ENHANCED VOCAL PERFORMANCE TRACKING USING MULTIPLE INFORMATION SOURCES

Lorin Grubb and Roger Dannenberg
{lgrubb, rbd}@cs.cmu.edu
School of Computer Science, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213

Abstract

We describe an enhanced statistical method for tracking vocal performances in real time. This approach integrates multiple measurements of the sound signal that are relevant to determining score position. The information is combined using a statistical model that incorporates distributions estimated from actual vocal performances. We present results from using different combinations of signal measurements to track both recorded and live performances. Results of this testing indicate some requirements for successfully using multiple information sources to improve automated vocal performance tracking.

1. Motivation

Many applications in computer music require the ability to listen to live or recorded musical performances by following along in a written score. This ability is necessary for systems executing tasks such as automated accompaniment (Dannenberg 1984; Vercoe 1984), interactive performance (Katayose *et al.* 1993), musical tutoring or coaching (Dannenberg and Joseph 1992), and automated performance analysis (Honing 1990). Typically, both the score position and tempo of the performer must be identified with high accuracy, high precision, and high frequency during the performance. Some tasks, such as accompaniment, also require the tracking results in real time—sometimes demanding latencies below 100 milliseconds.

Performance tracking for instrumentalists was first considered by Dannenberg (1984) and Vercoe (1984). More recently, systems that require the ability to accurately and efficiently track a vocalist have also been presented (Katayose *et al.* 1993) (Inoue, Hashimoto, Ohteru 1994) (Puckette 1995). Tracking a vocalist is a particularly challenging task. Vocal performances are highly fluid, making it difficult to reliably and easily distinguish performed notes. In addition, features extracted from the performance (such as fundamental pitch) often vary from what is notated in the score. Frequently, singers introduce these variations for expressive and stylistic purposes. For instance, operatic singers generally apply vibrato, where the sung pitch is intentionally made to oscillate around the pitch notated in the score. Finally, techniques for extracting relevant features from the sound signal, including pitch detection (Kuhn 1990) and vowel detection (Inoue, Hashimoto, Ohteru 1994), are never error-free.

Recently, a system using a robust statistical method for tracking vocalists has been demonstrated (Grubb and Dannenberg 1997). This approach incorporates statistical descriptions of both detected fundamental pitch in vocal performances and the expected progress of a singer through the score (*i.e.*, the expected amount of score performed based on recent tempo and elapsed time). The tracking system uses a probabilistic description of a performer's score position, making explicit the system's current uncertainty about the singer's location. This tracking method was used in a computer accompaniment system and a preliminary evaluation of its tracking ability was provided.

We present a new system that uses an extended statistical model for tracking vocalists. This method incorporates multiple measurements of the sound signal, including fundamental pitch, spectral features related to the phonetic content of the performance, and amplitude changes indicative of note onsets. Detecting multiple features in a performance can enable more robust tracking. For instance, vocal scores of Western classical music often contain several successive notes with the same pitch. In some cases, a single pitch may be repeated for several measures. By considering phonetic content of a performance, the system can track position and tempo changes even through measures where the notated pitch does not change. Likewise, vocal scores often contain melismas where many successive notes are sung on the same syllable. While phonetic content does not help determine position during melismas, detected pitch and onsets can enable the system to track the performer through these phrases. Finally, to the extent that deviations in pitch, onset, and phoneme production are not correlated, a system that incorporates all types of features will be more robust than a system that relies exclusively on one feature. This property also applies to detection and measurement errors due to the respective signal processing techniques for each feature.

In this paper, we first review the basic tracking model and show how it can be extended to incorporate multiple features reported during a single performance. The empirically derived probability distributions employed by the

Grubb, L. and Dannenberg, R. 1998. Enhanced Vocal Performance Tracking Using Multiple Information Sources. In *Proceedings of the 1998 International Computer Music Conference.* San Francisco: International Computer Music Association.

model are then described. We present results of applying several versions of the model to track recorded and live performances by experienced singers. Finally, we discuss the implications of these results, examining both the conditions necessary for multiple signal measurements to improve the accuracy of performance tracking and the areas of subsequent investigation most likely to lead to even more robust tracking of vocal performances.

