

USING CONTOUR AS A MID-LEVEL REPRESENTATION OF MELODY

by

Adam Taro Lindsay

S.B., Cognitive Science

S.B., Music

Massachusetts Institute of Technology (1994)

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning
in partial fulfillment of the requirements for the degree of

Master of Science in Media Arts and Sciences

at the

Massachusetts Institute of Technology

September 1996


© Massachusetts Institute of Technology 1996. All rights reserved.

Author



Program in Media Arts and Sciences,
August 20, 1996

Certified by



Whitman Richards
Professor of Cognitive Science
Program in Media Arts and Sciences
Thesis Advisor

Accepted by



Stephen A. Benton
Chair, Departmental Committee on Graduate Students
Program in Media Arts and Sciences

MASSACHUSETTS INSTITUTE
OF TECHNOLOGY

AUG 21 1996

LIBRARY

USING CONTOUR AS A MID-LEVEL REPRESENTATION OF MELODY

by
Adam Taro Lindsay

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning
on August 20, 1996, in partial fulfillment of the
requirements for the degree of Master of Science in
Media Arts and Sciences

Abstract

A psychological experiment is described, in which subjects were required to repeat short melodies presented over headphones. Subjects' performances were recorded and analyzed, revealing that errors in musical performance followed certain patterns across all subjects. An important finding was that errors remained constant regardless of note distance in time and frequency. This result suggests that subjects use an enhanced melodic contour as their basic representational form. A model based on this finding for use in computer-driven melody recognition is proposed. Query-by-humming, other applications, and future research directions are also discussed.

Thesis Supervisor: Whitman Richards

Title: Professor of Cognitive Science

USING CONTOUR AS A MID-LEVEL REPRESENTATION OF MELODY

by

Adam Taro Lindsay

Readers

Certified by

Barry Vercoe
Professor of Media Arts and Sciences
Program in Media Arts and Sciences
Research Advisor

Certified by

Peter Child
Professor of Music
School of Humanities and Social Sciences

TABLE OF CONTENTS

	Abstract	3
CHAPTER 1	INTRODUCTION	9
	Why mid-level representation?	10
	Expectations	11
	Preliminary results	13
	Contours	15
CHAPTER 2	BACKGROUND	17
	Representations in music cognition	18
	Representation in applications	19
CHAPTER 3	EXPERIMENT	21
	Subjects	22
	Procedure	22
	Equipment	23
	Stimuli	23
	Discussion of experimental strategies	25
CHAPTER 4	ANALYSIS	27
	Segmentation	28
	Pitch tracking	29
	Note-level segmentation	31
	Event-level estimates	31
CHAPTER 5	RESULTS	35
	Absolute pitch or interval?	36
	The benefits of intervals	39
	Accuracy of interval size	39
	Variances of intervals	41
	Accuracy of accumulated intervals	43

CHAPTER 6	REPRESENTATION AND MODEL	49
	Basic representation	50
	Basic model	51
	Extensions to the basic representation	54
	Covariance matrix	54
	Piecewise-linear accommodation to subjects	56
	Ten-dimensional expansion	58
	Sign-based contour	60
	Summary	60
CHAPTER 7	CONCLUSION	63
	Model benefits and shortcomings	63
	Benefits	64
	Limitations	64
	Applications	66
	Future research	66
	The end	67
CHAPTER 8	ACKNOWLEDGMENTS	69
APPENDIX	EXPERIMENTAL STIMULI	71
	BIBLIOGRAPHY	75

Imagine the following scene: a cocktail bar, a cocktail pianist playing in the corner, and a customer approaching with a request. The customer has had a few too many, and cannot remember the name of the song he wishes to hear. But he remembers “how the song goes”. “Da-da-dah,” the soused customer brays, “da-dee-da-da-duh.” The musician then breaks into a flawless version of the standard, “All of Me,” which is what the drunk wanted to hear.

The situation seems pretty mundane, but how does it happen? It is very clear that the person with the request cannot sing well, and his pitch and rhythm are wildly inaccurate. Yet the sounds that he “sings” contain enough information for the listener to determine the singer’s intention, that is, the idea the customer tried to convey using music. This thesis examines the issues involved in answering the question, “How does it happen?” by attempting to characterize what human singers do and having a machine listening system try to replicate a human listening system.

The cocktail bar example is slightly frivolous, but it illustrates a phenomenon that most people take for granted. Nearly everyone, including non-musicians, has an incredible capacity for recognizing countless tunes. These tunes, however, need not be heard precisely as the originals: they can be recognized despite all sorts of transformations. A common “transformation” of a melody is the distortion it goes through when sung. Few but the most highly trained singers in the most controlled situations can sing every note perfectly, yet all but the most monotonic renditions can be heard as musical, and often recognized.

What cognitive mechanisms might be in use when listening to a melody? We will concern ourselves with the information gained by a listener after pitch is perceived. Therefore, we assume that pitch is processed before any decision about melody is made. This assumption may not be entirely valid if some expectations of pitch affect processing, but will be sufficient for our experimental purposes, which we will see in Chapter 5. After characterizing singers’ typical errors, we will wrap the results into a mid-level representation for use in a melody recognition model.

1.1 Why mid-level representation?

Mid-level is a somewhat nebulous term. We use that term to describe our representation for a number of reasons. It is mid-level because the information with which it deals lies between raw sample data and a melody model. It is not low-level because the information is far abstracted from a representation of a sound wave or that gained from a cochlear model. Nor is it as precisely high-level as musical notion; our pitch information is continuous, therefore containing more precise information than discrete music notation, but containing less of the crucial symbolic information that makes structure apparent.

We also consider the system to be mid-level in the sense of approach to the data. Low-level systems are most commonly data-driven: patterns are derived from the original data, without overt regard for the high-level interpretation. High-level systems, in con-

trast, are often characterized as knowledge-driven: they take knowledge about the world and sort the data based on that information. Our approach straddles the line between the two, letting patterns in the data inform our knowledge of the world, but retaining a strict, sensible *musical* interpretation over an data-driven optimal representation without clear meaning.

In general, the mixed approach works well for this problem. We are dealing with very real human data: we measure direct musical utterances from the subjects. This approach requires a sensitivity to the human errors made (errors are what we seek to characterize), as well as the ability to tease the data apart to look for unexpected patterns. The resulting representation is rich enough to capture uniquely human expressions and errors, but is general enough to serve a variety of purposes, such as query-by-humming, and other applications discussed in Chapter 7.

