Pattern Recognition Letters 31 (2010) 1507–1513

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Modeling and segmentation of audio descriptor profiles with segmental models

Julien Bloit *, Nicolas Rasamimanana, Frédéric Bevilacqua

Ircam CNRS UMR STMS, 1 Place Igor Stravinsky, 75004 Paris, France

ARTICLE INFO

Article history:
Available online 10 November 2009

Keywords:
Hidden Markov models
Segmental models
Segmentation
Audio descriptors
Profiles
Music

ABSTRACT

We present a method to model sound descriptor temporal profiles using segmental models. Unlike standard HMM, such an approach allows for the modeling of fine structures of temporal profiles with a reduced number of states. These states, we called primitives, can be chosen by the user using prior knowledge, and assembled to model symbolic musical elements. In this paper, we describe this general methodology and evaluate it on a dataset made of violin recording containing *crescendo*/*decrescendo*, *glissando* and *sforzando*. The results show that, in this context, the segmental model can segment and recognize these different musical elements with a satisfactory level.

1. Introduction

The basic symbolic elements in modeling musical sound are often related to notes, i.e. events of constant pitch over a given duration. This assumption can however be limiting for a wide range of new music, from contemporary classical (Kaltenecker, 2001; Möller, 2009; Bevilacqua et al., 2006) to electroacoustic music (Smalley, 1997). In such cases, temporal variations of sound characteristics are central and should be emphasized for modeling.

Such considerations were already pointed out and addressed, as early as the mid-1900's, in Pierre Schaeffer's works on a generic ontology to describe musical sound (Schaeffer, 1966; Schaeffer et al., 1967). His categorization was deeply grounded on notions of temporal profiles of perceived spectral properties named *typomorphology*. Later, the role of temporal features was fully established from a perception standpoint (Grey, 1977; Mcadams et al., 1995).

From a computational point of view, sound categories defined by Schaeffer (summarized in Chion, 1983) are appealing. Models for some of these categories were recently implemented (Ricard and Herrera, 2004; Peeters and Deruty, 2008) using discriminative models. However, in these works, the aim was rather put on building parameters to condense the sound descriptors temporal profiles, which only captured specific aspects.

Similarly, we follow a morphological approach for musical sound modeling. Our general goal is to model temporal profiles of audio descriptors while relating them to a symbolic level. For this, we propose to use a Hidden Markov Model framework, which proved to be successful to model temporal data (Rabiner, 1989; Ghahramani, 2001; Young et al., 2006). The benefits of using such generative state models include the possibility to build in a straightforward manner hierarchical temporal structure, integrating prior knowledge. For example, in (Raphael, 1999; Ryynänen and Klapuri, 2008), a Hidden Markov Model (HMM) for constant pitch notes was incorporated into a higher level state model representing context knowledge (a score, a musicological model, or a perceptive model as in Vogel et al., 2005). Extensions to such model for constant pitch note were proposed in (Cont, in press) to account for multiple pitches in a single event such as *trills*. Although these works embedded temporal structures, the modeling still relied on the notion of successive steady values instead of accounting for a specific shape.

In this paper we propose to model sound descriptor profiles using a statistical framework called segmental models (SMs). SMs are a generalization of HMMs that address three of their principal limitations: (1) weak duration modeling, (2) assumption of conditional independence of observations given the state sequence and (3) restrictions on feature extraction imposed by frame-based observations (Ostendorf et al., 1996). In contrast to HMMs, SMs provide explicit state duration distributions, explicit correlation models and use segmental instead of frame-based features. Specifically, SMs allow for the adequate modeling of temporal profiles with structural flexibility (Ostendorf et al., 1996) and permit to incorporate expert knowledge at any structural level of the model. SMs, previously proposed for pattern matching (Ge and Smyth, 2000) or for handwriting modeling in (Artières et al., 2007) reveal to be effective when only partial training data is available. Furthermore, the explicit modeling of the duration proved to increase robustness to noisy conditions (Morris et al., 2002).

If SMs have been used and implemented in the audio community in the context of speech processing (Deng et al., 1994), they

* Corresponding author. Fax: +33 1 44 78 15 40.
E-mail address: julien.bloit@ircam.fr (J. Bloit).

0167-8655/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2009.11.003

were much less exploited to our knowledge for non-speech sound. In a previous study, we investigated preliminary aspects of the modeling power of SMs (Bloit et al., 2009). In this paper we report on a general methodology using SMs for the modeling of temporal sound features that can be related to a symbolic level. In particular, the modeling of music elements such as *crescendo*/*decrescendo*, *glissando*, *tremolo* and *sforzando* within a continuous stream is demonstrated. Nevertheless, as described, the approach we propose actually allows the user to build its own vocabulary using simple procedures.