2. A Stochastic View of a Singer's Score Position

The model for stochastic vocal performance tracking represents the vocalist's part as a sequence of *events* that have a fixed ordering, or at least a desired ordering. Each *event* is specified by:

1. A *relative length* that defines the duration of the event as indicated in the score, relative to other events.
2. An *observation distribution* that completely specifies the probability of every possible observation, or signal measurement, at any time during the event.

Note that the model does not require a one-to-one correspondence between a note and an event. Events can map to portions of a note or regions spanning note boundaries. The relative length of an event may be specified in musical beats for a fixed tempo. It may also be specified in a unit of time, requiring conversion of beats to "idealized time" using a fixed, idealized tempo. The length is assumed to be a positive real value, not necessarily an integer.

The observation distribution for each event specifies the probability of observing each possible value of a detected feature when the vocalist is performing that event. This distribution generally is conditioned on information provided in the score. For example, assume that pitch detection is applied to a performance. In this case, the observation distribution of a note scored for the pitch A-440 might specify the likelihood of the pitch detector reporting each possible pitch. In addition to the possibility of observing an A, there would be some probability of observing a G, or a B, or a B-flat, etc. As another example, observation distributions can also describe the likelihood of detectable spectral features that are correlated with sung vowels.

The vocalist's part in the score is viewed as a sequence of events, each event spanning a region of the number line. The *score position* of a singer is represented as a real number between zero and the sum of the lengths of all events in the score. Score position is thus specified in either idealized beats or idealized time, and can indicate the performer's location with precision finer than an event. At any point while tracking an actual performance, the position of the vocalist is represented by a continuous probability density function over score position, the *score position density.* The area under this function between two positions indicates the probability that the performer is within that region of the score. This concept is depicted in Figure 1. The area over the entire length of the score is always 1, indicating it is 100% likely that the performer is in the score. As the performance progresses and subsequent observations (signal measurements) are reported, the score position density is updated to yield a probability distribution describing the performer's new location.

To describe the performance tracking system, we will first summarize the basic mathematical model for updating the position density assuming one type of observation only. Next, we will extend the basic model to incorporate observations encompassing several simultaneous but distinct measurements of the sound signal. Finally, we will describe how the probability functions in the model were estimated from actual vocal performances.

3. The Basic Model for Estimating Score Position

The general approach to tracking the performer is conceptually simple. During a performance, observations are periodically made through sensors (signal processing techniques). For each observation, we use the current score position density and the observation distributions to estimate a new score position density. This updated density

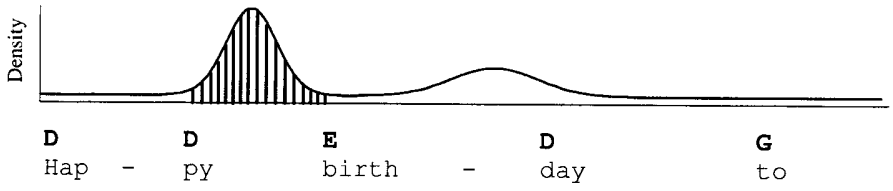

Figure 1: Example of a density function characterizing the score position of a performer. The area of the shaded region gives the probability that the performer is singing the second note.

characterizes the current location of the performer in the score. In practice, however, calculating a new score position density requires a number of simplifications, assumptions, and approximations.

The method of updating the score position density incorporates three values that are relevant to determining the new position of the performer, or the performer's *destination position*. First, since a performer's rendering of a musical score is highly sequential, it is important to consider the performer's location at the time of the previous observation, the performer's *source position*. Second, the observation most recently extracted from the performance will obviously provide information about the performer's current location. Finally, performers often attempt to maintain a consistent tempo, subject to relatively minor and gradual variations. An estimate of the performer's tempo in the recent past, along with the elapsed time since the position density was last updated, gives a useful prediction of how much score was recently performed. This prediction will be referred to as the *estimated distance*.