1.2 Expectations

What do we expect an average untrained singer to sing correctly? We expect the vestiges of rhythm and melodic contour to be correct. At the very least, we expect that when one note is supposed to be higher than the last, it will be sung higher. If a note is to be longer than the previous note, it will probably be sung longer. The notes will be approximate, but will follow the general trends displayed by the original intention.

What errors do we expect of someone singing a melody? We expect the tonic note to be somewhat approximate. Notes may vary from there, perhaps drifting sharp or flat, but most likely staying constant. The singers will have difficulty if they attempt to sing at the extremes of their ranges. Intervals are likely to be “squashed” (large intervals made smaller), and rests are likely to be “telescoped” (pauses made much shorter).

Given our expectations, what sort of data are we likely to see from singers? Since our intuitions say that extremes are distorted, we would expect a figure such as Figure 1. The plots therein predict a model in which standard error increases linearly with inter-

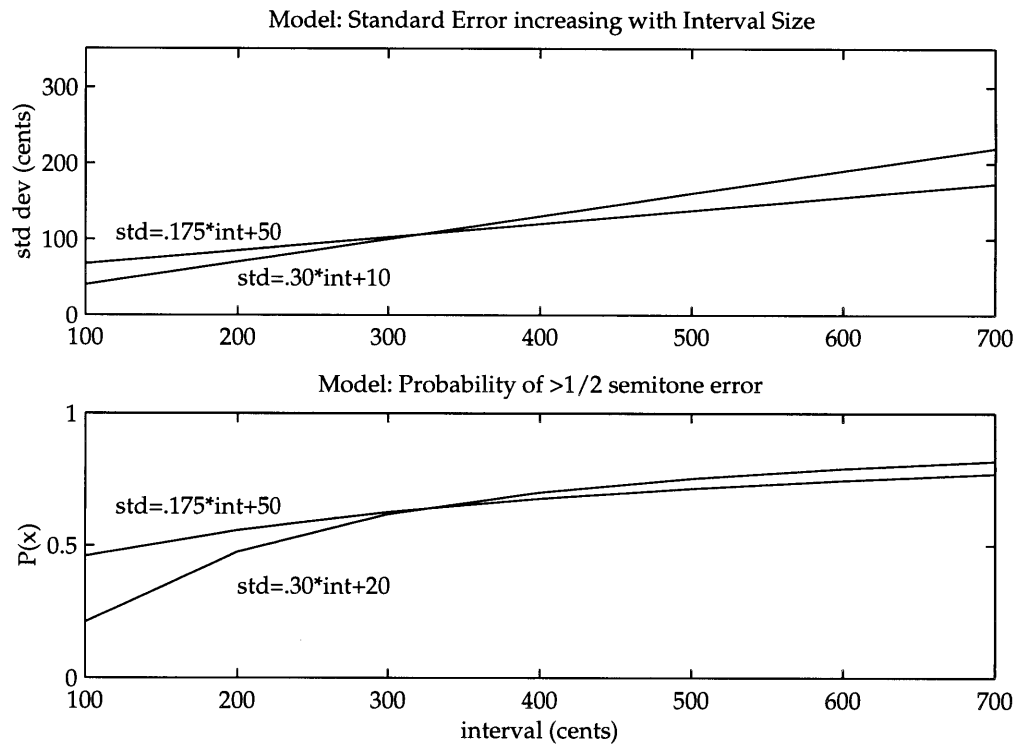


FIGURE 1. Possible models of error probability increasing with interval size. Top shows two possible linear increases of errors with intervals, and bottom shows the corresponding probabilities that the errors are too large to be correctly rounded. One semitone (equal temperament) is equal to 100 cents.

val size. We present two plausible error curves. According to our model, there is a constant percentage of a given interval in error, therefore, a larger absolute error as intervals get larger. The increasing absolute error results in an increasing chance of the error being larger than one quartertone (also 50 cents, or 1/2 semitone). We focus on quartertone errors because if we know the error to be less than that, obtaining musical intention would be as simple as rounding a sung pitch to the nearest semitone.

We would also expect most of the error to be carried over from one note to another: we mentioned singers drifting sharp or flat, above. If errors accumulate, then the chance of a quartertone error will be greater later in the stream of notes than earlier. Perhaps we would see the first interval and last interval in a series be more accurate than the ones in

the middle, observing a primacy/recency effect. Such expectations would result in a model such as Figure 2. The increase of the standard deviation of error with the square root of the number of notes results in a less dramatic rise in the probability of a large error than in Figure 1, but the effects are visible, nonetheless.

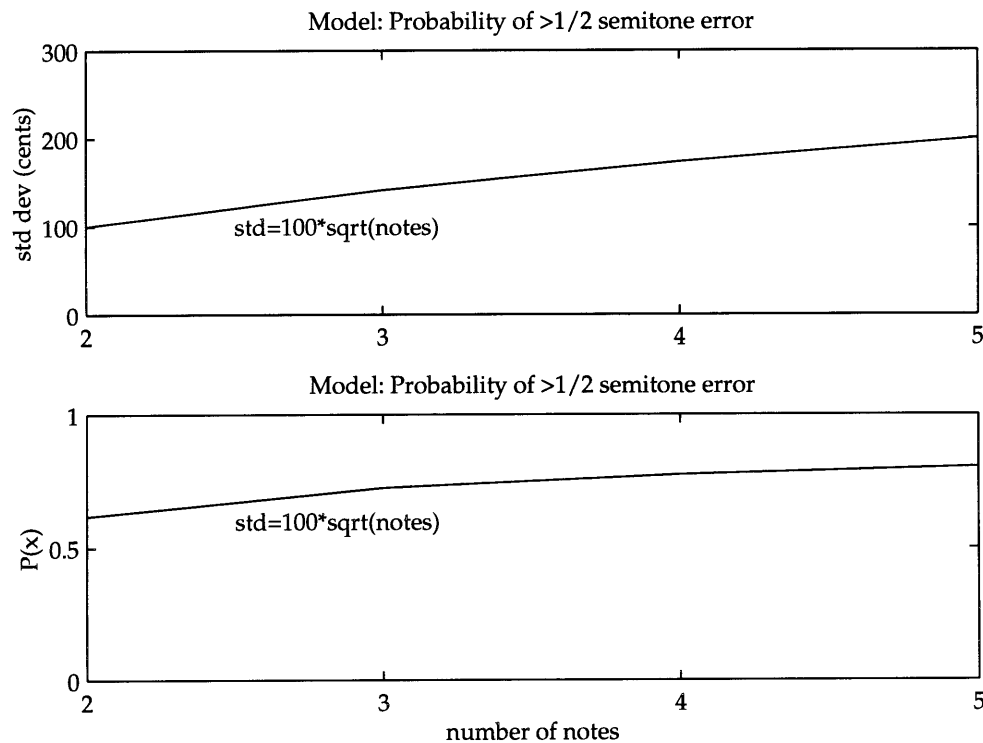


FIGURE 2. Possible model of error probability increasing with number of interceding notes. This model shows a less dramatic probability of a large error.

