FROM BOULEZ TO BALLADS: TRAINING IRCAM'S SCORE FOLLOWER

Diemo Schwarz *Arshia Cont* *Norbert Schnell*

Ircam–Centre Pompidou, 1, place Igor-Stravinsky, 75003 Paris, France

ABSTRACT

This paper describes our attempt to make the *Hidden Markov Model (HMM)* score following system developed at Ircam sensible to past experiences in order to obtain better audio to score real-time alignment for musical applications. A new observation modeling based on Gaussian Mixture Models is developed which is trainable using a learning algorithm we call *automatic discriminative training*. The novelty of this system lies in the fact that this method, unlike classical methods for HMM training, is not concerned with modeling the music signal but with correctly choosing the sequence of music events that was performed. Besides obtaining better alignment, new system's parameters are controllable in a physical manner and the training algorithm learns different styles of music performance as discussed. Experience with the piece *...explosante–fixe..* by Boulez, and with an advanced karaoke system that allows to sing ballads in free tempo with automatic accompaniment are given.

1. INTRODUCTION

Score following is the real-time alignment of a known musical score to the audio signal produced by a musician playing this score in order to synchronise the electronic part of the music to the performer, leaving him with all possibilities of expressive performance.

For an introduction and state of the art on score following and details of the system developed by Ircam's Real-Time Applications team, see [8]. A review of past attempts in score following literature, focusing on the adaptability and learning aspects of the algorithms, specially of importance for our work, is given in [3].

In this paper, we introduce a new learning algorithm used for Ircam's score follower. This training is used to adapt the follower to a certain instrument, or musician, or even a certain movement in a piece. It is also used to train on unexperienced singers in an advanced karaoke system that allows to sing ballads in free tempo with automatic accompaniment.

Section 3 gives an overview of our approach and objective for training leading to a new *observation modeling* for score following. After reviewing the proposed architecture in section 4, we introduce a learning algorithm called *automatic discriminative training* in section 5 which conforms to the practical criteria of a score following system. The novelty of this system lies in the fact that this method, unlike classical methods for HMM training, is not concerned with modeling the music signal but with correctly choosing the sequence of music events that was performed. In this manner, using a *discrimination* process we attempt to model class boundaries rather than constructing an accurate model for each class. Finally, in section 6 we demonstrate some results and evaluations of the new system and relate the experiences we had regarding a live orchestral performance with the new system for a piece by Boulez, and with the advanced karaoke application.

2. RELATED WORK

Probabilistic or statistical score followers, including the concept of training, are first described in [6]. The probability density functions (PDFs) should be obtained in advance and are good candidates for an automatic learning algorithm. Three different PDFs are used and alternative methods to obtain them are defined, using information based on intuition and experience, and information based on empirical investigations of actual performances. A total of 20 recorded performances were used and their pitch-detected and *hand-labeled* time alignment is used to provide an observation distribution for actual pitch given a scored pitch and the required *PDF*s are calculated from these hand-discriminated data.

In the HMM score following system [10], statistics (or features in our system's terminology) are trained using a *posterior marginal distribution* $\{p(x_k|\mathbf{y})\}$ to re-estimate the feature probabilities in an iterative manner. In this iterative training *signatures* assigned to each frame are used for discrimination but no parsing is applied beforehand.

3. APPROACH

Training in the context of score following is to adapt its parameters to a certain style of performance and a certain piece of music. We envision a system which adapts itself to correct parameters using a database of sound files of previous performances of the same piece or in the case of a creation, of recorded rehearsals. After the offline and automatic learning, the system is adapted to a certain style of performance, and thus provides better alignment with the score in real-time.

Figure 1 shows a general diagram of Ircam's score follower. The work presented here refines the *observation modeling* (upper block). The *decision and alignment block* (lower block) is described in detail in [7].