This paper is structured as follows. First, we describe the segmental method and explain key differences with standard HMM models. Second, we describe the method for building the vocabulary made of musical units. Finally, we report and discuss results obtained on violin recordings.

2. Modeling framework

In this section we first present the segmental model (SM), which can be considered as an extension of standard HMMs. Second, we discuss important differences between HMMs and SMs.

2.1. Segmental model

In a segmental model, we consider a joint distribution $p(y_1^T, l_1^N, q_1^N)$ over observation frames $y_1^T = [y_1, y_2, \ldots, y_t, \ldots, y_T]$, a sequence of hidden states $q_1^N = [q_1, q_2, \ldots, q_n, \ldots, q_N]$ with segment durations $l_1^N = [l_1, l_2, \ldots, l_n, \ldots, l_N]$, where N is the number of units in a state sequence, and T is the number of observed D-dimensional frames ($y_t \in \mathbb{R}^D$). The K possible states $\mathscr{S} = \{s_1, s_2, \ldots, s_k, \ldots, s_K\}$ usually can be seen as symbolic units. In our case, these are the states corresponding to the musical elements that the user chooses to model.

The hidden layer dynamics is modeled with three distributions:

1. a state prior distribution $\pi(i) = p(q_1 = s_i)\ \forall s_i \in \mathscr{S}$,
2. a state-duration distribution $p(l|s)$, where l belongs to a finite set of durations $\mathscr{L}$,
3. a transition probability matrix A with elements defined as $a_{ij} = p(q_n = s_i, q_{n+1} = s_j)\ \forall (i,j) \in [1, \ldots, K]^2$.

The state dynamics is semi-Markov, i.e. the state-distribution could be governed by any chosen law. In comparison, the exponential duration law in HMMs usually favors shorter state durations (Tóth and Kocsor, 2005).

The observation distribution differs from an HMM as illustrated on Fig. 1. Instead of emitting a single observation, a segmental state q_n with duration l_n emits a segment $y_{t_n - l_n + 1}^{t_n}$ with probability $p(y_{t_n - l_n + 1}^{t_n} | l_n, q_n)$ where t_n is the ending time of the nth segment. As a consequence, there is possibly fewer elements in q_1^N than in y_1^T.

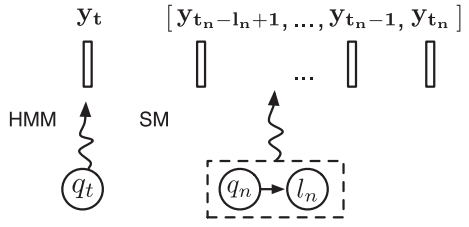

Fig. 1. Observation emission process for a single state. The HMM state emits one observation frame. A segmental model state emits an observation sequence, which distribution is also conditioned on segment duration l_n.

Considering this formulation, the joint distribution over the model variables can be expressed as:

$$p(y_1^T, l_1^N, q_1^N) = p\left(y_1^T | l_1^N, q_1^N\right) p\left(l_1^N | q_1^N\right) p(q_1^N). \tag{1}$$

The segmental model makes the assumption of independence between observation segments given state and duration pairs. For the first term in (1) this yields the following expression:

$$p\left(y_1^T | l_1^N, q_1^N\right) = \prod_{n=1}^{N} p(y_{t_n - l_n + 1}^{t_n} | l_n, q_n). \tag{2}$$

Moreover, the assumption of state conditional independence between successive segments durations allows to express the second term in (1) as:

$$p(l_1^N | q_1^N) = \prod_{n=1}^{N} p(l_n | q_n). \tag{3}$$

Finding the most likely state sequence for an observation sequence involves searching a state space that has a size of approximately $|\mathscr{S}| \times |\mathscr{L}|$ (*approximately* only because we can define a different distribution support for each distinct state in $\mathscr{S}$). This is done by finding

$$\left(\widehat{N}, \widehat{q_1^N}, \widehat{l_1^N}\right) = \underset{N, q_1^N, l_1^N}{\operatorname{argmax}}\, p\left(y_1^T | l_1^N, q_1^N\right) p\left(l_1^N | q_1^N\right) p(q_1^N), \tag{4}$$

which explores every possible segmentation. This is solved with an extension of the Viterbi algorithm (Deng et al., 1994; Russell, 2005), with an additional search dimension spanning every possible duration in the decoding lattice.

