Journal of New Music Research, 2014
Vol. 43, No. 3, 266–275, http://dx.doi.org/10.1080/09298215.2014.890630

Routledge
Taylor & Francis Group

Estimating Onset and Offset Asynchronies in Polyphonic Score-Audio Alignment

Johanna Devaney

The Ohio State University, USA

(Received 5 July 2013; accepted 29 January 2014)

Abstract

In order to use score-audio alignment to study expressive performance or guide signal processing algorithms, it is necessary to identify the onsets and offsets of all of the notes in sounded simultaneities (e.g. chords). This paper describes an algorithm to improve the accuracy of dynamic time warping-based score-audio alignment for monaural polyphonic recordings. The algorithm uses a multi-pass approach, where an initial dynamic time warping alignment is refined by a hidden Markov model to allow the identification of asynchronies between musical events that are notated as simultaneities in the score. By providing estimates of individual onsets and offsets in notated simultaneities, the multi-pass algorithm improves the median accuracy of the DTW alignment for individual onsets by 41 ms on average, from 118 to 77 ms, and for individual offsets by 6 ms on average, from 75 to 69 ms.

Keywords: performance, music alignment, dynamic time warping, hidden Markov models, voice

1. Introduction

Music alignment has been an active area of inquiry for over twenty-five years, although there remain a number of challenges to be addressed in terms of system accuracy and robustness. Work has been divided between the online problem of following a solo performer in real-time in order to generate a responsive accompaniment and the offline problem of matching a symbolic representation (such as MIDI) to a monaural recording of a polyphonic performance. Offline implementations have applications in digital library synchronization, signal processing, and expressive performance analysis. This paper presents an offline system that focuses on the challenge of measuring note onset and offset asynchronies between musical lines that are marked as simultaneities in the score, which is particularly important for signal processing and expressive performance applications. Research has shown that these asynchronies can range from 7 to 50 ms (Palmer, 1997). The current state of the art, as described in Section 2, is not sufficiently accurate to address this, especially for tones with non-percussive onsets, such as unfretted strings or the singing voice, which form two of the major types of ensembles in Western art music: string quartets and vocal ensembles/choirs.

After providing a summary of related in work Section 2, the paper details the specific challenges of identifying asynchronies between voices in monaural recordings of polyphonic performances in Section 3. Section 4 describes a novel technique for improving the accuracy of polyphonic score-audio alignment by using a multi-pass approach where a hidden Markov model is used to refine the offset–onset transitions identified by an initial dynamic time warping alignment between score and audio representations of the same music where at least one voice changes notes or re-articulates its current note. An evaluation of the algorithm and discussion of the results is presented in Section 5. Conclusions, potential applications of this algorithm, and possibilities for future work are presented in Section 6.

2. Background

Dynamic time warping (DTW) and hidden Markov models (HMMs) have been the most commonly used techniques for alignment (e.g. Orio and Schwarz (2001) for DTW and Cano, Loscos, and Bonada (1999) for HMMs), though recently more complex graphical models have been explored (e.g. Raphael, 2004). Dynamic programming, a technique for recursively

Correspondence: Johanna Devaney, School of Music, The Ohio State University, 1899 College Road, Columbus, OH 43210-1170, USA. E-mail: devaney.12@osu.edu

solving problems with a certain type of structure, was used in early score following (e.g. Dannenberg, 1984). DTW is a particular type of dynamic programming algorithm for aligning two sequences given a matrix of compatibilities between all pairs of elements in the sequences and rules for valid differences between the sequences. Graphical models take advantage of graph structure and the laws of probability to efficiently estimate model parameters and states. HMMs are a type of directed graphical model where a hidden state variable completely determines the distributions of both the observed variables at a moment in time and the next state in the sequence. DTW can be considered a constrained form of an HMM, where the state sequence always moves forward and each relative state transition has the same probability for all states. In contrast to an HMM, classic DTW lacks meaningful training procedures. Because of its constrained nature, DTW is better suited to offline applications where there is a known correspondence between the performance and the score as it is not as flexible as HMMs in dealing with performance errors in online contexts. Other approaches used in score following include particle filters (e.g. Otsuka, Nakadai, Ogata, & Okuno, 2011) and conditional random fields (CRFs) (e.g. Joder, Essid, & Richard, 2011). Particle filters, also known as sequential Monte Carlo methods, are related to HMMs in that they use Markov chains, but differ in that they use a continuous rather than discrete state space, which can be useful when modelling things like tempo. CRFs are similar to HMMs, but instead of learning to describe a sequence of observations using hidden state sequences, they learn to predict the probabilities of state sequences given an observation.

2.1 Online systems

The published history of score following began at the 1984 International Computer Music Conference (ICMC), where Dannenberg (1984) and Vercoe (1984) presented separate papers on the topic of automatic computer accompaniment of a live musician. Both Dannenberg and Vercoe were interested in creating accompaniment systems that were able to respond to live soloists. Vercoe's system was particularly motivated by a commission for a piece for flute and live electronics from IRCAM. A 'second generation' score follower developed by Baird, Blevins and Zahler (1990, 1993) segmented the musical score to improve the accompaniment systems' musicality. These early systems worked on symbolic representations of the performances and focused on either keyboards, which could generate symbolic data, or instruments with sensors affixed that could transmit symbolic data. Audio score following was motivated in the 1990s by the desire to automatically accompany the singing voice, which was addressed by both Puckette (1995) and by Grubb and Dannenberg (1997). In these early score following systems, either MIDI was transmitted directly from the instrument (Baird et al., 1990, 1993; Dannenberg, 1984; Vercoe, 1984) or a pitch-tracking algorithm was used to generate MIDI data (Puckette, 1995). The performance MIDI data was then aligned to a MIDI score, typically using string-matching techniques and/or dynamic programming until Stammen and Pennycook (1993) pioneered the use of a modified version of DTW for score following.

The use of stochastic methods in online music alignment was pioneered by Grubb and Dannenberg (1997), who used a stochastic approach for estimating a score position pointer in a vocal performance using probability density functions. Since the late 1990s, it has been more common for online systems to use HMMs and graphical models. Cano et al. (1999) used a left-to-right HMM-based approach for aligning monophonic music to model attack, sustain, release and silence states. Orio and Déchelle (2001) trained a multi-level HMM for polyphonic recordings that consisted of HMMs at both the song- and note-levels using the same features as Orio and Schwarz (2001). Raphael (2004) used a more generalized graphical model approach for score matching in order to address the problems that HMM-based systems have in modelling note duration. His two-level model was able to use the duration information in the score more effectively than a single-level model: one level modelled the pitch content in the signal and the other the notes and tempo-shifts.