1.3 Preliminary results

The previous figures reflect a basic assumption: errors accumulate with the difficulty of the task. We presume it to be more difficult to sing a large interval than a small one. It should be more difficult to sing several notes in tune than a pair of notes. As the results of our experiment will show, such assumptions are wrong. Our subjects made constant-sized errors across both of these conditions. Some small increases in error with the

degree of difficulty were observed, but nothing approaching the dramatic linear increases in these models.

We now compare some of the experimental results to come with the hypothetical models presented above in Section 1.2. Figure 3 shows our preliminary model against our experimental results. The standard deviations of the errors, and therefore the probabil-

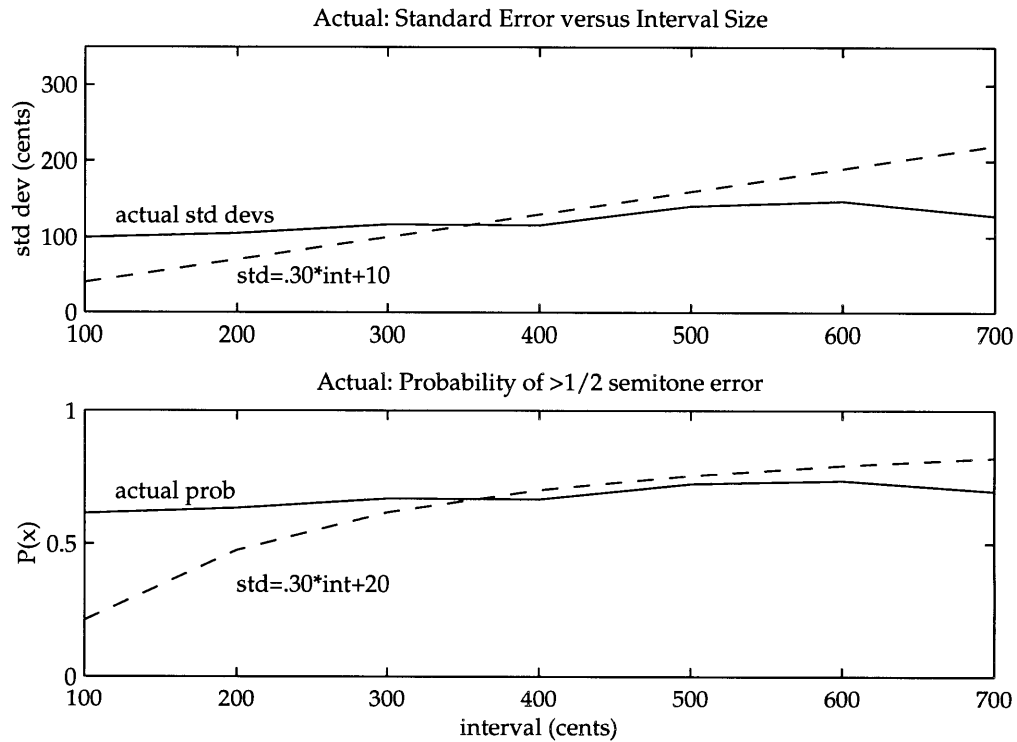


FIGURE 3. Experimental results are compared with the possible model established above. The standard deviations of the errors are essentially constant, showing that it is not more difficult to sing a wide interval than a narrow one.

ity of large errors, remain constant, no matter how large the interval is.

Similarly, though less dramatically, the number of intervals between two notes has no effect on the spread of the error. Figure 4 shows the experimental data to come against the model established in Figure 2.

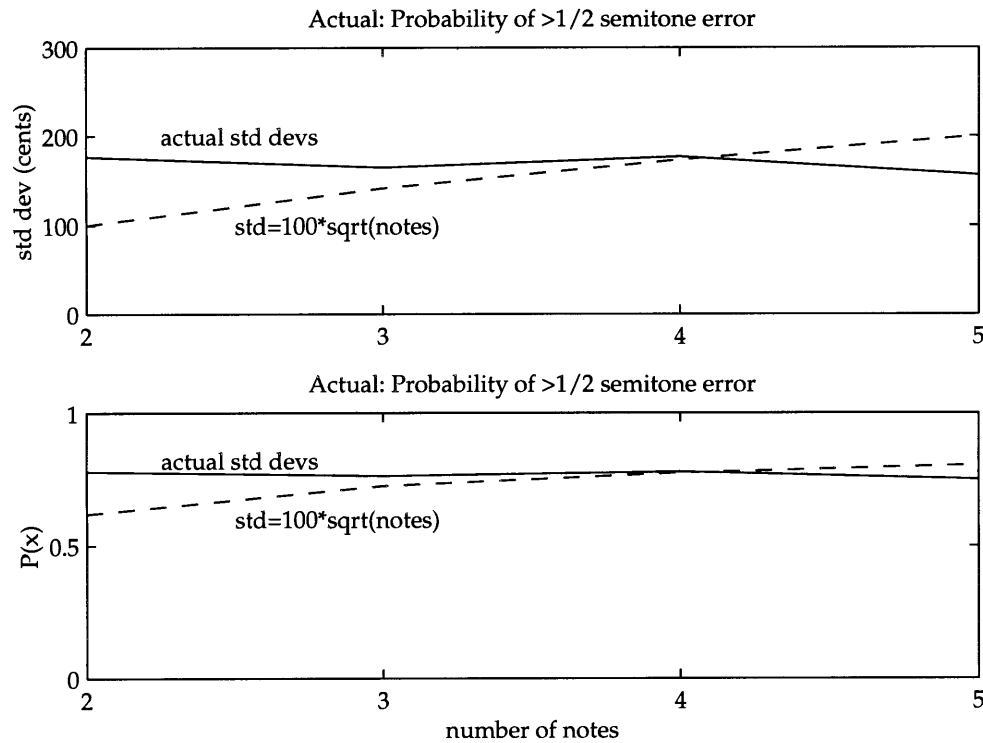


FIGURE 4. Experimental results compared with the model of error increasing with note distance in time. Although the distinction is slight, it is apparent that the experimental results reflect a constant probability of large error.