2.2. Motivations

The following example points the specificity of the segmental modeling framework compared to a classic HMM. From its description, we intend to give a qualitative insight on our motivations. Consider the classes $\{A, B, C\}$ shown in Fig. 2 characterized by specific temporal profiles, each being segmented in three successive parts: an initial constant phase, a transition step, and a final constant phase. Given the clear 3-phase structure of each class, it seems straightforward to build a three-state left-to-right model for each class.

Let us consider $D(s1, s2)$, the probabilistic divergence between two states. In the HMM case, the discrimination of class A and B would essentially rely on the high values given by $D(a1, b1)$ and $D(a3, b3)$, and much less on the $D(a2, b2)$ since the data in both states span over the same interval. For the same reasons, B and C could be easily discriminated. However, the discrimination between A and C would be more problematic since their only distinctive feature lies in their middle phase: a single HMM state for this phase would not suffice to discriminate between A and C, since they would both share similar distribution centered on close mean values. To solve this within a standard HMM framework, a significant larger number of states should be necessary to model the middle phase of the C model. However, the use of segmental states could discriminate between A, B and C, since the states would model the entire segment profiles and not only segment mean values.

3. Method

We present the followed methodology on a violin phrase dataset. Details on the set of primitives, and their associated duration distributions are given. We also explain how the primitives are assembled together to model elements of a musical vocabulary before presenting the evaluation procedure.

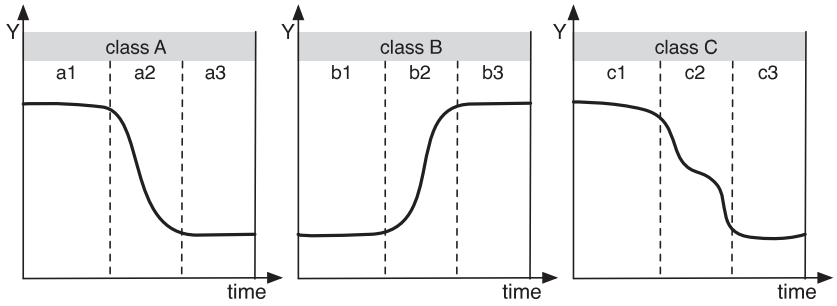

Fig. 2. Three classes $\{A, B, C\}$ defined by their specific time profiles, and their annotation into sub-unit, annotated as $[x1, x2, x3]$ for a given class X.

3.1. Dataset

We recorded a dataset with musical elements selected for their specific temporal profiles. Precisely, we defined a musical vocabulary composed of four pitch profiles (Fig. 3a),

- P_1: *upward glissando*,
- P_2: *downward glissando*,
- P_3: *constant pitch*,
- P_4: *tremolo*,

and four intensity profiles (Fig. 3b),

- I_1: *crescendo*,
- I_2: *decrescendo*,
- I_3: *constant intensity*,
- I_4: *sforzando*.

Crescendi (resp. *glissandi*) correspond to a continuous intensity (resp pitch) increase from one level to another. *Sforzando* consists in a sharp attack, a short decay followed by a longer sustain. *Tremolo* corresponds to two pitches alternating very rapidly.

We generated short music sketches by randomly combining these vocabulary elements. Each sketch contains two randomly chosen pitches, in a four-beat score. Each beat consists in a combination of one intensity profile I_i and one pitch profile P_j. No global dynamic levels (e.g. *piano* or *fortissimo*) are imposed but only dynamic variations (*crescendi* and *decrescendi*). Fig. 3c shows an example of such a generated music sketch.

The generated scores were interpreted by a violin player at a fixed tempo of 25 bpm. We recorded 25 sketches containing the various pitch and intensity profiles in random proportions. From the recording (sampled at 44,100 Hz), we extracted two sound descriptors, highly correlated to the musical notions of pitch and intensity, namely fundamental frequency (de Cheveigné and Kawahara, 2002) and loudness (Moore et al., 1997). These descriptors were computed on 46.4 ms windows, with a hop size of 5.8 ms, yielding an approximate frame rate $fr = 172$ Hz for the descriptors data. In order for the considered pitch profiles to be shift-invariant along the frequency axis, we use a logarithmic scale (unit: *cents*).

3.2. Set of primitives

We designed a set of primitives as elementary units of the vocabulary elements. One advantage of the proposed methodology is precisely to leave to the user the choice of these primitives. This choice allows one to emphasize particular profiles, that could be related to notation or semiotic level. In our case, our choice (see Fig. 4) was motivated by works by Bootz and Hautbois where they introduce a set of parametric temporal patterns (Bootz and Hautbois, 2007) related to Temporal Semiotic Units (Frey et al., 2008). We devised the primitives as follows:

- f_1: constant horizontal,
- f_2: linearly increasing,
- f_3: linearly decreasing,
- f_4: impulse up,
- f_5: impulse down,
- f_6: bell shape.

