. a first order logic regression tree that contains logical tests at its nodes and five-dimensional vectors at its leaves, each one containing an averaged prediction of duration ratio, onset deviation, energy, and a pair of values for legato. The accuracy of the prediction model obtained is better than the best of several propositional models we have investigated. Details of these models can be found in (Ramirez and Hazan 2005).

Intra-note level prediction. In order to construct a predictive model for intra-note features (e.g. amplitude shape), we have devised a learning scheme based on note classification and a posterior class prediction, briefly described as follows. Notes are represented with vectors containing five features considered important for classifying samples into different note qualities in terms of amplitude/loudness: log-attack time, energy level at the end of the attack segment, sustain duration normalized to note duration, sustain energy slope, and spectral centroid (Peeters 2004). Using some of the features annotated for each note, we perform k-means clustering of the whole sample database, adding to the annotations the cluster of each note. Then, we build a note cluster prediction model based on a first order logic decision tree, which gives us a prediction of the note class given the score descriptors outlined at the beginning of this section.

4 Audio synthesis

Our system generates the audio sequence based on the predictions of the performance model and the annotated sample database. First, most suitable samples are selected taking into account the output of the performance model and the cost of the transformations to be applied. Selected samples are first transformed to fit the predicted note characteristics applying global note amplitude transformation, pitch shift and time stretch. After that, samples are concatenated by means of amplitude, pitch, and spectral shape interpolation applied to the resulting note transitions.

Sample retrieval. Using the note description given by the performance model, which includes a cluster prediction, the system performs sample retrieval inside the predicted cluster in two steps. First, the system re-classifies each predicted note into one of the four articulation groups (see Section 2) depending on its new context by looking at the output note sequence. In a second step, a search is performed within the notes of the predicted cluster belonging to such articulation group using a euclidean feature-weighted distance vector. For every note sample to be retrieved, a initial feature set consisting on MIDI pitch, duration, energy, and spectral centroid is used to compute the distance vector. Then, some features will be added depending on the articulation group to which the predicted note belongs, leading to a variable length feature vector. For the first articulation group (isolated note), no new features are added. For the second articulation group, two features (corresponding to the right side transition) are added: legato descriptor and pitch interval respect to the next note. Analogously, for the third and fourth articulation groups, the same pair of features are added for the left side transition, and for both transitions respectively. If a candidate note sample not presenting sustain segment would need to be time-stretched by an amount exceeding the duration of one frame, it is discarded from the search due to the fact that time stretch is going to be applied only in the sustain segment.

Sample transformation and concatenation. Once the note samples have been selected from the predicted cluster as outlined before, the system uses spectral processing techniques (Amatriain et al. 2002) for transforming each retrieved note sample in terms of amplitude, pitch and duration to match, in the same terms, the target description given at the output of the performance model. After that, samples are concatenated following the note sequence given at the output of the performance model. Note global energy prediction is applied first as a global amplitude transformation to the sample, since the energy envelope quality is already included in the cluster prediction. Then, pitch transformation is applied by shifting harmonic regions of the spectrum by an amount equal to the pitch interval between retrieved sample and predicted note, keeping the original spectral shape. After that, time stretch is applied inside the limits of the sustain segment by repeating or dropping the required number of frames to match the predicted duration. As a final step, the system restores the possible timbre discontinuities occurring within resulting transitions, in the neighborhood of the junction point of each pair consecutive notes (no silence in between). This is carried out by smoothing amplitude and pitch contours (approximated by a third order spline), and interpolating spectral shapes in order to avoid too sharp timbre changes.

5 Conclusion and future work

In this paper, we have presented an approach to investigate the synthesis of jazz saxophone expressive performances based on concatenating note samples. For our first experiments, we used the same set of saxophone expressive recordings for training the performance model and for constructing the database. Although the system is still in a very preliminary state, first synthesis results are promising. We carried out some tests synthesizing the same songs used for training the performance model and for building the sample database, resulting in an acceptable perceptual similarity at higher tempi. The synthesis of pieces not present in the training set showed our model to be somehow specific. Abrupt timbre changes during transitions lead us to think of incorporating the transition as a sample unity, while note-to-note sudden loudness/brightness changes demonstrated us that, in terms of classification, both feature selection and weighting, as well as the number of clusters, could become a subject of study for further developments of our system. Further work also includes studying pitch contour in order to model 'portamento-like' transitions and pitch modulations (e.g. vibrato) occurring within sustain segment. Moreover, due to the difficulties of evaluate the quality of the expressivity/naturalness of the synthesized performance, we must carry out more extended auditory tests in order to be able to tune or improve our system. This work should be considered as a step towards a methodology for the automatic creation of both the performance model and the sample database needed to carry out expressive synthesis, leaving clear other duties like trying out other instruments and other classification approaches.

References

Amatriain, X., J. Bonada, A. Loscos, and X. Serra (2002). Spectral processing. *DAFX Digital Audio Effects, ed. Udo Zőlzer*.

Bonada, J. and A. Loscos (2003). Sample-based singing voice synthesis based in spectral concatenation. In *Stockholm Music and Acoustics Conference*, Stockholm, Norway.

Danneberg, R. and I. Derenyi (1998). Combining instrument and performance models for high quality music synthesis. *Journal of New Music Research*.

de Poli, G., S. Canazza, C. Drioli, A. Rodà, and A. Vidolin (2004). Modelling and control of expressiveness in music performance. In *Proceedings of the IEEE, Vol.92*.

Dubnov, S. and X. Rodet (1998). Study of spectro-temporal parameters in musical performance, with applications for expressive instrument synthesis. In *IEEE International Conference on Systems Man and Cybernetics*, San Diego, USA.

Gómez, E., M. Grachten, X. Amatriain, and J. Arcos (2003). Melodic characterization of monophonic recordings for expressive tempo transformations. In *Proceedings of Stockholm Music Acoustics Conference*, Stockholm, Sweden.

J. Arcos, R. d. M. and X. Serra (1997). Saxex: a case-based reasoning system for generating expressive musical performances. In *Proceedings of the International Computer Music Conference*, Thessaloniki,Greece.

Maestre, E. and E. Gómez (2004). Automatic characterization of dynamics and articulation of monophonic expressive recordings. *Proceedings of the 118th AES Convention*.

Narmour, E. (1990). *The Analysis and Cognition of Basic Melodic Structures: The Implication Realization Model*. University of Chicago Press.

Peeters, G. (2004). A large set of audio features for similarity and classification. *CUIDADO IST Project Report*.

Ramirez, R. and A. Hazan (2005). Modeling expressive music performance in jazz. In *Proceedings of the Florida Artificial Intelligence Research Society Conference*, FL, USA.

Ramirez, R., A. Hazan, and E. Maestre (2005). Intra-note features prediction model for jazz saxophone performance. In *Proceedings of the International Computer Music Conference*, Barcelona, Spain.

Schwarz, D. (2000). A system for data-driven concatenative sound synthesis. In *Proceedings of the COST G-6 Conference on Digital Audio Effects*, Verona, Italy.