4. OBSERVATION MODELING

Observation in the context of our system consists of calculating features from the audio spectrum in real-time and associate the desired probabilities for low-level HMM states. Low-level states in our system are *attack*, *sustain* and *rest* for each note in the score. Spectrum features are *Log of Energy*, *Spectral Balance* and *Peak Structure*

[*] This work was done in the framework of the project SemanticHIFI, funded by the European Commisson. See http://shf.ircam.fr/.

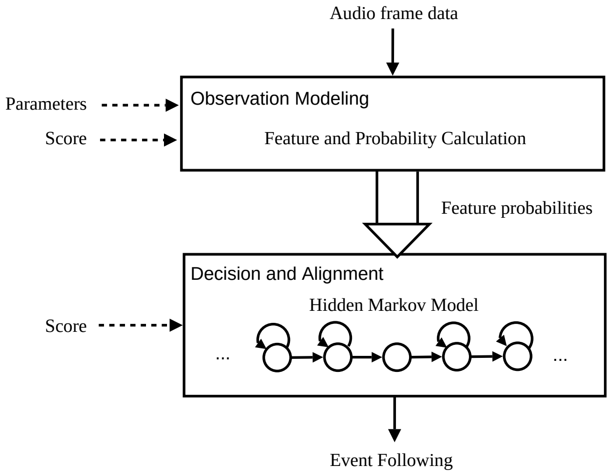

Figure 1. General diagram of the score following system

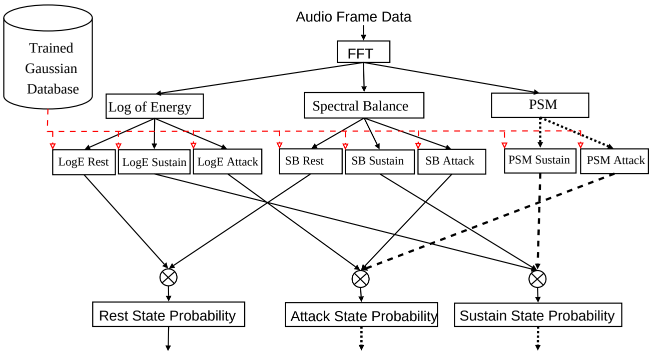

Figure 2. Probability observation diagram

Match (PSM). We will not go into implementation details of the mentioned features which are described in [1, 8, 9], but focus on the learning aspect of the architecture.

The observation process can be seen as a dimension reduction process where a frame of our data, or the FFT points, lies in a high dimensional space, $\Re^J$ where J is 2048. In this way, we can consider the features as vector valued functions, mapping the high dimensional space into a much lower dimensional space, or more precisely to $2 + N$ dimensions where N is the number of different notes present in the score for the PSM feature. Another way to look at the observation process is to consider it as a probability mapping between the feature values and low-level state probabilities. A diagram of the observation process is demonstrated in figure 2.

In this model, we calculate the low-level feature probabilities associated with each feature which in terms are multiplied to obtain a certain low-level state feature probability. As an example, the *Log of Energy* feature will give three probabilities *Log of Energy for Attack*, *Log of Energy for Sustain* and *Log of Energy for Rests*.

In order to calculate probabilities from features, each of the 8 low-level state feature probabilities is using probability mapping functions from a database of stored trained parameters. They are derived from Gaussians in forms of *cumulative distribution functions (CDFs)*, inverse cumulative distribution functions or PDFs depending on the heuristics associated with each feature state. This architecture is inspired by Gaussian Mixture Models. Note that the dimension of each model used is one at this time.

By this modeling we have assumed that the low-level states' attributes are global which is not totally true and would probably fail in extreme cases. However, due to a probabilistic approach, training the parameters over these cases would solve the problem in most cases we have encountered. Another assumption made is the conditional independence among the features, responsible for the final multiplication of the feature as in Figure 2.