3.3. Model topology

The primitives were concatenated to build profile models of the vocabulary elements. Formally, this consisted in defining a topology for the segmental model. The pitch and intensity profile classes were therefore defined as follows:

- P_1 and I_1: left–right topology with primitives $\{f_1; f_2; f_1\}$.
- P_2 and I_2: left–right topology with primitives $\{f_1; f_3; f_1\}$.
- P_3 and I_3: single state with primitives $\{f_1\}$.
- P_4: left–right topology with primitives $\{f_4; f_1; f_5; f_1\}$.
- I_4: left–right topology with primitives $\{f_6; f_1\}$.

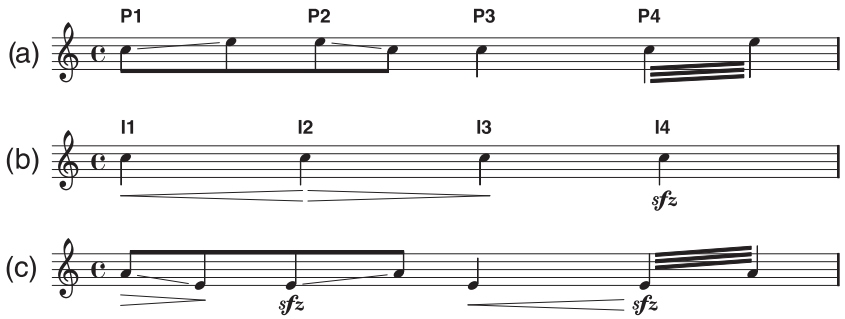

Fig. 3. Vocabulary of four profiles on the pitch dimension (a) and four profiles on the intensity dimension (b). A sketch example combines these profiles (c).

Fig. 4. Set of primitives.

Such a topology is relatively intuitive to define. For pitch and loudness *glissando* classes $P_{1,2}$ and $I_{1,2}$, the succession of primitives $\{f_1;f_2;f_1\}$ model the continuous progression from one pitch (resp. intensity level) to another pitch (resp. intensity level), while P_3 and I_3 represent a constant period. For pitch class *tremolo* P_4, we interleaved short constant segments between primitives f_4 and f_5 to authorize some temporal flexibility. Since the state duration distribution is explicitly defined by a non-exponential distribution, we did not allow self-transitions for states.

The profile models were further merged into a global profile-sequence model, with a topology allowing transitions between all the profile models (see Fig. 5). A uniform prior distribution was set for states entering a profile model. Notice that two independent profile-sequence models were built for pitch and intensity, in order to allow different segmentations depending on the considered musical dimension.

3.4. Model distributions

In order to capture the shape of a primitive, along with its characteristic duration, we used a trajectory model to represent the state conditional observation probability (right term in Eq. 2). For a given state s_k with duration l, the trajectory model is a sequence of l Gaussians with means μ_1^l sampled on a primitive shape, and a shared variance term σ_k. The observation likelihood is approximated, as done in (Artières et al., 2007):

$$-logp(y_1^l|l,s_k) = \frac{1}{2}\sum_{i=1}^{l}\frac{(y_i-\mu_i)^2}{\sigma_k^2}. \quad (5)$$

Because we wanted to recognize shapes without depending on their absolute spatial values, we modeled the data's first order time derivative

For the state duration distributions, we simply set a uniform distribution defined on a set of possible durations $\mathscr{L}_k$ for each state. This avoids to bias the segment duration towards a shorter durations as it would be the case with an exponential law. It also allows the user to decide which segments are likely to be stretched or not. This is crucial for musical data where the specific structure of a profile calls for non-linar stretching along its primitives. For example, the cases of vibrato, glissando and ornement typically follow different schemas when stretched: for the vibrato, more cycles should be added to the vibrato, for the glissando, the whole shape should be stretched while for the ornement, the main note should be stretched but not the ornement (Desain and Honing, 1992).

The model parameters where manually set by observation on a single example for each profile, outside the testing database. The primitives $f_{1,2,3}$ are characterized by their constant slope. Observation distribution for their corresponding states in the global model were thus set with a unique value for μ_1^l.

Primitives $f_{4,5,6}$ were defined by sampling a representative example. In contrast to the previous primitives, we wanted here to benefit from fine shape details of actual sampled data.