Given these three variables—previous position, most recent observation, and estimated distance—the current location of the performer can be specified stochastically by the following conditional probability density:

$$f_{I \mid D,V,J}\,(i \mid d, v, j)$$

i = the performer's destination position
d = the estimated distance
v = the observation
j = the performer's source position

Unfortunately, directly defining this multidimensional function for every score would be challenging. Also, the previous score position of the performer is never known with certainty, so the value of at least one conditioning variable, j, should also be described statistically. To provide a model that is feasible to define and implement, we make several simplifying assumptions. The assumptions permit an approximation to the conditional density. First, an estimate of current location based on prior location and estimated distance is calculated through convolution:

$$f_{I \mid D}\,(i \mid d) = \int_{j=0}^{\|Score\|} f_{I-J \mid D}\,(i-j \mid d) \cdot f_{Source}(j)\,\partial j \qquad (1)$$

Next, this estimate is updated to account for the most recent observation using a calculation similar to Bayes' Rule:

$$f_{I \mid D,V}\,(i \mid d, v) = \frac{f_{V \mid I}\,(v \mid i) \cdot f_{I \mid D}\,(i \mid d)}{\int_{k=0}^{\|Score\|} f_{V \mid I}\,(v \mid k) \cdot f_{I \mid D}\,(k \mid d)\,\partial k} \qquad (2)$$

The result is a score position density conditioned on both the estimated distance and the most recent observation. Note that if d and v represent fixed values, the result is a one-dimensional function over score position. In addition, actual computation of the score position density for real applications can be restricted to a small window of score at any instant (*i.e.*, over a restricted range of values for i). The window can move through the score over time, following the performer. Details of all assumptions necessary to apply the simplified model, as well as efficient methods for numerically computing the model, are provided by Grubb and Dannenberg (1997).

4. Incorporating Multiple Observations

The model for estimating score position includes a density function, $f_{V \mid I}$, specifying the likelihood of observing any possible sensor output for every score position. In the basic model we assumed that the observation is a single value reported by a single sensor or signal processing algorithm. This model for estimating score position can be easily extended to incorporate observations from multiple sensors, providing that two conditions are satisfied. First, the model assumes that the observation, v, is independent of the performer's source position, j, and the estimated distance, d. This assumption implies that a function specifying the likelihood of any value for v does not change if the value of either j or d changes. All observations must satisfy this condition. Second, it must be possible either to specify a joint probability density for multiple, simultaneous observations or to approximate this joint density by assuming independence. Thus, for two different sensor observations, *V1* and *V2*, we must either directly estimate the likelihood of all pairs of values for *V1* and *V2* (the function $f_{V1,V2 \mid I}$) or use the following approximation:

$$f_{V1,V2 \mid I} = f_{V1 \mid I} \cdot f_{V2 \mid V1,I} \approx f_{V1 \mid I} \cdot f_{V2 \mid I}$$

The joint density can then replace the observation density for a single observation, $f_{V \mid I}$, in Equation 2.

Calculating the position density using multiple sensors is very similar to calculating the density using a single sensor. The previous score position density and the distance density, $f_{I\text{-}J \mid D}$, are convolved as per Equation 1. Next, the result of convolution and the joint observation density are multiplied and normalized, giving the final position density. Note that technically the model does not require all sensors to report observations simultaneously. However, the time between interleaved reports cannot be too small. Otherwise, unacceptable errors result from numerical integration. The minimum time depends on both the functions in the integral and the sample rate. Also, the accuracy of position estimation may vary with the frequency of observations. Position estimation also depends upon how accurately the observation density functions model the actual distributions during real performances. In the following sections, we discuss data collection and analysis to implement the vocal performance tracking system.

5. Data Collection and Modeling

Three density functions are assumed to be pre-defined prior to each calculation of a new score position density function: f_{Source}, the stochastic estimate of the performer's source position; $f_{I\text{-}J \mid D}$, the probability that the performer has actually performed an amount of score I-J given D, a prediction of the amount of score performed; and $f_{V\mid I}$, the probability of making observation V when the performer is at position I. While the first function is a prior calculation of the score position density, the second and third functions must be specified before using the model.