5. TRAINING THE SCORE FOLLOWER

In an ideal training, the system runs on a huge database of *aligned* sound files and adapts its parameters to the performance. In this case, the training is usually supervised and is dependent on system architecture. However, in a concert setup with rehearsals and performances, such an ideal procedure would not be possible [1]. In this context, the training will be offline and would use the audio data recorded during rehearsals to train itself.

5.1. The automatic discriminative training

In score following we are not concerned with estimating the joint density of the music data, but are interested in the posterior probability of a musical sequence using the acoustic data. More informally, we are not finally concerned with modeling the music signal, but with correctly choosing the sequence of music events that was performed. Translating this concern to a local level, rather than constructing the set of PDFs that best describe the data, we are interested in ensuring that the correct HMM state is the most probable (according to the model) for each frame.

This leads us to a *discriminative training* criterion. This criterion has been described in [11] among others. Discriminative training attempts to model the class boundaries —learn the distinction between classes— rather than construct as accurate a model as possible for each class. In practice this results in an algorithm that minimizes the likelihood of incorrect, competing models as well as maximizing the likelihood of the correct model.

While most discriminative training methods are supervised, for portability and practical issues, it should be automatic if not unsupervised. For this reason, we introduce an automatic supervision over training by constructing a *discrimination knowledge* by an algorithm which forces each model to its boundaries and discriminates feature observations.

Figure 3 shows a diagram of different steps of this training. The inputs of this training are audio files plus a music score. There are two main cores to this system: *Discrimination* and *Training*.

5.2. Discrimination

Using discrimination, we aim to distinguish low-level states in the feature domain. In this process, as part of the training, a set of states and their corresponding observations is obtained without actually segmenting or labeling the performance. The *Yin* algorithm [4] is used as the

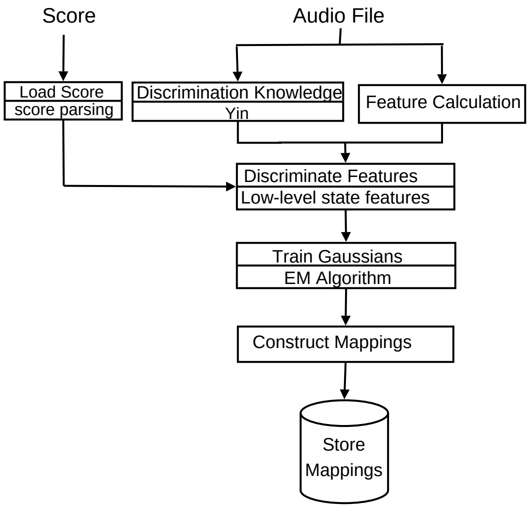

Figure 3. Automatic Discriminative Training Diagram

base knowledge. *Yin* is originally a monophonic fundamental frequency estimator and provides fairly good measures of aperiodicity, which discriminates *rest* and *note* events. If the detected note meets a minimum time length of about 20 frames, the neighbourhood of the starting index is marked as *attack*, and the rest as *sustain*.

Because of the noisiness of *Yin*'s measurement, it is not being used in the first place in the system itself. However, this noisiness will be covered during training due to the statistical nature of the algorithm.

This work is comparable to unsupervised model adaptation algorithms in speech where model parameters are adjusted on the basis of unlabeled training data by making a preliminary recognition. In this way, *discrimination knowledge* refers to unsupervised labeling of the audio file associated with HMM low-level states.

5.3. Training

Having all features discriminated, we are ready to train the *Gaussians*. We evade using fitting algorithms due to robustness issues and use an *EM Algorithm* [5] to construct the *Gaussians* on observed discriminated features.

The result of the training is a set of PDFs that correspond to each low-level state feature. This data is stored in a database which is used in the score follower's *observation* block as shown in Figure 2.

6. EVALUATION AND APPLICATIONS

Evaluation of a score following system is a wide topic. In [8], *objective* and *subjective* evaluation was discussed, suggesting a framework for evaluation of different existing systems. Figure 4 gives evaluation results for the proposed system. Overall, the stability of the system has increased and noisiness has been reduced, essential features for real-time following. Alignment is also improved in general and specially for fast phrases.