Cont (2006) used hierarchical HMMs to model the notes, chords, and rests in the lower level and the temporal relationship between the lower-level events and the score in the upper level models. He later used this approach as the basis for his 'anticipatory' score following system, which makes predictions about the future to inform its current decision through the use of a hidden hybrid Markov/semi-Markov model (Cont, 2010). Other related approaches include the use of dynamic HMMs in a Bayesian framework for score position pointer estimation by Peeling, Cemgil and Godsill (2007), the use of particle filtering by Otsuka et al. (2011), and Montecchio and Cont (2011b) for estimating song pointer position and tempo simultaneously. Duan and Pardo (2011) presented a generalized model for polyphonic score following, also based on particle filtering, that did not require training for specific instruments.

Online systems require robust and accurate note location estimates, accounting for deviations from the score, such as ornamentation or mistakes, and low-computational cost, in order to provide low-latency. Typically, online systems consider estimates within 250 ms (Cont, 2010) to 300 ms (Cont, Schwarz, Schnell, & Raphael, 2007) of the actual note to be considered correctly aligned. Overall, while there are certain instruments for which score following is effective, there remain others for which it needs to be improved, especially those with non-percussive onsets (e.g. violins (Cont, 2010)).

2.2 Offline systems

Offline music alignment can be used for a number of applications, including digital library synchronization, signal processing, and expressive performance analysis. Digital music libraries contain both score-based and acoustic-based representations of music. Dunn, Byrd, Notess, Riley and Scherle

(2006), and Damm, Fremerey, Kurth, Müller and Clausen (2008) synchronized these representations with one another to create a multi-modal browsing experience using music alignment. Such synchronization allowed for both fast indexing of the recordings and a multi-modal experience where the score position is shown in real-time while the recording plays. Offline music alignment has been used to create a reference for signal processing algorithms (Basaran,Cemgil, & Anarım, 2011; Dannenberg, 2007; Montecchio & Cont, 2011a; Smit & Ellis, 2009; Woodruff, Pardo, & Dannenberg, 2006). It is also useful in expressive performance studies because of its ability to achieve greater precision in identifying note onsets and offsets than blind estimation algorithms, an application idea first described by Scheirer (1995) and more recently employed in toolkits by Dixon and Widmer (2005) and Devaney, Mandel and Fujinaga (2012).

As noted above, DTW is commonly used in offline systems. Orio and Schwarz (2001) used DTW for monophonic and polyphonic music alignment by modelling note attacks and silences. Soulez, Rodet and Schwarz (2003) subsequently improved the robustness of the algorithm by also modelling note sustains. Müller, Kurth and Roder (2004) later implemented an alignment algorithm that used a less constrained version of DTW in order to allow for matching when there is information in the audio signal that is not present in the score (e.g. ornamentation). In the following year, Müller, Mattes and Kurth (2006) presented a multi-scale DTW-based algorithm, which allowed better alignment for performances with significant structural variation from the reference MIDI file. Müller and Ewert (2008) later used this algorithm for assessing structural similarities between two pieces of audio by computing a joint structural analysis on the pieces.

More recently, Joder, Essid and Richard (2010) built on the multi-scale DTW model in Müller et al. (2006) with a hierarchical pruning method in which some pruning parameters were set adaptively and which resulted in both increased speed and accuracy. They also explored the use of CRFs through three formulations of the model's state transition function (Joder et al., 2011). Overall they found a direct relationship between accuracy and complexity, with the most accurate being the most computationally expensive.

There has been a limited interest in multi-pass approaches to refine an initial DTW alignment for increased accuracy. In previous work (Devaney, Mandel, & Ellis, 2009), the author and collaborators used acoustic features of the singing voice to guide an HMM to refine a DTW alignment based on Orio and Schwarz (2001). This work described an algorithm for aligning monophonic recordings of the singing voice where the initial DTW alignment served as a prior for the HMM, which refined the alignment using aperiodicity and power measurements along with fundamental frequency estimates as observations. The algorithm also identified the transient and steady state sections of the note and decreased the overall median alignment error in the initial DTW alignment for both onsets and offsets from 52 to 28 ms. Similarly, Niedermayer and Widmer (2010) used non-negative matrix factorization (NMF) to refine the estimate of the onsets of individual notes within vertical simultaneities in piano performances obtained by the DTW alignment approach in Müller et al. (2004). The piano is well suited to the use of NMF because the individual notes sound at a consistent pitch during a performance and contain characteristic inharmonicities in their spectrum. Niedermayer and Widmer evaluated their algorithm against accuracy metrics of 10 and 50 ms deviation from the ground truth. They used 10 ms for the first metric based on findings in Friberg and Sundberg (1993) that 10 ms is the perceptual limit for humans' ability to differentiate between onset times. The 50 ms metric was included to allow for comparison with earlier results. They found that their system improved the number of notes within the 10 ms threshold from 40.0% to 49.8%, but that there still remained a number of outliers for both this and the 50 ms metric. Note offsets were neither estimated nor evaluated.

Niedermayer and Widmer (2010) demonstrated the need for a high level of accuracy for expressive performance applications, one that exceeds the current requirements for both score following, as discussed above, and digital libraries. Research into the timing of asynchronies between notated simultaneities (either between fingers on the piano or between different performers in an ensemble) can range from 7 to 50 ms (Palmer, 1997). For digital music libraries, where the alignment is used to either visually link the score to the audio during playback or to find a particular section of the piece, the required alignment precision may be even lower and may range from note-level, at its most precise, to the bar-level, for certain applications such as syncing audio to a score for multi-modal playback. For a piece in 4/4 that is performed at 120 BPM, this would translate to 500 ms for quarter note-level precision or 2 s for bar-level precision. Evaluations of digital libraries are typically made in terms of overall usability or retrieval accuracy, rather than onset estimation accuracy. In contrast, signal processing applications, where alignment assists in source separation for such tasks as F_0 estimation and audio mixing, ideally requires a level of accuracy comparable to expressive performance applications.