Although initially perplexing, the implications of the experimental results become much more useful than our expectations for our intended purposes. They allow for a more elegant model of human vocal performance than would be allowed by the above expectations. Such simplicity is well worth the surprise.

1.4 Contours

The underlying discussion through this chapter, indeed, this entire thesis, has been how to approach amateur singers' approximations to melodies. A common musical term which could be applied to this area is melodic contour, and different interpretations of that term will be discussed in the next chapter.

Chapter 3 and Chapter 4 are concerned with an experiment designed to refine our notion of contour in the context of human singing: the former detailing the experiment itself, and the latter detailing the various forms of data analysis. Chapter 5 discusses the results and implications of the experiment, leading to a workable melody recognition model in Chapter 6. Chapter 7 summarizes, presents potential applications for this research, and outlines future research directions.

For a musician, the word “contour” is difficult to define. All musicians have a sense of what it is, and can give examples, but few can sum up the concept well in a sentence or two. Simply put, melodic contour is “the up-ness and down-ness of the notes in a melody.” This definition is acceptable as a starting point, but does nothing to capture the *gestalt*, neither of the term or the actual melodic phrase. Contour involves a metaphor of motion. Where does the melody *go*? Where does the line begin and where does it end? How fast does it rise and fall? The idea of flowing, continuous motion is powerful, made only more striking when one considers that, as they are normally heard, melodies are discrete steps. As we note below, Gjerdingen [1994] treats this aspect of apparent motion quite nicely.

2.1 Representations in music cognition

Contour in the music cognition literature has been represented by a sign that one note is higher than, lower than, or the same as the previous note. This ternary (+/-/0, also up/down/same, or u/d/s) representation has been accepted as the standard definition of melodic contour, drawing upon analogies with early visual perception systems. Dividing contour into these three discrete steps has endured despite its limitations when compared to a musician's intuitive definition of contour, and despite the limitations of analogous visual systems.

Handel [1989] summarizes, stating that contour is "the sequential pattern of +, -, and 0." Researchers distinguish contour from interval representation in articles such as Dowling [1984] and Edworthy [1985]. Dowling notes that inexperienced listeners represent melodies as sequences of intervals, more experienced listeners use a scale-step representation, and professional musicians are capable of using a flexible representation scheme.

The alternate path in music representations has been to take a structural approach to melodies. Deutsch and Feroe [1981] present a hierarchical representation for melodies. The model takes such phenomena as "chunking" into consideration (indeed, chunking, or grouping into simpler units, is one of its main procedures), but assumes that pitch is already perceived and that a high-level structure is established. Lerdahl and Jackendoff [1984] present a hierarchical musical grammar that encompasses large musical structures as well.

Gjerdingen [1994] takes the opposite approach from the above discrete representations. Given a discrete representation of a melody, he essentially smooths it to make it a continuous percept. Although this continuous representation is opposite from the direction we wish to go, it serves as a reminder that our perception of melody is not necessarily discrete in time and frequency. Cole [1974] warns that a precise notation comes at the price of limiting melody to "that which can be notated by a series of fixed pitches." Melodies have a shape, which is illustrated by a musical notation used in Tibetan Buddhist

chant, seen in Figure 5. This culture has retained the notion of melodies being continuous movements. It is in this sense that we use the word “contour” through most of this thesis.

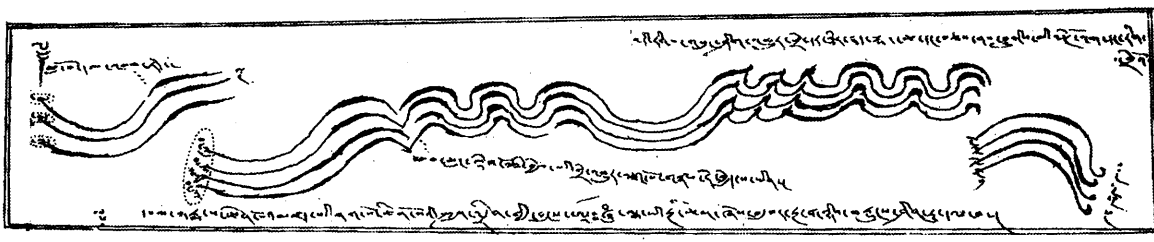


FIGURE 5. An excerpt from a *dbyangs-yig*, a Tibetan Buddhist songbook. Note the clarity of the notational curve. The rendering of the parallel lines is an ornamental notation because this is the beginning of a song. (Reproduced from Kaufmann [1975]).

2.2 Representation in applications

The primary application of this field of study has been in indexing databases by content, specifically, query-by-humming. Not surprisingly, the representations used in the applications have been very influenced by research in music cognition. Recent papers (Ghias et al. [1995], for example) use the simple ternary system of contour to index into a database of melodies.

It would seem that this u/d/s representation has been a success since the systems that use it appear to work well as presented. However, such systems require ten or more notes as input (which is a sizable portion of many melodies), and by their nature cannot distinguish between two melodies with the same up-down contour, even if the notes are vastly different. There is also no accounting for global change, which is illustrated in Figure 6. The sign representation may be identical for two melodies, but since no sense of interval size is indicated, one does not know where the melody goes over time. In short, this ternary representation does not capture the richness of what musicians instinctively call contour.

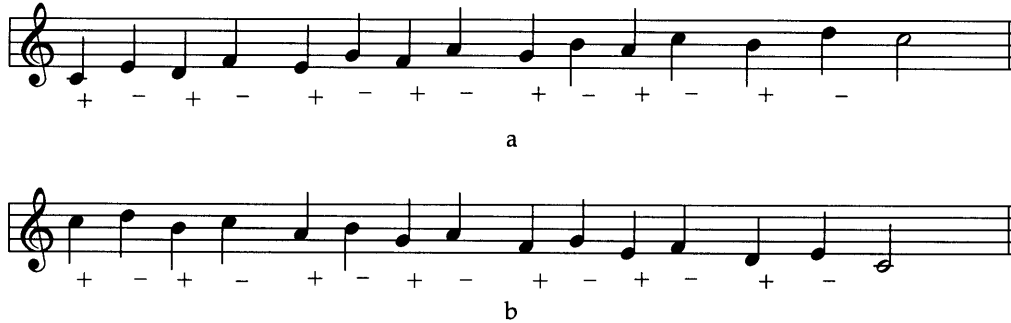


FIGURE 6. The shortcomings of a simple contour representation. These two sequences are indistinguishable using a ternary (+/0/-) representation.