One important outcome of this learning algorithm is the ability to model and differentiate different styles of performance of a piece specific to different musicians, and to specific mouvements. For a detailed discussion of this feature, see [1, 2, 3].

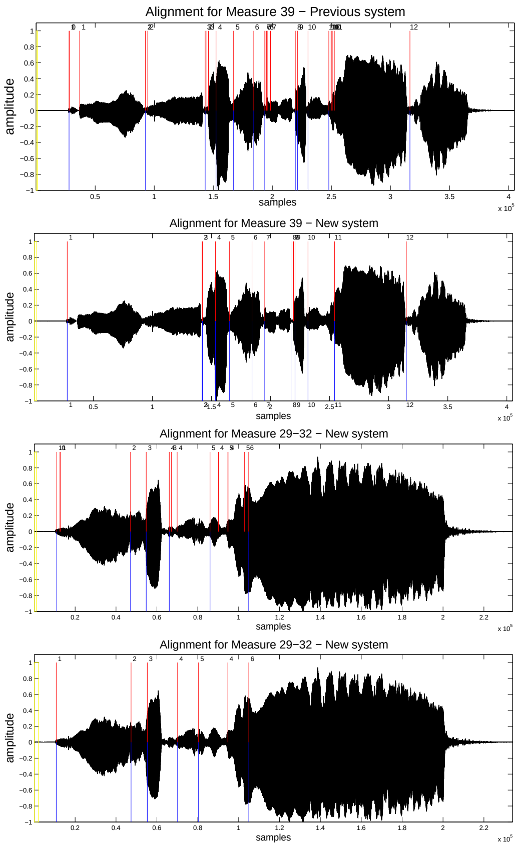

Figure 4. Alignment results of the previous system and new system, using the proposed observation block and training, on measure 39 and 29–32 of part I of *En Echo*. Vertical lines with numbers demonstrate the segmentations associated with the note number in the score.

6.1. Electroacoustic Music

Our score following system had its successfull première in concert with the piece *...explosante–fixe...* by Pierre Boulez for three flutes, orchestra, and electronics. Before, the piece used to be performed with a specially constructed Midi-flute that outputs the pressed keys when sound is present, but was very unwieldy and error prone. Now, the solo flutist can use her own instrument.

The performance of the score follower was almost perfect during rehearsals, but a little less in the concert. The remaining problems were early triggers in long pauses where the orchestra played louder than on the training recordings, two missed cues, the passages with lots of repeated notes, which were followed by hand, and one part with trills with a fifth jump, played very soft and breathy. Retraining was not done on the rehearsal recordings, since the follower performed sufficiently well and its errors were predictable and could easily be reacted to.

In one fast passage, the followed note information was directly used to drive a harmoniser, which proves that the follower provides a tight synchronisation.

6.2. Automatic Accompaniment of Ballads

In the framework of the SemanticHifi project* [13] we developed a demo application that allows users of a Hifi system of the future to interact with the system and its music collection. The score following application developed for this live demonstration allows for automatic synchronization of a pre-composed accompaniment to a hobby singer. The melody and the accompanying chords of the ballad *Autumn Leaves* were chosen for the demonstration. During the performance, melody is sung into a microphone connected to the system, which plays the accompaniment precisely synchronized to the singer's performance. The demonstration was a musically convincing experience for the singer as well as for the audience.

The application used in the demonstration mainly consists of the score following module and an accompaniment module connected to a General MIDI synthesizer. The score following module receiving the audio input from the singer continuously estimates the current position in the performed song, output as a cue number. This output is used by the accompaniment module to look up the chord sequence associated to the cue, which is sent to the MIDI synthesizer. The accompaniment module can be schematized as a finite state machine advancing in the pre-composed sequence of chords driven by the output of the score following module and an internal timer. The internal timer is adjusted to the singer's rhythm and allows advancing in the chord sequence in the case that the singer pauses (according to the score or by error).