3. Asynchrony in polyphonic performances

In earlier work (Devaney & Ellis, 2009), the author and a collaborator demonstrated that in various applications DTW was not sufficient by itself to estimate the timing of asynchronies in notated simultaneities. Evaluation was performed on a hand-annotated forty-second excerpt of multi-tracked recordings of the 'Kyrie' from Guillaume de Machaut's *Notre Dame Mass* (shown in Figure 1) and three different tests were performed. In the first test, each line of the monophonic recording was aligned to each part's MIDI file, in the second, the four parts were aligned simultaneously to the polyphonic composite of the entire MIDI file, and in the third, each part's MIDI file was aligned to the polyphonic composite. The multi-tracked recordings allowed for testing on both individual and composite tracks. Two measures were used to assess the accuracy

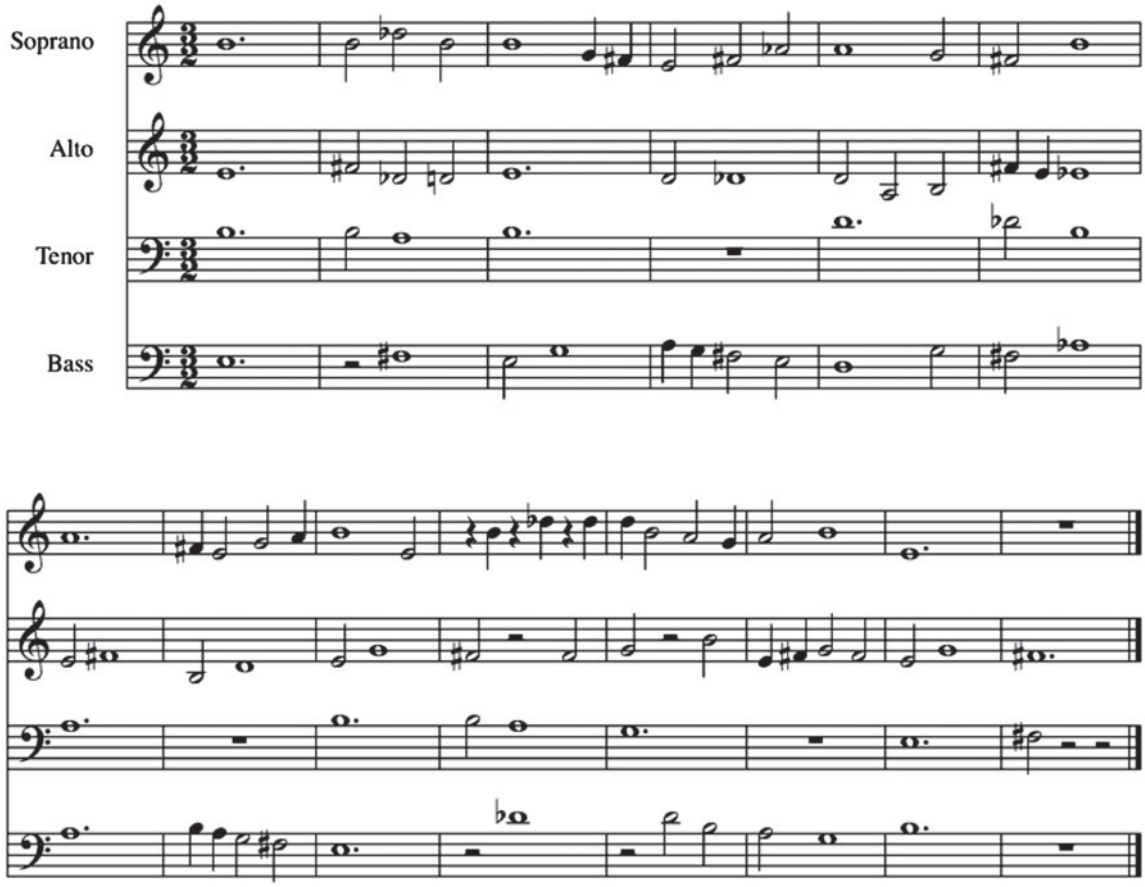


Fig. 1. Score of 'Kyrie' from Guillaume de Machaut's *Notre Dame Mass*. Reproduced from Devaney and Ellis (2009).

of the alignment: the first tallied the number of alignments that are within 100 ms of the ground truth's onsets and offsets and the second averaged the amount that the alignments were off from the ground truth. The results demonstrated that the individual alignments (test one) performed comparably to the simultaneous alignment (test two). In both of these tests the DTW alignment algorithm was able to consistently find the relevant notes in the audio signal, but the determination of the exact location of onsets and offsets was not always accurate (especially for asynchronies between simultaneously performed notes in the simultaneous alignment). The alignment of individual lines against the composite signal in test three did not prove to be a viable option for addressing asynchrony due to the DTW alignment algorithms' tendency to become lost when aligning a single line in a monaural recording of a polyphonic performance. Figure 2 provides a visual example of the problem with this approach. At the notated simultaneity between the soprano and the bass around 13.3 s the alignment is locked to the onset of the soprano's note, which, in the performance, is approximately 30–40 ms behind the onset of the bass' note. Also, the offset of the tenor note occurs approximately 100 ms before the other voices' offsets. A multi-step extension to the DTW approach for simultaneous alignment used in test two was suggested as a promising way to address the issue of onset and offset asynchrony in notated

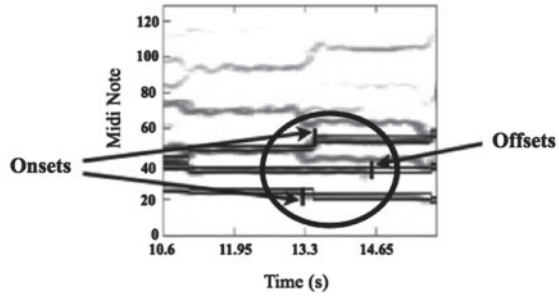


Fig. 2. Example of errors in estimating onset and offset asynchrony. The boxes indicate the DTW estimates and the vertical lines and arrows identify the actual locations of the onsets and offsets in the performance. Reproduced from Devaney and Ellis (2009).

simultaneities, though this was not designed or implemented at that time.