We analyzed actual vocal performances to estimate these densities. We recorded performances by live vocalists singing with live accompanists, and isolated the vocal part by using a highly directional microphone placed in close proximity to the singer. These recordings were analyzed for pitch content, spectral envelope, and amplitude changes as well as tempo. The vocalists had at least one full year of university training. They performed Western classical music that either was familiar to them or was part of their current program of study. The performers were given one opportunity to rehearse prior to recording. A subset of 20 performances was used to determine empirical density functions. These recordings contained 2 performances by each of 10 singers encompassing all primary voice types. Each singer performed two different works. The pieces represented a variety of styles and genres. While all 20 recordings were used to estimate the distance density, only 18 were used to estimate densities for observed pitch, spectral envelope, and amplitude changes. Two of the recordings contained a low-level background hum. The 18 recordings included 2 performances by each of 9 singers.

Although the observation density, $f_{V\mid I}$, can specify a different likelihood at every point in the score, it is impossible to collect sufficient data to model such a function. Consequently, generalizations will be made. For instance, the distribution for actual reported pitch will be based upon only the pitch written in the score. Score positions with the same scored pitch will have the same observation distribution. Subsequent definition of observation densities will specify one or more conditioning variables that replace the variable i. Score positions associated with the same values of the new conditioning variables will share identical observation distributions.

6. Fundamental Pitch

We have used a pitch detection method described in (Kuhn, 1990). We constructed our detector to use a sample rate of 16 KHz and to report pitch at a rate of 10 Hz during a sustained tone. A simple amplitude threshold is used to roughly separate pitched signal from unpitched signal and silence. Pitched portions of a vocal performance generally exhibit greater amplitude than portions where the singer is resting or producing unvoiced consonants. Eighteen of the recorded performances were played from a DAT tape and processed by the pitch detector. The output was parsed by hand in order to align the reported pitches with the notes in the scores. Next, the distance (in semitones) between the detected pitch and the scored pitch was calculated for each detector output.

A histogram of these differences was then generated. For every scored pitch, an observation distribution of actual pitch conditioned on scored pitch can be generated by mapping the scored pitch to the zero difference bin and all other notes to the bin corresponding to their distance from the scored note. Thus for a scored pitch of middle-C, middle-C maps to 0, D above middle-C maps to +2, and B below middle-C maps to -1. This distribution is used to describe observed fundamental pitch. It is conditioned only on the pitch written in the score.

7. Spectral Envelope

To obtain an approximation of the spectral envelope, we compute a short time log power spectrum using around 8 ms of signal that has been sampled at 16KHz, pre-emphasized to remove the natural spectral tilt in singing, and

multiplied by a Hamming window. Four such spectra from consecutive regions of signal are averaged and the resulting spectrum is normalized. We then extract the peak values occurring in fifteen partitions of the frequency axis. Each partition spans roughly one-third of an octave, and the partitions in total cover frequencies from 200 to 5000 Hz. Note that this representation essentially skims the tops of the harmonics present in the log spectrum, and attempts to capture properties of the vocal tract transfer function significant for perception of phonemes. As with pitch detection, we apply a threshold attempting to distinguish pitched signal (particularly vowels) from unpitched signal and silence. This thresholding synchronizes extraction of spectral features with pitch detection. The spectral information is obtained at a rate of 10 Hz during a sustained tone, using the most recent 32 ms of signal.

Eighteen of the recorded performances were processed by the spectral envelope extractor. We used the output to generate a vector quantization codebook with 128 entries. A simple Euclidean distance metric was used Next, digitized recordings of the performances were parsed by hand to identify phonetic boundaries. The time-stamped spectral representations were time-aligned with the phonetic parse. Phonetic transcriptions of all pieces were based on the IPA (International Phonetic Alphabet) commonly taught to students of singing. The recorded pieces included examples in English, Italian, and German. Distributions over the vector codebook were calculated for each phoneme occurring in the transcriptions. Note that while vowels are of primary interest, the thresholding used by the detector does not completely discard all consonants. Thus models for consonants as well as vowels were developed, but they are based only on observations actually reported by the detector. These distributions were used as the observation densities for observed spectral envelope (vector codebook entry) conditioned on scored phoneme.

For initial testing of the tracking method, we have assumed that pitch and spectral envelope are independent. We approximate the joint density by multiplying the individual densities. In reality, this approximation is certainly not accurate. The observed spectrum is influenced by the fundamental and associated partials. However, if the model data contains examples of each phoneme over most possible pitches, this assumption may not hinder tracking. Explicit modeling or better approximations of the joint density might ultimately yield better tracking.