We note that a hierarchical version of sign-based contour would go far towards capturing a musical sense of contour. Using another level of +/-/0 signs, grouping two notes at a time, would resolve the ambiguity in Figure 6. This approach is a possible future research direction, but it becomes implicit in our interval representation, outlined below and detailed in Chapter 6: by combining intervals, global trends become apparent.

Kageyama et al. [1993] appear to use a sort of scale-step representation, using more information than simple ternary contour, though the details are somewhat sketchy in the available proceedings paper. Despite such sophisticated techniques as dynamic time warping, their system requires over a dozen notes as an index, and has limited accuracy. It is clearly a working system, however, with an extensive 500 song database.

We will develop a definition of contour which involves not only the signs of intervals, but approximations to the intervals themselves, therefore more strictly qualifying as an interval representation. The details of our representation will be informed by the results of the experiment detailed in the next chapter, which examines human behavior when reproducing melodies.

We wish to characterize the natural musical response as described in the introduction. We see the unrehearsed musical utterance as a window on internal representation. There will always be mistakes in musical expression, but if we determine the features of the errors that remain constant across all levels of ability, we see part of the underlying representation. Put another way, if a feature appears in all renditions, it probably corresponds to something in the mind. Furthermore, the pattern of errors will also reflect properties of the underlying representation.

So, in order to characterize normal human performance in recreating melodies, we had to devise a situation natural enough for subjects to approach in a “casual” musical way. Naturally, a casual setting would not be rigorous enough for experimental purposes. As a result, we considered a call-and-response paradigm. It is basic enough for any subject to understand, especially in a musical or speech domain (“repeat after me”), but it is also justifiable in psychological terms as a simple stimulus-response pair.

We also note the importance of recreating entire musical phrases, not individual notes. The two tasks are different and should not be confused.

3.1 Subjects

Six subjects volunteered to participate in this experiment. Five subjects were between 21 and 30 years of age, and the sixth was over 50. There were two males and four females. There was a representative combination of musical backgrounds among the subjects: there was one non-musician with less than 5 years musical exposure, four amateur musicians (one vocalist and three instrumentalists) with more than 12 years of experience, and one conservatory-level singer with 20 years of training.

3.2 Procedure

The subjects were brought into a soundproof booth. There they filled out the requisite forms and were told of the basics of the experiment:

“This is an experiment in melody. You will sit in front of the computer with the headphones on. There will be a series of musical notes played through the headphones for you to repeat. Immediately after the series of notes end, you are to sing them towards the microphone on top of the computer monitor. There is a loudness meter on the computer screen; please sing as loud as you can without letting the meter show red. The first, practice trials are single notes. There will be a pause after the note for you to sing, and the next note will come automatically. I can answer any questions you have about the procedure after the practice set.”

The experimenter was present for the single-note practice trials. The subjects understood the instructions well enough to begin the task, but invariably were caught by surprise by the beginning of the second note. The trial run was sufficient training for them to settle into the “rhythm” of the trial set.

3.3 Equipment

The computer mentioned above was an Apple Power Macintosh 8100/110 running Opcode's StudioVision AV. The trial sets were MIDI sequences with pauses in between phrases (as described below in Section 3.4). The sequencer recorded directly onto hard disk from the Apple Plaintalk microphone placed in front of the subject. This microphone was designed to receive such utterances from two feet away. The entire trial made up one soundfile to be later segmented as described in Chapter 4.

Each of the trials was synthesized by a Boss DS-330 MIDI module using the "Piano" timbre and presented through AKG-K240 headphones.

3.4 Stimuli

The stimuli were designed with several factors in mind. With respect to the subjects, the melodic phrases had to be relatively easy to remember in order to be sung, as well as be constrained in pitch range enough to be sung. In order to be suitable stimuli for the purpose of later analysis, they had to cover a wide variety of melodies, and have a relatively equal distribution of intervals.

In keeping with the need for ease of singing by the subjects, the stimuli were designed to make sense in a tonal context (if the phrases themselves were not strictly diatonic). Five-note phrases seemed to make the most sense in this case: they were easily spanned by short-term memory, but were long enough to be "melodies", or at least melodic fragments.

The phrase length of five notes also worked well with the constraint of variety of melodies. We judged the variety of melodies by number of different possible sign-only contours. For five note phrases, there were sixteen distinct up-down contours (five notes yield four intervals: $2^4=16$). By carefully writing two five-note phrases for each of the sixteen contours, we were able to distribute each of fourteen intervals (± 7 semitones)

among 32 trials. For further details, including a complete list of musical stimuli, see the Appendix.

The resulting phrases each were interpretable in a tonal context; however, because of the limitation of including each chromatic interval between an ascending perfect fifth and a descending perfect fifth, the trials taken as a whole did sound a bit odd. In particular, there seemed to be a preponderance of the tritone: since ± 6 semitones had to be as common as every other interval, the tritone appeared more frequently than normal in traditional western music. As a result of this effect, these somewhat unusual phrases were more difficult to sing than an average melodic fragment, but they were by no means impossible to sing.

Because the trials were synthesized, we were able to take advantage of MIDI pitch bend. That is, each trial was altered by a random amount between -200 and +200 cents. By continuously randomizing the pitch bend value of each trial, we could eliminate any consistent tonal base across multiple trials. The microtonal alteration thus set up an orthogonal pitch axis (Handel [1989]) for each trial.

Although it may be argued that requiring subjects to replicate microtonal variations, even between trials, is unfairly difficult, the five note phrases were tonal enough to set up a salient context within the trial. Furthermore, this issue is largely irrelevant to the analysis, since almost all of the statistics were derived from the intervals presented (rather than the absolute pitches), which were strictly chromatic.

The trials were presented at a tempo of $\text{♩} = 240$, or four notes per second. A five note trial was then 1.25 seconds long. This fast presentation was to ensure that the responses were immediate and not over-considered. The goal was to get a basic reaction from the subjects, not an accurate rendition.

3.5 Discussion of experimental strategies

This issue of getting a basic reaction is at the heart of our strategy for the experiment. As we stated at the beginning of this chapter, we wish to characterize a typical spontaneous musical utterance. Because most non-musicians and many amateur musicians are shy about their singing voices, we wanted an approach that prevented the subjects from thinking too hard about their singing. This did not necessarily mean making the trials too easy: if the phrases were overly simplistic, there would be not enough errors to analyze later. In short, we wanted a fair number of errors from the subject. Nearly all of the subjects' renditions of the trial phrases were recognizably similar to the stimulus, yet none were note-perfect. We sought to uncover what is "similar enough" about the subjects' responses to generalize to a resemblance between any spontaneous musical utterance and the utterer's musical intention.