4. Algorithm for addressing asynchrony

To address the issue of onset and offset asynchrony in notated simultaneities, this presents a multi-pass algorithm (DTW/HMM) that uses both DTW and an HMM to estimate the location of the onsets and offsets for each voice in a monaural recording of a polyphonic performance. The algorithm was

developed and tested for the singing voice, one of the hardest paradigms for music alignment, but can easily be extended to other instruments by retraining the HMM on appropriate recordings. The algorithm was designed for recordings with one voice per part, and thus is appropriate for smaller vocal ensembles rather than a choir. As in Devaney et al. (2009), discussed above in Section 2.2, DTW is used in the first pass to obtain a rough estimate of the note locations and an HMM is used in the second pass to refine the offset–onset transitions between groups of 'simultaneous' notes in the DTW alignment in order to estimate the location of the onsets and offsets for each voice. The HMM assumes that the DTW is roughly correct and only looks at the audio 125 ms before and after the onset identified by the DTW alignment, thus it is only able to correct errors in the DTW alignment by a maximum of that amount. A visual representation of the DTW alignment allows for detection of gross errors in the DTW alignment, which can be manually corrected.

The multi-pass algorithm uses the DTW alignment method described in Orio and Schwarz (2001), which creates an idealized harmonic template for each note in the score and calculates the peak spectral difference between the template and the audio in order to build a similarity matrix through which the DTW algorithm finds the best path. The output of the DTW is a series of note transition times, $\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_T$. The algorithm then uses the HMM to find the best sequence of note offsets and onsets for each voice in the recording using the processed output of a constant-Q filter bank, X, as observations. The DTW used Dan Ellis' (2003) MATLAB code and the HMM was implemented with Kevin Murphy's (1998) HMM toolbox for MATLAB. Müller and Ewert's (2011) MATLAB toolbox was used for calculating the filter bank and for estimating the tuning of the recording. Tuning estimation was particularly important for running the algorithm on *a cappella* vocal ensembles because they do not necessarily sing with a strict tuning reference.

4.1 HMM states and transitions

In the HMM, each voice present in the signal can be in one of three sub-states: Note1, NoteOff, and Note2, denoted $s \in \{-1, 0, 1\}$. Note1 reflects the time spent on the note being sung at the beginning of the offset–onset transition. NoteOff reflects the time spent after the first note and before the second note, when no note is being sung. Note2 reflects the time spent on the note being sung at the end of the offset–onset transition. The offset for the first note is calculated from the point in time when the HMM changes from the Note1 state to the NoteOff state and the onset for the second note is calculated from the point in time when the HMM changes from the NoteOff state to the Note2 state.

Each state in the HMM is a composite with one sub-state per voice, $S = [s_1 \ldots s_N]^T \in \{-1, 0, 1\}^N$, where N is the number of parts in the offset–onset transition. A monophonic offset–onset transition would consists of just three states and as the number of voices increases the number of states increases as 3^N. Thus, a two-part offset–onset transition consists of nine states (shown in Figure 3), a three-part offset–onset transition consists of 27 states, and a four-part offset–onset transition consists of 81 states. This exponential growth is not a problem for small ensembles such as vocal trios or quartets. Also, since the HMM only operates on small chunks of audio, the computational cost is low. State transitions and paths are restricted so that only one voice may change sub-states at a time and sub-state paths must go through the 'off' state. Figure 3 shows the possible paths that the HMM may take in a two-voice offset–onset transition and which sub-state each voice is in for each HMM state. Self-loop transition probabilities are 10 times more likely than each other allowable state transition, although the actual probabilities are dependent on the number of outgoing states. Experimentation revealed that the results are not very sensitive to the transition probabilities; rather it is the legal state paths that are important.

4.2 HMM observations

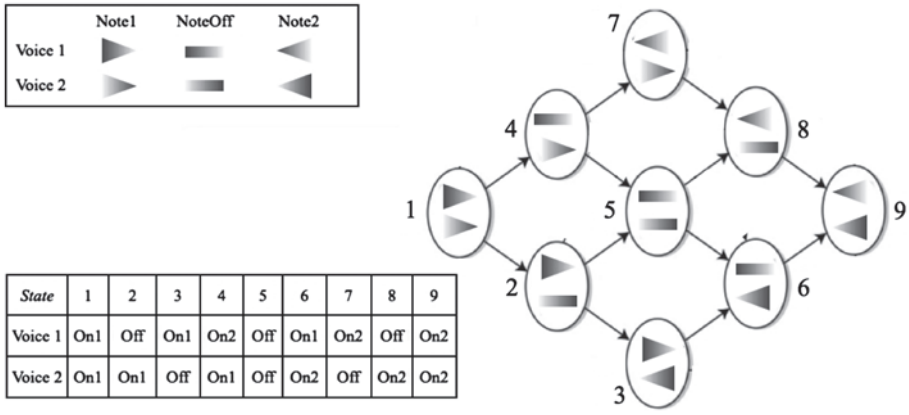
The likelihood of each sub-state is independent within each state, and is parametrised by which notes are on and which are off. The HMM's observation model is a single, multivariate Gaussian per state, with two observation dimensions for each voice, one for Note1 and one for Note2. The probabilities for the presence of Note1, NoteOff, and Note 2 are:

$$p\left(s_v^{(n)} = -1\right) = p_{\text{on}}\left(x_v^{(n)}(t)\right) p_{\text{off}}\left(x_v^{(n+1)}(t)\right),$$
$$p\left(s_v^{(n)} = 0\right) = p_{\text{off}}\left(x_v^{(n)}(t)\right) p_{\text{off}}\left(x_v^{(n+1)}(t)\right),$$
$$p\left(s_v^{(n)} = 1\right) = p_{\text{off}}\left(x_v^{(n)}(t)\right) p_{\text{on}}\left(x_v^{(n+1)}(t)\right),$$

where $p_{\text{on}}(x) = N(x; \mu_{\text{on}}, \sigma_{\text{on}})$ and $p_{\text{off}}(x) = N(x; \mu_{\text{off}}, \sigma_{\text{off}})$. In the case when a voice re-articulates the same note, $p(s_v^{(n)} = -1) = p(s_v^{(n)} = 1)$.

The parameters for the observations were calculated using hand annotations of a multi-tracked four-part recording of the 'Kyrie' from Machaut's *Notre Dame Mass* (shown in Figure 1). Since each singer was recorded separately, the note onsets and offsets could be precisely hand-annotated in the individual tracks and the observations could be calculated in the composite signal. The means and covariances were modelled with a single Gaussian for when a note was present, $p_{\text{on}}(x_v^{(n)}(t))$, and a different Gaussian for when one was not, $p_{\text{off}}(x_v^{(n)}(t))$, with separate parameters for male and female voices. These models were evaluated and found to be robust on audio files with different levels of amplification.