8. Amplitude Changes Indicative of Note Onsets

Sung vowels often exhibit a significant change in amplitude when preceded by consonants, a rest, or a breath. An amplitude threshold is used in both fundamental pitch and spectral envelope detection. Our simple onset detector reports an onset when 30 ms of signal with amplitude below this threshold precedes a detected pitch. 30 ms is more than twice the pitch period of the lowest note sung by a bass, but typically less than the shortest consonants appearing between vowels. Also, detected onsets must be spaced by at least 150 ms, otherwise the detector does not report the second onset. This constraint helps to prevent duplicate reports during note transitions. 150 ms is almost always shorter than the duration of sung notes not in a melisma. The detector reports presence or absence of an onset for every observation of fundamental pitch and spectral envelope. Providing that the amplitude threshold is reasonable, this simple approach can reliably detect a useful number of note onsets.

The estimated distribution for detected onsets contains two important factors (conditioning variables). First, detection of an onset almost always occurs during or just after note transitions. Onsets are rarely reported during a sustained tone. The chance of detecting the onset of a note depends upon whether the vowel immediately follows a rest or breath, consonants, or another vowel. In addition, onsets are sometimes detected immediately before a rest. Second, when a group of consonants precedes a vowel, the number of voiced consonants in that group affects the likelihood of detecting an onset. The presence of fewer voiced consonants increases the likelihood that an onset will be observed. The distribution for detected onsets specifies the likelihood of detecting onsets according to these conditions: whether a transition between notes contains a rest or breath, consonants, or neither; and when consonants are present, both the number of consonants and the number of voiced consonants present.

The onset detector was applied to eighteen of the recorded performances. Note that during collection of all three types of observations, the amplitude threshold was held consistent per recording. Detected onsets were associated with notes in the score based on the phonetic parses previously mentioned. Probabilities were estimated for each of the conditions just described by calculating the percent of instances when an onset was detected. Note that for each condition, Pr(*No Onset*) = 1 - Pr(*Onset*). For an initial performance tracking system, we have assumed that the detected pitch, spectral envelope, and detected onsets are independent. We approximate the joint density by multiplying the individual densities for all three types of observations.

9. Model Application and Results

To assess the tracking model, we have used it to track both recorded and live performances. The tracking model was used as part of an automated accompaniment system. To produce a point estimate of the performer's position, the system determines the 100 ms region of score that most likely encompasses the performer's current location. This region is the 100 ms portion of the score position density function containing the most probability. The center of this region is used as a best estimate of the vocalist's current position.

To quantitatively evaluate the tracking system, we examine the difference between the time when the system estimates a particular position and the time that the vocalist was actually at that position. During a performance, all of the system's position estimates are recorded with a time stamp indicating the point in the audio signal last processed. The recorded performance is also digitized and parsed by hand to locate the onset time of each note in the score. The start of the vowel in each syllable is taken to indicate the onset. For a melisma, changes in pitch are used instead. The recorded position estimates are time aligned with the transcript. We calculate differences between the time at which the system first estimated a position within each score note and the time at which the performer was actually at that position. Graphs and statistics of this data are examined to assess the tracking system's accuracy, subject to the errors introduced by human generated transcripts and time alignment.

Table 1 presents summary statistics for eight recorded vocal performances. These performances were not used when estimating the density functions. Four sets of observation types were applied—fundamental pitch, fundamental pitch and spectral envelope, fundamental pitch and onsets, and all three types of observations. The values reported for each performance are averages over three trials. Statistics were calculated using all estimated positions and after removing the 5% outliers by absolute value. Note that the system combining all observation types produced smaller standard deviations for 6 of 8 performances when all data is considered and for all 8 performances with outliers discarded. The average results across the eight performances also improved. Observing spectral envelope and pitch did not produce as noticeable an improvement. Standard deviations are smaller for only 3 performances when all data is considered and for 5 performances when outliers are removed. The average with outliers removed also improved. Observing note onsets as well as pitch yielded smaller standard deviations for 5 performances when all data is considered and 7 performances with outliers discarded.