Primitives were assigned duration distributions and linearly resampled according to each possible length. In our case, we chose to allow for a relatively large duration range for $f_{1,2,3}$ ([0.230; 2.5] s). The range distribution for the impulse primitives $f_{4,5}$ is short ([0.025; 0.075] s), which was consistent with the fact they essentially model sharp transitions. For f_6, the duration range is medium, i.e. between in the other mentioned cases ([0.375; 1.12] s).

3.5. Evaluation tasks

The proposed approach was evaluated on two independent recognition tasks performed on the recorded dataset:

- task *T1*: recognize the four pitch profiles within a sketch.
- task *T2*: recognize the four intensity profiles within a sketch.

As such, the recognition tasks included two sub-tasks, namely segmentation of the classes within a continuous stream and classification. Each sketch is decoded twice: once with the pitch model, and once with the loudness model (Fig. 5). The decoding is achieved through maximizing the most likely state sequence (see Eq. 4). The resulting sequence of state-index q_1^N and durations l_1^N are subsequently mapped to the corresponding profile labels and associated time tags. This information is then handled to the evaluation step.

Evaluation is carried out using segmentation metrics defined for the MIREX 06 Score Following contest (Cont et al., 2007), as well as usual phone recognition metrics as presented in (Young et al., 2006). Each sketch in the test set was manually annotated with time tags and profile name labels as a reference. The evaluation metrics are computed as follows. A one-second tolerance window is centered on the reference time tag. For each reference segment, we attribute a detection label depending on what is detected inside the tolerance window: if a new event with the correct label is detected, we register a *hit* (H); if it has a wrong label, we register a *substitution* (S). If the correct label start within the reference segment region, but outside the tolerance interval, it is a *late* (L), and if no label starts within the tolerance interval, we register a

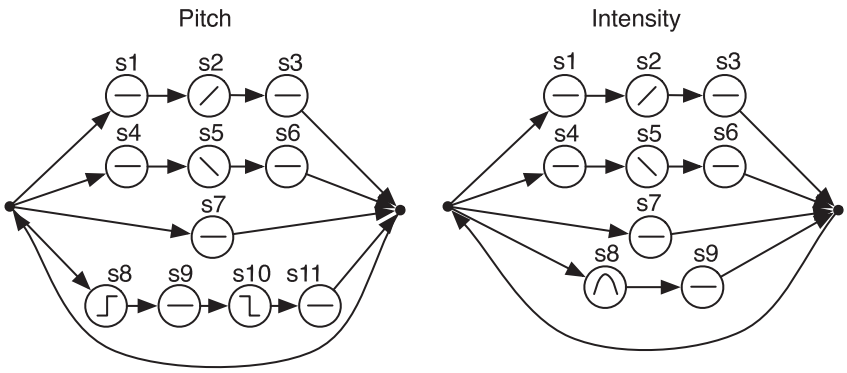

Fig. 5. Global topology for the sequence model of pitch profiles, and the sequence model of intensity profiles.

deletion (D). Any additional event starting during the reference was counted as an *insertion* (I).

4. Results

Typical recognition results are illustrated on Fig. 6 for one excerpt of a recorded sketch. In this figure, segmentation results are displayed for both pitch and intensity profiles. This sequence directly indicates the two levels of modeling we set: the level of primitives (dashed lines), and the level of classes (solid lines), grouping the primitives and corresponding to symbolic music elements we wanted to model. Precisely, each primitive expresses different phases of the class. On this particular example, we can see that the *crescendo* (class I_1) has a longer increase phase with shorter constant phases, while an opposite behavior is found for pitch (class P_1). The detailed view for the pitch class P_4, displayed on Fig. 7, illustrates the ability to capture the periodic aspect of this class, indicated by the clear repeated sequence of states $\{s_8; s_9; s_{10}; s_{11}\}$.

We found that the segmentation was relatively robust to various artefacts that were present in the descriptor profiles. For example, loudness values were drastically altered by the note tremolo occurring simultaneously, which was supposed to appear only in the pitch estimation (classes I_1 and P_4 on Fig. 6). Such an effect however did not influence the segmentation and recognition. This shows that an adequate choice of primitives can adequately add robustness. Pitch estimation also suffers from artefacts, especially on class P_4. Nevertheless, the SMs were able to segment correctly, excepted for large errors which forces the system to insert incorrect elements.

The evaluation metrics are reported for all tested sequences on Fig. 8. They show that classes were correctly recognized in 86% of the cases for pitch and 63% of the cases for loudness. For the pitch profiles, occasional substitutions, deletions and late detections were found (resp. 2%, 3% and 9%). However, the number of insertions was relatively important ($mean_I = 82.5\%$). As already mentioned above, pitch estimation errors typically led to insertions, especially at note onsets where pitch values are largely erroneous.