The observations, $x_v^{(n)}(t)$, for the HMM are calculated from a constant-Q filter bank decomposition of the signal with one filter per semitone, $X(f, t)$. Guided by the DTW alignment, a power measurement (in decibels) is summed over a 3-semitone span around the fundamental of the MIDI note (the fundamental plus/minus one semitone) for both the ending note, $f_v^{(n)}$, and starting note, $f_v^{(n+1)}$, in each offset–onset transition identified by the DTW alignment,



State	1	2	3	4	5	6	7	8	9
Voice 1	On1	Off	On1	On2	Off	On1	On2	Off	On2
Voice 2	On1	On1	Off	On1	Off	On2	Off	On2	On2

Fig. 3. The schematic on the right shows all of the possible states and state transitions for a scenario when two voices change. A summary of the combinations of note ons and offs for each state is shown in the table in the lower left.

$x_v^{(n)}(t) = \sum_{f=f_v^{(n)}-1}^{f_v^{(n)}+1} X(f,t)$, where $t_n - 0.125 \leqslant t \leqslant t_n + 0.125$. The inclusion of additional harmonics in this sum did not improve results. This is likely due to overlap between the voices from the ± 1 semitone range used to calculate the observations. These sub-state likelihoods are then combined into the composite state likelihoods assuming that they are independent, and then the Viterbi path is found through the HMM to perform the alignment.

The algorithm automatically identifies onset–offset transitions as places in the MIDI file where at least one voice changes notes or re-articulates its current note. For each offset–onset transition, an HMM is constructed based on the score to use the appropriate state space size and parameters. Only those voices changing notes are included in the HMM alignment process for a given transition. Thus, for a four-part ensemble, a transition may include one, two, three, or four voices. For the corner case where one voice moves to a note already being held by another voice, the DTW onset estimate is used for the moving voice. This is because the observations for the HMM cannot distinguish between the held voice and moving voice on the same note. (This corner case occurs in the Benedetti exercise used in the evaluation in Section 5.)

5. Evaluation of algorithm

The algorithm was tested on two sets of the recordings. The first set was four recordings of a three-part exercise (alto-tenor-bass) by Giambattista Benedetti, shown in Figure 4. The second set was seven recordings of the first verse of Michael Praetorius' 'Es ist ein Ros' entsprungen', a four-part piece shown in Figure 5. As with the Machaut recording, these recordings were multi-tracked and hand annotated by the author and another person, which allowed for an accurate determination of the onset and offset for each note. The evaluations were run on the composite signals, where the three or four multi-track parts were mixed down to a monaural audio file.

The accuracy of the algorithm was evaluated by two metrics. The first is the number of onsets and offsets within a fixed interval of the ground truth. Following from Niedermayer and Widmer (2010), this was calculated for both 10 ms (Table 1) and 50 ms (Table 2) thresholds. Across the combination of both test sets, the DTW/HMM algorithm improved the DTW alignment for the onsets and offsets for both the 10 ms (from 2% to 9% of the notes for the onsets and 6% to 11% for the offsets) and 50 ms (from 11% to 39% of the notes for the onsets and from 23% to 40% for the offsets) thresholds. This general trend held for the Praetorius test set, although the percentage of notes within the threshold was proportionally much higher for the DTW/HMM algorithm versus the DTW than for the Benedetti test set. This is likely due to the increased complexity of the musical texture, which could lead the performers to greater timing asynchrony than the more chordal nature of the Benedetti. In the Benedetti test set, however, the DTW outperformed the DTW/HMM algorithm for the offsets within 10 ms, by 17% to 11%. An examination of the individual voices shows that the largest performance difference between the DTW alignment and the DTW/HMM occurred in the alto voice (27 onsets within 10 ms for the DTW algorithm versus nine for the DTW/HMM algorithm) is likely due to the fact that this voice changes notes most frequently and thus acted as an anchor for the DTW alignment such that this voice was most accurate at the expense of the other voices.

The second measurement of accuracy is the 2.5th, 25th, 50th, 75th, and 97.5th percentiles of the difference between the predictions and the ground truth, which provide information about the distribution of errors (Table 3). Overall the DTW/HMM algorithm improves the median (50th percentile) onsets results for both test sets, which shows that on average the algorithm is improving the accuracy of the alignment. For offsets, there is only an improvement for the median with the DTW/HMM for the Bendetti test set, although the difference for the Praetorius test set between the DTW and the DTW/HMM is less than 6 ms. For the Benedetti test set, the DTW/HMM algorithm also improves the 75th and 97.5th

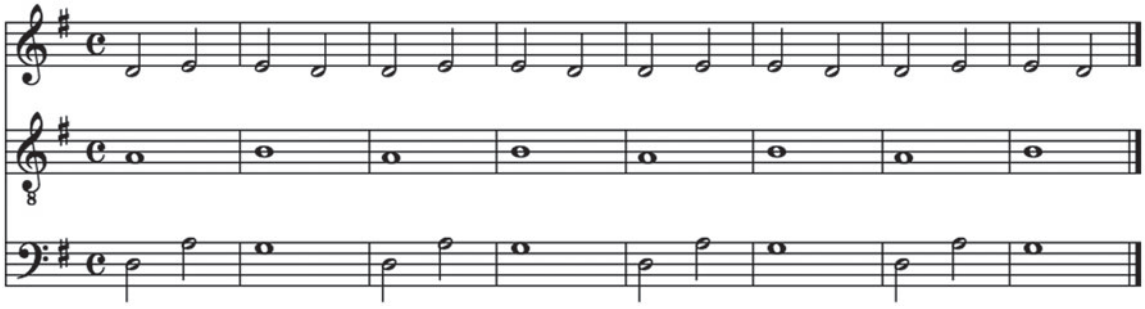


Fig. 4. Score of exercise by Giambattista Benedetti used for evaluating the alignment algorithm.