In the case that the singer sings out of tune or an unexpected melody, the score following module adjusts as well as possible the output position in the song by waiting and advancing. The module turns out to perform robust musical accompaniment in a number of situations usually judged as difficult to handle for automatic accompaniment systems, such as the singer deviating from the score and singing out-of-tune. However, the robustness of the score following module to the latter has still to be enhanced.

7. FUTURE WORK

The robustness of the score following module to out-of-tune singing has still to be enhanced by replacing the spectral matching approach of PSM [9] with a measure that allows to calculate the match *and* the deviation from the correct pitch.

Training is implemented in *Matlab*, but clearly the next step is to integrate it into the Max patch, which would allow faster setup times for new pieces, less and more fluid rehearsals when a difficult section can be quickly re-trained during a break.

8. CONCLUSION

In this paper we presented a new approach for the *observation modeling* of our statistical HMM score follower which can articulate specific behavior of the musician in a controllable manner.

Using this approach, a learning algorithm called *automatic discriminative training* was implemented which conforms to the practical criteria of a score following system. The novelty of this system lies in the fact that this method, unlike classical methods for HMM training, is not concerned with modeling the music signal but with correctly choosing the sequence of music events that was performed. The proposed training is independent of the system's architecture and has led to improvements in real-time alignment. The system tends to model the margins of different styles of performance to a good extent and moreover, might be a point of departure for further studies in the context of learning algorithms for audio signal processing.

Our trainable score follower has proven its viability in a concert performance, and its flexibility in an automatic accompaniment application for ballads. It is implemented in Max/MSP using the FTM [1] enhancements [12].

9. ACKNOWLEDGMENTS

We are grateful to Philippe Manoury, Serge Lemouton and Andrew Gerzso without whose valuable comments and presence during test sessions the project could not have advanced. We would like to acknowledge Nicola Orio's ground laying work as the founder of this research project.

10. REFERENCES

[1] Arshia Cont. Improvement of observation modeling for score following. Master's thesis, University of Paris 6, Ircam, Paris, 2004.

[2] Arshia Cont, Diemo Schwarz, and Norbert Schnell. Training Ircam's score follower. In *AAAI Fall Symposium on Style and Meaning in Art, Language and Music*, 2004.

[3] Arshia Cont, Diemo Schwarz, and Norbert Schnell. Training Ircam's score follower. In *ICASSP*, 2005.

[4] Alain de Cheveigne and H. Kawahara. YIN, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.*, 111:1917–1930, 2002.

[5] A.P. Dempster, N. M. Laird, and D. B. Rubin. maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society*, 39(B):1–38, 1977.

[6] L. Grubb and R. B. Dannenberg. A stochastic method of tracking a vocal performer. In *Proc. ICMC*, 1997.

[7] Nicola Orio and François Déchelle. Score Following Using Spectral Analysis and Hidden Markov Models. In *Proc. ICMC*, Havana, 2001.

[8] Nicola Orio, Serge Lemouton, Diemo Schwarz, and Norbert Schnell. Score Following: State of the Art and New Developments. In *New Interfaces for Musical Expression (NIME)*, Montreal, 2003.

[9] Nicola Orio and Diemo Schwarz. Alignment of Monophonic and Polyphonic Music to a Score. In *Proc. ICMC*, Havana, 2001.

[10] C. Raphael. Automatic Segmentation of Acoustic Musical Signals Using Hidden Markov Models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 21(4), 1999.

[11] S. Renals, N. Morgan, H. Bourlard, M. Cohen, and H. Franco. Connectionist probability estimators in HMM speech recognition. *IEEE Transactions Speech and Audio Processing*, 1993.

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

[13] Hugues Vinet. The SemanticHifi project. In *Proc. ICMC*, Barcelona, Spain, 2005.

[1] http://www.ircam.fr/equipes/temps-reel/maxmsp/ftm.html