Interestingly, the analysis actually reveal some interpretation features that were not expected by our choice of the class structure. For example, when playing two successive upward *glissandos* (P_1), the violinist would perform a short downward *glissando* in between, not indicated in the score. Such an effect clearly appears in the signal and was correctly detected and labeled by the segmental model (see P_2 insertion on Fig. 6).

Results on loudness classes exhibited more errors ($mean_L = 5.5\%$, $mean_D = 9.5\%$, $mean_I = 12.5\%$). More substitutions were however performed ($mean_S = 22\%$) than for pitch. However, substitutions were dependent on the sequence as can be seen on Fig. 8. These errors can be explained by the difference between the score and the instrumentalist performance: for instance, the preparation for a

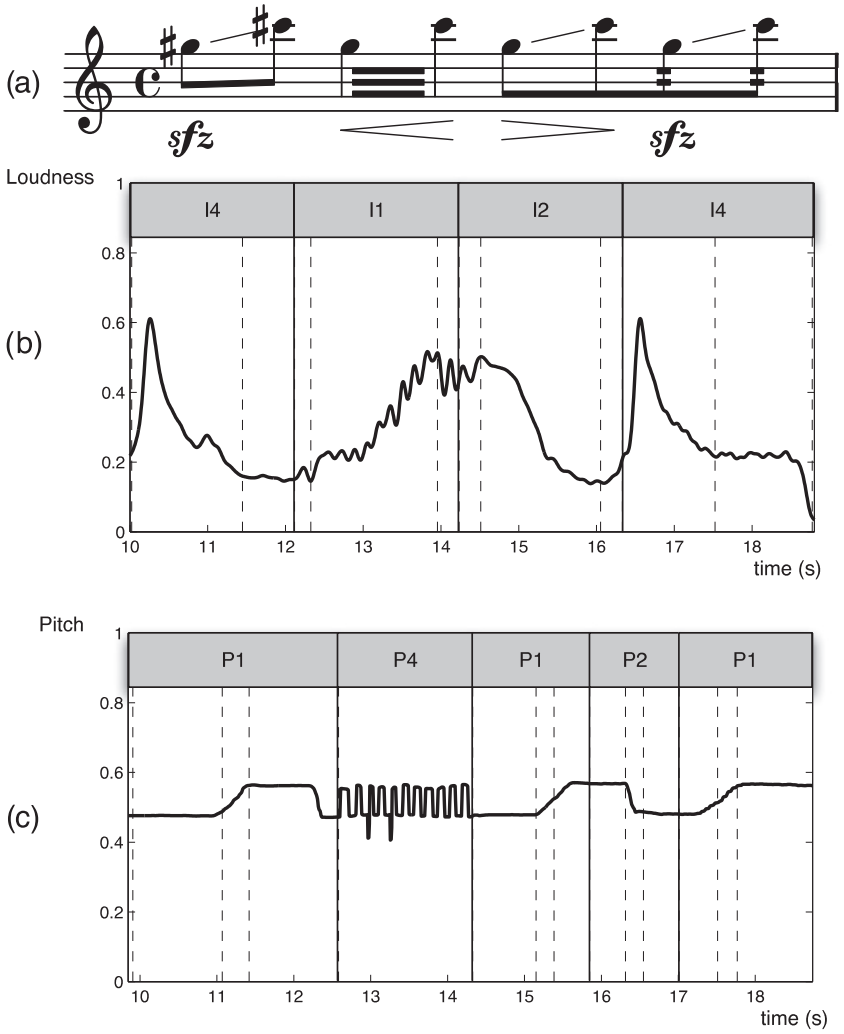

Fig. 6. A four-beat music sketch (a) with its resulting segmentation into intensity profiles (b) and pitch profiles (c).

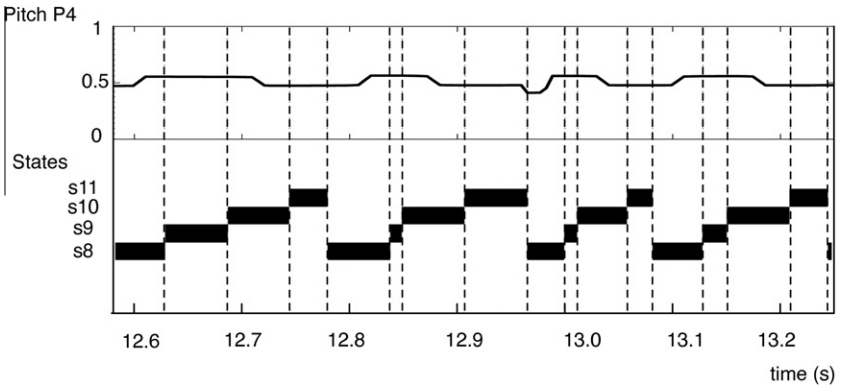

Fig. 7. A detail view from the decoded state sequence for pitch profile P4 from Fig. 6c.