Thus the trials were tonal, had one of sixteen (binary) contours, and had an equal distribution of intervals. They were presented with relatively small pitch perturbations held for the entire trial. The phrases, as presented, were easy enough to be repeated by any subject, but difficult enough that no subject could be completely accurate.

The trials were presented in a pseudo-random, interleaved order. No trial was presented close in time to another with similar contour, nor was a trial presented close to its inverted contour. There was a three second pause between trials, which was ample time for an alert subject to repeat the phrase and prepare to hear the next.

Given our rich data of subjects singing responses to melodic phrases, how do we interpret them? The prevailing assumption is that the subjects' responses are approximations to the stimulus phrases, in a way similar to a spontaneous rendition being an approximation to the intended result. Here we see the connection between musical intention and musical expression: the pitches *expressed* do not necessarily match the pitches *intended*.

Although one may instinctively try to examine the actual pitches sung and compare them with the stimulus pitches, a more appropriate method, as explained in Section 5.2, is to examine the pitch intervals, or, the differences between two pitches. We will show that subjects make more errors singing absolute pitches than relative pitches and that by using relative pitches, many useful musical side-effects, such as key independence, result. In order for us to see those results, we must first get the data into a usable form.

The data analysis and matrix manipulation system, MATLAB, was invaluable in processing the data. Although its characteristics as an interpreted scripting language limit its further use with a real-time system for applications, it served well as a research tool to establish the parameters necessary for such a future system.

4.1 Segmentation

As noted in Section 3.3, the entire trial sets were recorded as one long sound file. While it was easy to collect the data this way, it was not easy to process. The multi-megabyte file was not even able to be loaded into MATLAB in full. The first segmentation script had to read in manageable chunks of sound, separate the trials from the silence (and each other), and save each trial in a pre-determined place.

Our first MATLAB script accomplished these goals. The script's function is summarized in Figure 7. It would read in a portion (five seconds) of the trial set sound file (in aif-format). It then squared and low-passed the segment, to reflect the magnitude of sound during those five seconds. MATLAB then hopped through the file (not testing the loudness at every single sample, rather, every fifty samples because the rate of change of power was slow enough), testing if the magnitude rose above a threshold of 0.001 (-30 dB). When the threshold was crossed, it signified the beginning of the next trial in the sequence. When the magnitude fell below the same threshold, the script marked it as a possible end point for the phrase, but checked the magnitude for the next half-second, to determine if it rose above threshold again. If not, then it was the true end of the phrase; if the sound got louder within that period, the subject hesitated but continued.

Having marked the beginning and end of one trial in the original five-second segment, the script saved the phrase (according to the order it came in during the trial set) in a filename reflecting the original order of the trials (as shown in the Appendix). If the end of the segment was found before the end of the trial was reached, the script re-loaded five seconds starting from just before the beginning of the trial. In any case, as one trial

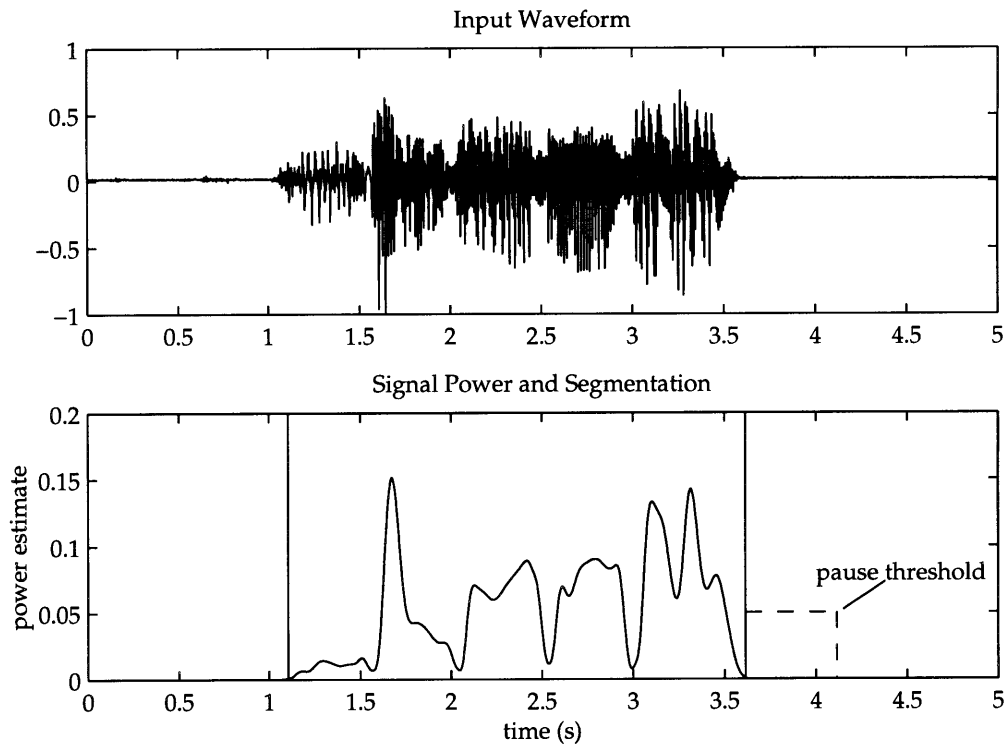


FIGURE 7. Summary of large-scale trial segmentation script. MATLAB read in a five second section of sound, as represented by the waveform in the top plot. The bottom plot shows an estimate of the power at each point, along with the estimated beginning and end of the trial, as computed by the script. Also note the 0.5 second threshold after the end, to account for the possibility of a hesitation.

was found and saved, the next five-second segment was loaded. This approach to segmenting the trial set into manageable “chunks” was very successful.

4.2 Pitch tracking

Once the individual trials were separated, we had to pitch track and note-segment the five-note phrases. The pitch tracker was one designed and used by Brown [1992]. It was a relatively robust constant-Q transform pitch tracker. After a frequency analysis and converting to a logarithmic spacing in the frequency domain, the pitch tracker smoothed across several frames to improve stability against spurious octave errors and

dropouts (which are a common problem for every frequency detector). The pitch tracker then used a pattern-recognition technique to obtain precise pitch values: it matched the frequency profile at every time-slice with an expected harmonic profile (here, usually three to six harmonics).