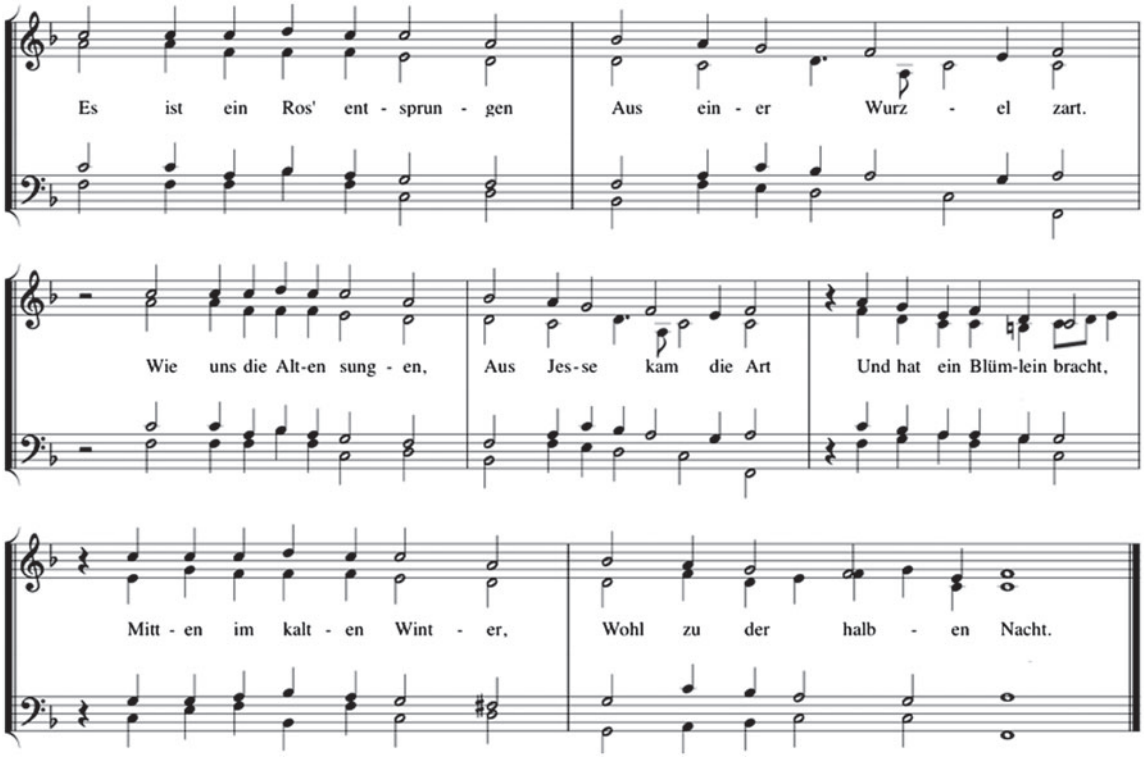


Fig. 5. Score of Michael Praetorius' 'Es ist ein Ros' entsprungen' used for evaluating the alignment algorithm.

Table 1. The number of onsets and offsets predicted by the alignment within 10 of the ground truth. DTW indicates the values for the original DTW alignment and DTW/HMM indicates the values for the multi-pass algorithm. Bold typeface indicates the better results for each vertical category.

	Number of onsets within 10 ms					
Vocal Part	**Benedetti (228)**		**Praetorius (1252)**		**Total (1480)**	
(# of notes)	***On***	***Off***	*On*	***Off***	***On***	***Off***
DTW	14 (6%)	**38 (17%)**	9 (1%)	47 (4%)	23 (2%)	85 (6%)
DTW/HMM	**20 (9%)**	25 (11%)	**107 (9%)**	**140 (11%)**	**127 (9%)**	**165 (11%)**

percentile results for both onsets and offsets, which means that it is correcting the outliers; although, since the algorithm can only correct the DTW by 125 ms in either direction, it can only improve alignment errors by this amount. The DTW/HMM algorithm improves the 25th percentile results for onsets in the Benedetti test set and both the onsets and offsets in the

Table 2. The number of onsets and offsets predicted by the alignment within 50 ms of the ground truth. DTW indicates the values for the original DTW alignment and DTW/HMM indicates the values for the multi-pass algorithm. Bold typeface indicates the better results for each vertical category.

	Number of onsets within 50 ms					
Vocal Part	**Benedetti (228)**		**Praetorius (1252)**		**Total (1480)**	
(# of notes)	***On***	***Off***	*On*	***Off***	***On***	***Off***
DTW	88 (39%)	102 (45%)	81 (6%)	247 (20%)	169 (11%)	347 (23%)
DTW/HMM	**119 (52%)**	**109 (48%)**	**465 (37%)**	**477 (38%)**	**584 (39%)**	**386 (40%)**

Table 3. 2.5th, 25th, 50th, 75th, and 97.5th percentiles for the discrepancy between the onset and offset alignments and the ground truth. DTW indicates the values for the original DTW alignment and DTW/HMM indicates the values for the multi-pass algorithm. Bold typeface indicates the better results for each vertical pair of values.

			Percentiles				
			2.5	**25**	**50**	**75**	**97.5**
Benedetti	***Ons***	**DTW**	**3.1**	29.2	66.7	122.5	759.5
		DTW/HMM	3.8	**19.2**	**47.7**	**100.1**	**653.2**
	Offs	**DTW**	**1.0**	**21.8**	64.8	141.9	752.9
		DTW/HMM	1.5	25.1	**53.4**	**131.3**	**677.0**
Praetorius	***Ons***	**DTW**	18.2	87.2	142.0	244.3	**991.15**
		DTW/HMM	**2.8**	**30.8**	**82.6**	**206.2**	1071.8
	Offs	**DTW**	4.2	35.2	77.2	**173.0**	**1180.0**
		DTW/HMM	**1.8**	**23.2**	**74.5**	213.3	1250.4
Total	***Ons***	**DTW**	9.8	63.5	117.8	209.3	**831.4**
		DTW/HMM	**3.1**	**29.0**	**77.4**	**196.4**	1043.0
	Offs	**DTW**	2.1	31.8	74.7	**167.3**	**1112.8**
		DTW/HMM	**1.7**	**23.2**	**68.9**	206.9	1217.2

Praetorious test sets, meaning that for these, the good alignments get better. The DTW/HMM algorithm also improved the 2.5th percentile for the Praetorius test set but not for the Benettedi test set, though for the Bendetti it was only worse than the DTW on the order of 1 or 2 ms.