Table 2 presents summary statistics for six live performances. Two instances of the tracking method were evaluated—one using pitch and one using all observation types. Only performance six showed a drastic increase in the standard deviations. Two of three significant errors for this trial are due to spurious detection of onsets. Adjusting the amplitude threshold and repeating the performance would have improved the results. Note that this performance significantly influences the average standard deviations. Otherwise, statistics for the live performances are comparable to those for the recorded performances.

Table 1: Results of tracking recorded performances using different types of observations. Sample standard deviations for differences between the earliest time of an estimated position within each note and the time the performer was actually at that position. 5% outliers were removed based on absolute value.

	Pitch		Pitch & Spec. Env.		Pitch & Onsets		All Observations	
Performance	All Data	No Outliers	All Data	No Outliers	All Data	No Outliers	All Data	No Outliers
1	208 ms	155 ms	203 ms	121 ms	288 ms	194 ms	175 ms	96 ms
2	94 ms	52 ms	99 ms	59 ms	87 ms	36 ms	60 ms	34 ms
3	132 ms	97 ms	176 ms	108 ms	131 ms	68 ms	154 ms	76 ms
4	278 ms	209 ms	323 ms	195 ms	299 ms	194 ms	355 ms	170 ms
5	292 ms	159 ms	186 ms	89 ms	243 ms	128 ms	200 ms	72 ms
6	111 ms	75 ms	117 ms	67 ms	113 ms	63 ms	90 ms	56 ms
7	138 ms	85 ms	114 ms	65 ms	131 ms	66 ms	94 ms	60 ms
8	151 ms	108 ms	179 ms	117 ms	138 ms	71 ms	147 ms	78 ms
Average	175 ms	117 ms	175 ms	103 ms	179 ms	103 ms	159 ms	80 ms

Table 2: Results of tracking live performances using different types of observations. Sample standard deviations for differences between the earliest time of an estimated position within each note and the time the performer was actually at that position. 5% Outliers were removed based on absolute value.

	Pitch		All Observations	
Performance	All Data	No Outliers	All Data	No Outliers
1	102 ms	60 ms	97 ms	51 ms
2	181 ms	111 ms	154 ms	94 ms
3	90 ms	71 ms	79 ms	50 ms
4	104 ms	62 ms	80 ms	48 ms
5	112 ms	77 ms	116 ms	69 ms
6	179 ms	116 ms	318 ms	133 ms
Average	128 ms	83 ms	141 ms	74 ms

Generally, the statistics calculated over all time differences are influenced by a few outliers of large magnitude. The outliers indicate situations where the tracking system cannot disambiguate position based on the observations, or where the performer has made a large, sudden, and unexpected tempo change. Our models for spectral envelope often help position estimation, but are not complete enough to avoid causing an occasional confusion. Likewise, the amplitude thresholding sometimes leads to onset recognition where no onset is expected.

As an alternative assessment, the results of paired comparisons tests are presented in Table 3. These tests paired the trials using only pitch with trials using all observation types. For each pair, the time differences reported for the same note were subtracted. T-tests were run on these "differences of differences" to assess whether observed improvements or degradations were more than just random. Table 3 shows the number of trials that improved or degraded for various P-values when using all observation types. Note that for the recorded performances, many trials improved while none degraded. Only one live performance trial degraded significantly—performance six.

Mecca reports that, when playing a simple scale against a metronomic pulse at constant tempo, accompanists produce timing errors with standard deviations in the range of 5 to 48 ms (Mecca 1993). Melodic timing errors due to motor noise commonly are reported in the range of 10 to 100 ms (Desain and Honing 1992). While it is difficult to draw firm conclusions, it is probably safe to say that performance of our tracking system is not as good as a human listener. However, in many cases it is sufficient for useful accompaniment or analysis systems. For instance, our accompaniment system is not designed to jump immediately to the estimated position, but most often just increases or decreases the tempo beyond the singer's current tempo. If the outliers are not too large and occur independent of one another, only a few will significantly affect accompaniment. Also, in some trials, many of the outliers occurred in unaccompanied sections of a piece—either recitativos or cadenzas. Note that recorded performances 1 and 4 contained such sections. During testing there were no examples where the system became permanently lost or caused the live performer to stop. Finally, even in situations where problems do arise, tracking

Table 3: Results of paired comparison t-tests for recorded and live performances. Values indicate the number of trials that improved or degraded when using all observations versus using only fundamental pitch. Tests considered the difference between time differences (estimate time minus actual time) paired by note.