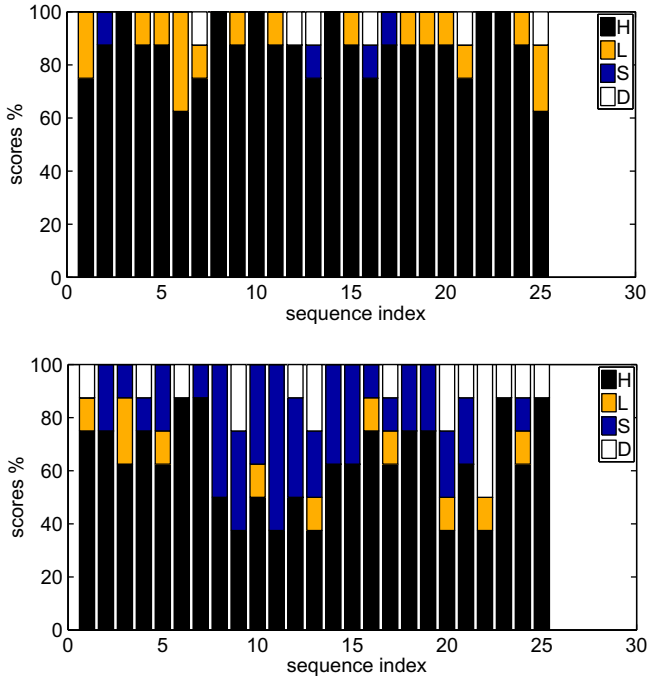

Fig. 8. Stack view for evaluation metrics of all sequences (with a one-second tolerance interval), as a percentage related to the number of reference profiles per sequence: (a) scores on task $T1$; and (b) scores on task T2 (H = hit, L = late, S = substitution, D = deletion).

sforzando may imply an intensity raise of the previous stroke, whatever class it may pertain to.

5. Conclusion

In summary, we proposed a method to model sound descriptors as temporal profiles, using a modeling framework based on segmental model, still rarely exploited for non-speech audio. Actually, our approach is well suited for morphological sound description.

We described our methodology for modeling and segmenting in a particular case of violin phrases, and carried on an evaluation. The results showed promising results, while indicating specific limits of the current model. This suggested future refinements of the global sequence model, simply built by assembling elementary profile models at the signal level. For example, additional context modeling could be taken into account. Also, these two sound descriptors were considered as independent streams, while they both relate to the same performance. Fusing the data streams could yield better segmentation results, by relying on data correlations, or compensating for noise in one of the streams. Besides, we could apply this modeling framework to multidimensional descriptors. Extending the segmentation towards an online version could also be considered with a short-time Viterbi algorithm (Bloit et al., 2008).

Finally, although no statistical learning was used to optimize the model parameters, the evaluations demonstrated that the directly interpretable model structure allows for user-centric definition of the model primitives. Such a flexibility represents a promising feature of this method.

Acknowledgements

We would like to acknowledge Thierry Artières and Xavier Rodet, Norbert Schnell and Pierre Machart for fruitful discussions. This work has been partially supported by the European Commission 7th Framework Programme SAME Project (No. 215749) and the ANR project Interlude.

References

Artières, T., Marukatat, S., Gallinari, P., 2007. Online handwritten shape recognition using segmental hidden Markov models. IEEE Trans. Pattern Anal. Machine Intell. 29, 205–217.

Bevilacqua, F., Rasamimanana, N., Fléty, E., Lemouton, S., Baschet, F., 2006. The augmented violin project: Research, composition and performance report. In: Proc. Internat. Conf. on New Interfaces for Musical Expression (NIME).

Bloit, J., Rodet, X., Mars, 2008. Short-time Viterbi for online HMM decoding: Evaluation on a real-time phone recognition task. In: ICASSP. Las Vegas.

Bloit, J., Rasamimanana, N., Bevilacqua, F., 2009. Towards morphological sound description using segmental models. In: DAFX. Como, Italy.