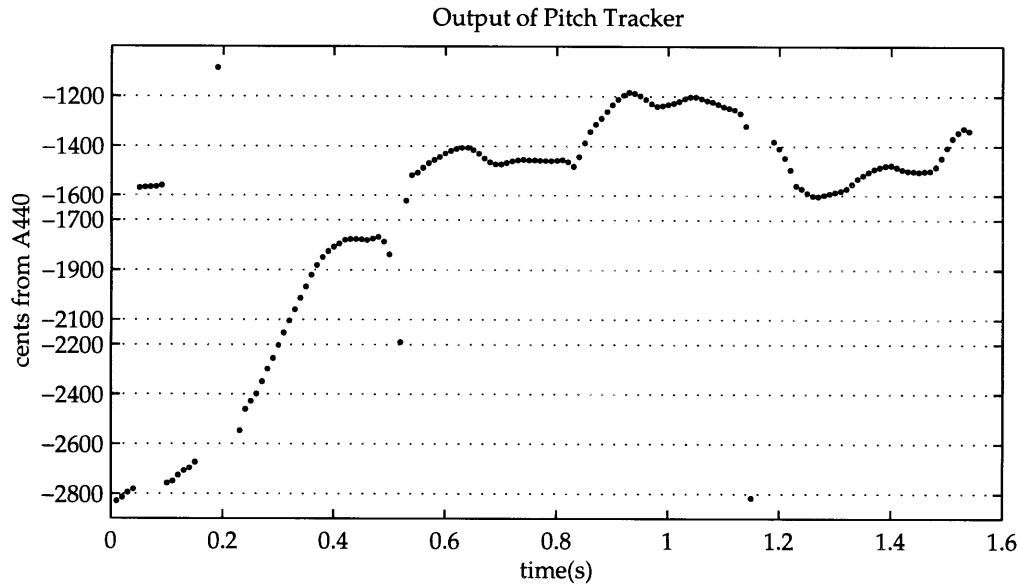


FIGURE 8. Output of Brown's [1992] pitch tracker. The above figure is a fair representation of the sort of basic data we obtained. Note the octave error in the first 100 msec, and dips in frequency at the note transitions.

Although the pitch tracker is capable of very fine frequency resolution, and is relatively stable, the speech-like qualities of the fast five-note phrases "confused" the pitch tracker often. Despite our efforts to smooth the data, there were occasional drop-outs and octave errors, artifacts for which we had to adjust later. Figure 8 demonstrates a typical pitch track for a trial by a musically experienced subject.

Given a continuous (100 frequency values per second) stream of pitch data, we had to find a way to compare it with a discrete set of five note values. Because we did not know *a priori* where one note ended and the next began, we had to determine such things before comparing one set of values with the other. We could compare the sung

pitches with the stimuli by matching the “rhythms” (starting and ending times) with each other, or discretize the frequency stream into pitch estimates for each note for comparison with the stimuli. We did both, but relied more heavily on the latter, as discussed in the following section.

4.3 Note-level segmentation

After having segmented the notes on a phrase level, we segmented them on a note level, also. We squared and lowpassed the signals as before to get an energy estimate. We converted to a logarithmic (decibel) scale at this point for the purpose of easy interpretation. Taking advantage of the explicit instructions to the subjects to sing syllables such as “da” (or anything separating the notes with a consonant), we noticed a very characteristic dip in energy between the notes. Therefore, treating the discrete-time difference as an approximation to the derivative, we compared the second difference in energy to a threshold of 1 dB/centisec², as shown in Figure 9.

This kind of automatic segmentation worked well in most cases. Different subjects required different lengths (of 150-200 msec) of inhibition to eliminate double-hits after each found note boundary. In some (less than 1%) of the segmentation cases, the system failed to find a note boundary, and we entered a guess based on the information in the pitch track. It was satisfying to note that for every energy contour where a note boundary was visible to the eye, the system was able to find a suitable boundary: the only failures were when the subject did not clearly articulate a consonant between notes.

4.4 Event-level estimates

Once the trials had been separated from each other to be pitch-tracked, and further divided into notes, we could obtain estimates of pitch for each note sung. Although the pitch estimate process discarded much of the information in the pitch track, we retained all that was necessary for comparison with the intended pitches (that is, the stimuli

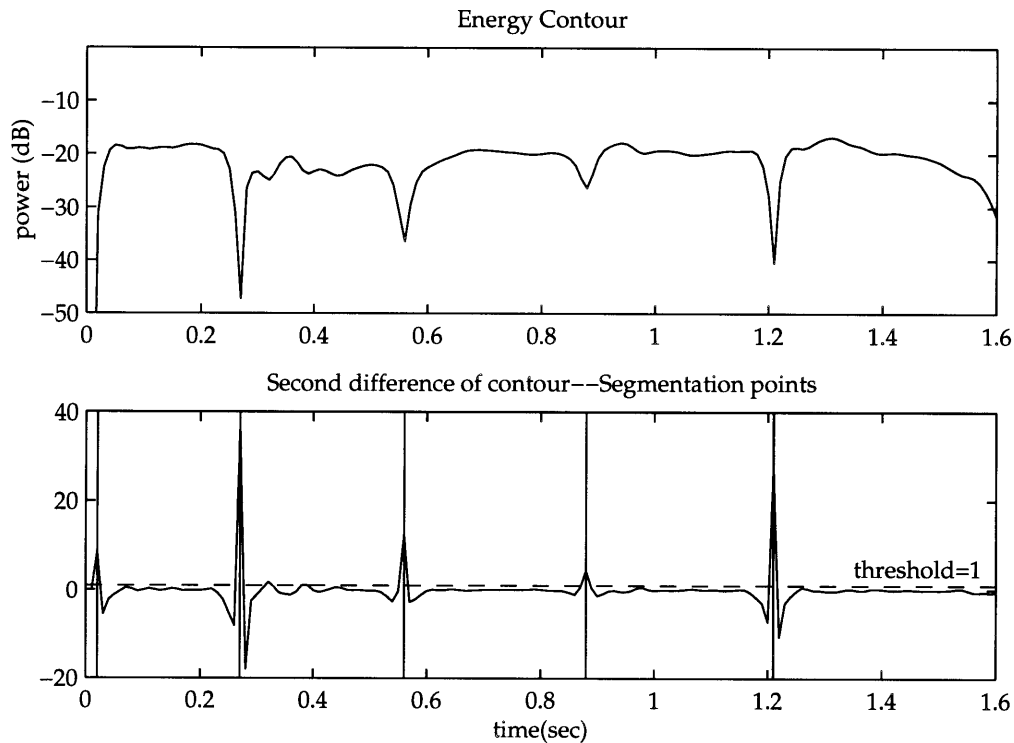


FIGURE 9. Segmentation on a note level. The above plot shows the energy of the phrase. Note the abrupt positive change in slope of energy between notes. The second plot compares this second difference with an experimentally-produced threshold of 1.0. Not shown is the 200 msec inhibition on secondary "hits" after a note boundary is found.

which the subjects attempted to repeat). Thus, the data with which we primarily worked were continuous-valued (in the frequency domain) pitches discretized in the time domain to individual events.