6. Conclusions

This paper has presented an offline multi-pass score-audio alignment algorithm for estimating note onset and offset asynchronies in monaural recordings of the singing voice between notes marked as simultaneities in the corresponding score. In the algorithm, an HMM is used to refine a DTW alignment with power estimates derived from a constant-Q filter bank decomposition of the signal as observations. The parameters for the HMM were trained on a multi-tracked vocal ensemble recording. The algorithm was evaluated on the mixed-down audio of a set of three- and four-part multi-track vocal performances, which were hand annotated with ground truth onsets and offsets. By providing estimates of individual onsets and offsets in notated simultaneities, the multi-pass algorithm improves the median accuracy of the DTW alignment for onsets by 41 ms on average, from 118 to 77 ms, and for offsets by 6 ms on average, from 75 to 69 ms. This algorithm can be used for expressive performance applications, both for estimating timing relationships between voices and for guiding signal-processing estimation of other performance parameters, such as pitch and dynamics. The algorithm can also be used for guiding other types of signal processing algorithms and source separation.

One future direction for this research is to generalize the algorithm to work for a range of instruments, not specifically the singing voice. This can be done with the current observations by creating a training set for the type of instrument one wishes to add. A more robust option is to explore other types of observations. One possibility is the use of a template-model similar to the one used in the DTW algorithm by Orio and Schwarz (2001), described above, either with the peak spectral distance used by Orio and Schwarz (2001) or the Kullback-Leibler divergence used by Cont (2010). This approach has the advantage of not requiring training, since the templates are fixed, theoretically making it usable for any instrumentation. Another area of future work is combining multiple features for the observations, in addition to the power measurements currently used, which could account for overlapping notes between voices. In earlier work (Devaney et al., 2009), the author and collaborators found that a combination of aperiodicity and power measurements along with fundamental frequency estimates as observations provided the best results for refining a DTW alignment of monophonic sung audio using an HMM. While these specific features are not available in a polyphonic context it is likely that the use of a combination of features in this algorithm would improve results.

Acknowledgements

The author would like to acknowledge the contributions of her collaborators in earlier stages of this work, specifically Michael Mandel, Dan Ellis, and Ichiro Fujinaga.

Funding

The author would like to acknowledge the financial support provided by the Fonds québécois de la recherche sur la société et la culture (FQRSC) and Social Sciences and Humanities Research Council of Canada (SSHRC) in the initial stages of the work. The underlying research materials for this article can be accessed at http://www.ampact.org.

References

Baird, B., Blevins, D., & Zahler, N. (1990). The artificially intelligent computer performer: The second generation. *Interface / Journal of New Music Research, 19*(2), 197–204.

Baird, B., Blevins, D., & Zahler, N. (1993). Artificial intelligence and music: Implementing an interactive computer performer. *Computer Music Journal, 17*(2), 73–79.

Basaran, D., Cemgil, A.T., & Anarım, E. (2011). Model based multiple audio sequence alignment. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (pp. 13–16). Piscataway, NJ: IEEE Press.

Cano, P., Loscos, A., & Bonada, J. (1999). Score-performance matching using HMMs. In *Proceedings of the International Computer Music Conference* (pp. 441–444). Ann Arbor, MI: Michigan Publishing.

Cont, A. (2006). Realtime audio to score alignment for polyphonic music instruments using sparse non-negative constraints and hierarchical HMMs. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 185–188). Piscataway, NJ: IEEE Press.

Cont, A. (2010). A coupled duration-focused architecture for realtime music to score alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*(6), 974–987.

Cont, A., Schwarz, D., Schnell, N., & Raphael, C. (2007). Evaluation of real-time audio-to-score alignment. In *Proceedings of the International Conference on Music Information Retrieval* (pp. 315–316). Vienna: Austrian Computer Society (OCG).

Damm, D., Fremerey, C., Kurth, F., Müller, M., & Clausen, M. (2008). Multimodal presentation and browsing of music. In *Proceedings of the International Conference on Multimodal Interfaces* (pp. 205–208). New York: ACM.

Dannenberg, R. (1984). An on-line algorithm for real-time accompaniment. In *Proceedings of the International Computer Music Conference* (pp. 193–198). San Francisco: International Computer Music Association.

Dannenberg, R. (2007). An intelligent multi-track audio editor. In *Proceedings of the International Computer Music Conference* (pp. 89–94). Ann Arbor, MI: Michigan Publishing.

Devaney, J., & Ellis, D.P.W. (2009). Handing asynchrony in audio-score alignment. In *Proceedings of the International Computer Music Conference* (pp. 29–32). Ann Arbor, MI: Michigan Publishing.

Devaney, J., Mandel, M.I., & Ellis, D.P.W. (2009). Improving MIDI-audio alignment with acoustic features. In *Proceedings of the Workshop on Applications of Signal Processing to Acoustics and Audio* (pp. 45–48). Piscataway, NJ: IEEE.

Devaney, J., Mandel, M., & Fujinaga, I. (2012). A Study of Intonation in Three-Part Singing using the Automatic Music Performance Analysis and Comparison Toolkit (AMPACT). In *Proceedings of the International Society for Music Information Retrieval Conference* (pp. 511–516). Porto, Portugal: FEUP Edições.

Dixon, S. & Widmer, G. (2005). MATCH: A music alignment toolkit. In *Proceedings of the International Conference on Music Information Retrieval* (pp. 492–497). London: Queen Mary, University of London. Available online at: http://ismir2005.ismir.net/proceedings/1002.pdf

Duan, Z., & Pardo, B. (2011). A state space model for online polyphonic audio-score alignment. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (pp. 197–200). Piscataway, NJ: IEEE.

Dunn, J.W., Byrd, D., Notess, M., Riley, J., & Scherle, R. (2006). Variations2: Retrieving and using music in an academic setting. *Communications of the ACM, 49*(8), 53–58.

Ellis, D.P.W. (2003). Dynamic Time Warp (DTW) in Matlab. Available from http://www.ee.columbia.edu/dpwe/resources/matlab/dtw/

Friberg, A., & Sundberg, J. (1993). Perception of just noticeable time displacement of a tone presented in a metrical sequence at different tempos. In *Proceedings of the Stockholm Music Acoustics Conference* (pp. 39–43). Stockholm: Royal Swedish Academy of Music.