P-Value per Performance	Recorded Performances (24 trials, 8 recordings)		Live Performances (6 trials, 6 performances)	
	Improved	Degraded	Improved	Degraded
< 0.01	9	0	1	1
< 0.05	14	0	1	1
< 0.10	16	0	2	1
< 0.15	17	0	3	1

can be improved if performers provide score modifications indicating previously unmarked tempo, pitch, and phonetic changes.

A tracking system using fundamental pitch, estimated tempo, and elapsed time is reasonable at following performances of Western classical music by trained singers. Statistical models for pitch are good at discriminating score position, easy to define, and feasible to estimate accurately. Extending the tracking system to observe spectral envelope and note onsets (amplitude changes) often improves tracking. However, it is difficult to define and estimate comprehensive statistics for spectral envelope. The approach used by Inoue and colleagues (1994), where vocalists must sing each vowel before the system accompanies them, may be helpful. However, the dependence of the spectral envelope on the pitch may limit use of this technique. Also, the simple threshold used to detect onsets could be replaced by a more sophisticated onset detector, but probably would introduce a more complex statistical estimation problem. Better statistical models that address these problems could enhance vocal performance tracking. Also, precise models of tempo changes at cadences, cadenzas, and fermatas could improve tracking.

10. Summary

We have presented an extended statistical method of tracking a vocal performer. It incorporates multiple sources of information relevant to score following, including observed fundamental pitch, spectral envelope, amplitude changes associated with note onsets, estimated tempo, elapsed time, and the notated score. Implementation of the model is based upon empirical density functions estimated from actual performances. Testing of the system has demonstrated robust and useful tracking, although probably not at the level of a human listener. The statistical model for tracking permits incorporation of multiple observations obtained from distinct measurements of the sound signal. This approach supports incremental development and testing, allowing for possible extensions and subsequent comparative examinations of tracking systems that incorporate different types of observations.

References

Dannenberg, R. 1984. An On-Line Algorithm for Real-Time Accompaniment. In *Proceedings of the 1984 International Computer Music Conference*, 193-198. San Francisco: ICMA.

Dannenberg, R. and Joseph, R. 1992. Human-Computer Interaction in the Piano Tutor. In *Multimedia Interface Design*, 65-78. M. Blattner and R. Dannenberg, eds. New York: ACM Press.

Desain, P. and Honing, H. 1992. Tempo Curves Considered Harmful. In *Music, Mind, and Machine*, 25-40. P. Desain and H. Honing, eds. Amsterdam: Thesis Publishers.

Grubb, L., and Dannenberg, R. 1997. A Stochastic Method of Tracking a Vocal Performer. In *Proceedings of the 1997 International Computer Music Conference*, 301-8. San Francisco: ICMA.

Honing, H. 1990. POCO: An Environment for Analyzing, Modifying, and Generating Expression in Music. In *Proceedings of the 1990 International Computer Music Conference*, 364-8. San Francisco: ICMA.

Inoue, W., Hashimoto, S., and Ohteru, S. 1994. Adaptive Karaoke System—Human Singing Accompaniment Based on Speech Recognition. In *Proceedings of the 1994 International Computer Music Conference*, 70-77. San Francisco: ICMA.

Katayose, H., Kanamori, T., Kamei, K., Nagashima, Y., Sato, K., Inokuchi, S., Simura, S. 1993. Virtual Performer. In *Proceedings of the 1993 International Computer Music Conference*, 138-145. San Francisco: ICMA.

Kuhn, W. 1990. A Real-Time Pitch Recognition Algorithm for Music Applications. *Computer Music Journal* 14:60-71.

Mecca, M. 1993. Tempo Following Behavior in Musical Accompaniment. Masters Thesis, School of Computer Science, Carnegie Mellon University.

Puckette, M. 1995. Score Following Using the Sung Voice. In *Proceedings of the 1995 International Computer Music Conference*, 175-178. San Francisco: ICMA.

Vercoe, B. 1984. The Synthetic Performer in the Context of Live Performance. In *Proceedings of the 1984 International Computer Music Conference*, 199-200. San Francisco: ICMA.