Bootz, P., Hautbois, X., 2007. Les motifs temporels paramétrés [configuration of temporal motives]. In: Rix, E., Formosa, M. (Eds.), Vers une sémiotique générale du temps dans les arts, Actes du colloque les Unités Sémiotiques Temporelles (UST), nouvel outil d'analyse musicale: Théories et Applications [Toward a General Semiotics of Time in Arts. Proceeding of Temporal Semiotic Units (TSU), New Tools for Musical Analyses: Theory and Applications]. Delatour-France, Sampzon, pp. 139–176.

Chion, M., 1983. Guide des objets sonores. INA/GRM, Buchet/Chastel.

Cont, A., in press. A coupled duration-focused architecture for realtime music to score alignment. IEEE Trans. Pattern Anal. Machine Intell..

Cont, A., Schwarz, D., Schnell, N., Raphael, C., 2007. Evaluation of real-time audio-to-score alignment. In: Internat. Symp. on Music Information Retrieval (ISMIR). Vienna, Austria.

de Cheveigné, A., Kawahara, H., 2002. Yin, a fundamental frequency estimator for speech and music. J. Acoust. Soc. Amer. 111 (4), 1917–1930.

Deng, L., Aksmanovic, M., Sun, X., Wu, C., 1994. Speech recognition using hidden Markov models with polynomial regression functions as nonstationary states. IEEE Trans. Speech Audio Process. 2 (4), 507–520.

Desain, P., Honing, H., 1992. Time functions function best as functions of multiple times. Comput. Music J. 16 (2), 17–34. <http://www.jstor.org/stable/3680713>.

Frey, A., Marie, C., Prod'Homme, L., Timsit-Berthier, M., Schön, D., Besson, M., 2008. Temporal semiotic units as minimal meaningful units in music? An electrophysiological approach. Music Perception 26 (3), 247–256.

Ge, X., Smyth, P., 2000. Deformable Markov model templates for time-series pattern matching. In: SIGKDD. ACM Press, pp. 81–90.

Ghahramani, Z., 2001. An introduction to hidden Markov models and bayesian networks. Internat. J. Pattern Recognition Artif. Intell. 15, 9–42.
Grey, J., 1977. Multidimensional perceptual scaling of musical timbres. J. Acoust. Soc. Amer. 61, 1270–1277.
Kaltenecker, M., 2001. Avec Helmut Lachenmann. Van Dieren, Paris.
Mcadams, S., Winsberg, S., Donnadieu, S., De Soete, G., Krimphoff, J., 1995. Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. Psychol. Res. (58), 177–192.
Möller, M., 2009. New sounds for flute. <http://www.sfz.se/flutetech> (accessed 05.04.09).
Moore, B., Glasberg, B., Baer, T., 1997. A model for the prediction of thresholds, loudness, and partial loudness. J. Audio Eng. Soc. 45 (4), 224–240.
Morris, A., Payne, S., Bourlard, H., 2002. Low cost duration modelling for noise robust speech recognition. In: ICSLP. pp. 1025–1028.
Ostendorf, M., Digalakis, V., Kimball, O.A., 1996. From HMMs to segment models: A unified view of stochastic modeling for speech recognition. IEEE Trans. Speech Audio Process. 4, 360–378.
Peeters, G., Deruty, E., 2008. Automatic morphological description of sounds. In: Acoustics 08. SFA, Paris.
Rabiner, L., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. In: Proc. IEEE. pp. 257–286.
Raphael, C., 1999. Automatic segmentation of acoustic musical signals using hidden Markov models. IEEE Trans. Pattern Anal. Machine Intell. 21 (4), 360–370.
Ricard, J., Herrera, P., 2004. Morphological sound description computational model and usability evaluation. In: AES Convention.
Russell, M., 2005. Reducing computational load in segmental hidden Markov model decoding for speech recognition. Electron. Lett. 41 (25), 1408–1409.
Ryynänen, M., Klapuri, A., 2008. Automatic transcription of melody, bass line, and chords in polyphonic music. Comput. Music J. 32 (3), 72–86.
Schaeffer, P., 1966. Traité des objets musicaux. Seuil.
Schaeffer, P., Reibel, G., Ferreyra, B., 1967. Solfège de l'objet sonore. INA/GRM.
Smalley, D., 1997. Spectromorphology: Explaining sound-shapes. Org. Sound 2, 107–126.
Tóth, L., Kocsor, A., 2005. On naive Bayes in speech recognition. Internat. J. Appl. Math. Comput. Sci. 15 (2), 287–294.
Vogel, B., Jordan, M., Wessel, D., 2005. Multi-instrument musical transcription using a dynamic graphical model. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing (ICASSP).
Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., 2006. The HTK Book (for HTK Version 3.4). Cambridge University Engineering Department.