Grubb, L., & Dannenberg, R. (1997). A stochastic method of tracking a vocal performer. In *Proceedings of the International Computer Music Conference* (pp. 301–308). Ann Arbor, MI: Michigan Publishing.

Joder, C., Essid, S., & Richard, G. (2010). An improved hierarchical approch for music-to-symbolic score alignment. In *Proceedings of the International Society for Music Information Retrieval Conference* (pp. 39–44). Utrecht: Universiteit Utrecht.

Joder, C., Essid, S., & Richard, G. (2011). A conditional random field framework for robust and scalable audio-to-score matching. *Transactions on Audio Speech and Language Processing, 19*(8), 2385–2397.

Kurth, F., Müller, M., Damm, D., Fremerey, C., Ribbrock, A., & Clausen, M. (2005). Syncplayer: An advanced system for multimodal music access. In *Proceedings of the International Conference of Music Information Retrieval* (pp. 381–385). London: Queen Mary, University of London. Available online at: http://ismir2005.ismir.net/proceedings/1025.pdf

Montecchio, N., & Cont, A. (2011a). Accelerating the mixing phase in studio recording productions by automatic audio alignment. In *Proceedings of the International Society for Music Information Retrieval Conference* (pp. 627–632). Canada: International Society for Music Information Retrieval. Available online at: http://ismir2011.ismir.net/papers/OS7-3.pdf

Montecchio, N., & Cont, A. (2011b). A Unified Approach to Real Time Audio-to-Score and Audio-to-Audio Alignment Using Sequential Montecarlo Inference Techniques. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (pp. 193–196). Piscataway, NJ: IEEE.

Müller, M., & Ewert, S. (2008). Joint structure analysis with applications to music annotation and synchronization. In *Proceedings of the International Conference on Music Information Retrieval* (pp. 389–394). Canada: International Society for Music Information Retrieval. Available online at: http://ismir2008.ismir.net/papers/ISMIR2008_115.pdf

Müller, M., & Ewert, S. (2011). Chroma toolbox: MATLAB impelmentations for extracting variants of chroma-based audio features. In *Proceedings of the International Society for Music Information Retrieval* (pp. 215–220). Canada: International Society for Music Information Retrieval. Available online at: http://ismir2011.ismir.net/papers/PS2-8.pdf

Müller, M., Kurth, F., & Roder, T. (2004). Towards an efficient algorithm for automatic score-to-audio synchronization. In *Proceedings of the International Conference on Music information Retrieval* (pp. 365–372). Barcelona: Universitat Pompeu Fabra. Available online at: ismir2004.ismir.net/proceedings/p067-page-365-paper136.pdf

Müller, M., Mattes, H., & Kurth, F. (2006). An efficient multiscale approach to audio synchronization. In *Proceedings of the International Conference on Music information Retrieval* (pp. 192–197). Victoria, BC: University of Victoria. Available online at: http://ismir2006.ismir.net/PAPERS/ISMIR0615_Paper.pdf

Murphy, K. (1998). Hidden Markov Model (HMM) Toolbox for Matlab. Available from http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html

Niedermayer, B. & Widmer, G. (2010). A multi-pass algorithm for accurate audio-to-score alignment. In *Proceedings of the International Society for Music Information Retrieval Conference* (pp. 417–422). Canada: International Society for Music Information Retrieval. Available online at: http://ismir2010.ismir.net/proceedings/ismir2010-71.pdf

Orio, N., & Dèchelle, F. (2001). Score following using spectral analysis and hidden Markov models. In *Proceedings of the International Computer Music Conference* (pp. 151–154). Ann Arbor, MI: Michigan Publishing.

Orio, N., & Schwarz, D. (2001). Alignment of monophonic and polyphonic music to a score. In *Proceedings of the International Computer Music Conference* (pp. 155–158). Ann Arbor, MI: Michigan Publishing.

Orio, N., Nakadai, K., Ogata, T., & Okuno, H.G. (2011). Incremental Bayesian audio-to-score alignment with flexible harmonic structure models. In *Proceedings of the International Society for Music Information Retrieval Conference* (pp. 525–530). Canada: International Society for Music Information Retrieval. Available online at: http://ismir2011.ismir.net/papers/PS4-9.pdf.

Palmer, C. (1997). Music performance. *Annual Review of Psychology, 48*, 115–138.

Peeling, P., Cemgil, T., & Godsill, S. (2007). A probabilistic framework for matching music representations. In *Proceedings of the International Conference on Music Information Retrieval* (pp. 267–272). Vienna: Austrian Computer Society (OCG).

Puckette, M. (1995). Score following using the sung voice. In *Proceedings of the International Computer Music Conference* (pp. 175–178). Ann Arbor, MI: Michigan Publishing.

Raphael, C. (2004). A hybrid graphical model for aligning polyphonic audio with musical scores. In *Proceedings of the International Conference on Music information Retrieval* (pp. 387–394). Barcelona: Universitat Pompeu Fabra.

Scheirer, E. (1995). *Extracting expressive performance information from recorded music* (Master's thesis). Massachusetts Institute of Technology, Media Laboratory, Cambridge, MA, USA.

Smit, C., & Ellis, D. (2009). Guided harmonic sinusoid estimation in a multi-pitch environment. In *Proceedings of the Workshop on Applications of Signal Processing to Audio and Acoustics* (pp. 41–44). Piscataway, NJ: IEEE.

Soulez, F., Rodet, X., & Schwarz, D. (2003). Improving polyphonic and poly-instrumental music to score alignment. In *Proceedings of the International Conference on Music information Retrieval* (pp. 143–148). Baltimore, MD: Johns Hopkins University.

Stammen, D., & Pennycook, B. (1993). Real-time recognition of melodic fragments using the dynamic timewarp algorithm. In *Proceedings of the International Computer Music Conference* (pp. 232–235). Ann Arbor, MI: Michigan Publishing.

Vercoe, B. (1984). The synthetic performer in the context of live performance. In *Proceedings of the International Computer Music Conference* (pp. 199–200). Ann Arbor, MI: Michigan Publishing.

Woodruff, J., Pardo, B., & Dannenberg, R. (2006). Remixing stereo music with score-informed source separation. In *Proceedings of the International Conference on Music Information Retrieval* (pp. 314–319). Victoria, BC: University of Victoria. Available online at: http://ismir2006.ismir.net/PAPERS/ISMIR06102_Paper.pdf