For us to reach this precise pitch-event representation, we needed to find an average pitch for each time interval demarcated by the note boundaries. The solution suggested by various frequency-modulated pitch perception experiments (such as Brown & Vaughn [1996]) was a magnitude-weighted mean frequency. Unfortunately, the quality of the output of the pitch tracker was not sufficient to support such an approach. The tracker made no continuity assumptions about its input, and a single octave error was enough to significantly skew the result of the mean.

With these limitations in mind, we chose to take the median pitch value for each note. Given 20 to 30 values for each note over the time held, with the data devoid of pitch track errors, the median was extremely close to the magnitude-weighted mean. In the case of an error-ridden pitch track, the median was a far better approximation to the ideal pitch track than the mean. The typical performances of these two methods are illustrated in Figure 10.

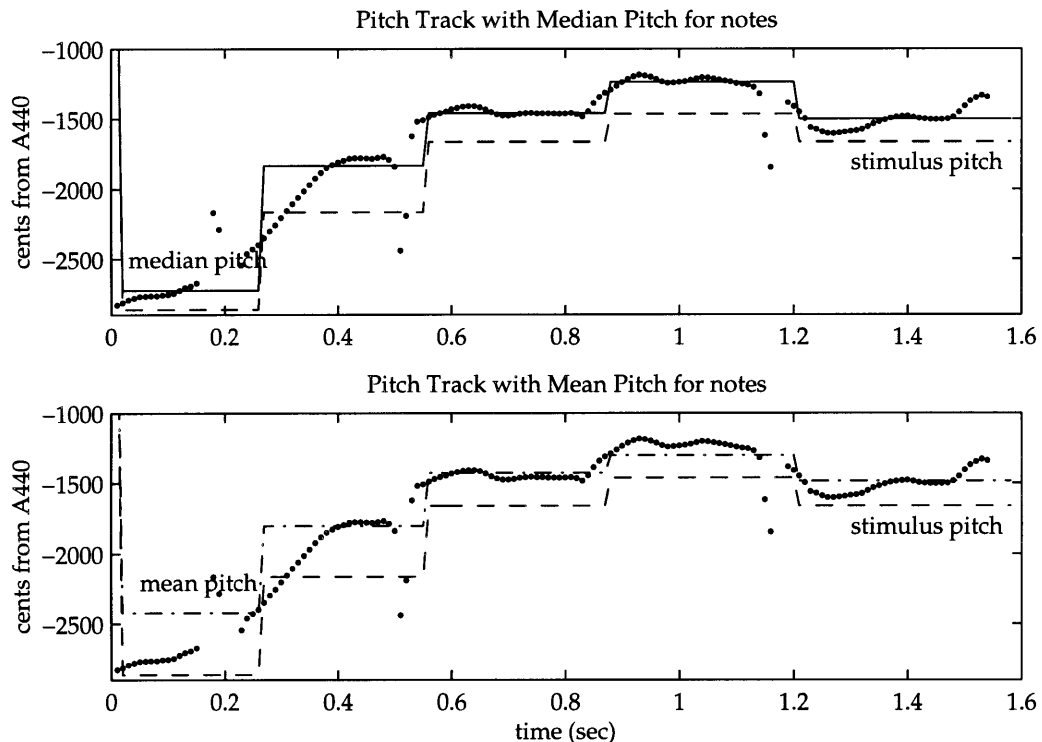


FIGURE 10. Event-level pitch estimates using median and mean. The octave errors and drop-outs in the first note, although minimized, adversely affect the mean-based approximation. The median remains a good approximation. In the case of the last note, when the pitch track is free of errors, the median and mean estimates are very close.

By applying the median estimate to each note, we obtained a simple list of each note in each trial, to be compared with the list of respective stimulus notes. With this continuous-pitch/discrete-time representation in place, we proceeded to look at the experimental results for each subject.

Now that we have an event-based approximation of a subject's sung pitch for each stimulus note, we must turn to determining the relationship between the two. Are our initial expectations correct? Do people actually make regular mistakes? Is there enough information remaining after the errors are taken into account to determine the singer's intention? These questions will be addressed in this and the next chapter.

By looking at pitch plots such as the one in Figure 10 on page 33 (top), we begin to see some common characteristics in the subjects' musical utterances. We note that the entire phrase is about two semitones sharp compared to the stimulus pitches. However, the median pitch estimates follow the same shape as the stimulus, with the same approximate intervals. The second note, for example, may be higher in relation to the stimulus than the first note was, but the third note "gets back on track." All of these observations will be borne out in the following chapter, by looking at all the pitches for all subjects.

5.1 Absolute pitch or interval?

Since we have good approximations of pitch for each note, we must decide whether absolute pitch or relative intervals will form the basis for our representation. Our bias in this matter has been evident from the start, but so far we have not ruled out the possibility of absolute pitch being a valid representation for our purposes.

Absolute pitch is the more easily comprehended representation. People sing notes, not intervals. Interpretation quickly becomes non-intuitive when one attempts to consider intervals and how they interact: when an interval is “sharp”, it means something different depending on whether the interval is ascending or descending; two notes in succession are easier to understand than two intervals in succession; errors in absolute pitch are independent from each other. All of these are valid arguments against using interval size in our representation, but, as we will show here and in Section 6.3, relative intervals are superior for our purposes.

The first piece of evidence lies in considering all of the notes and intervals in aggregate form for each subject. As described in Section 3.4 and the Appendix, the stimuli were designed with this in mind: each interval was to be equally represented across all trials. Similarly, the stimulus absolute pitches were relatively evenly distributed. By looking at simple scatterplots of the given pitches versus sung pitches (Figure 11) and of given intervals versus sung intervals (Figure 12), we can get a sense of the accuracy in each.

We do see in comparing Figure 11 with Figure 12 that for each subject, there is a larger amount of “scatter” for the absolute pitch than for the intervals. There is a higher correlation for each of the interval plots than for the corresponding absolute pitch plots.

Although these correlations do not allow for a simple t-test comparison, we observe that five of the six subjects’ correlations are less than their corresponding interval correlations, which occurs with a $p < 0.11$. That figure is not very strong, but we will not let it discourage us. The basic idea is that the error is greater for any given note than for any given interval.

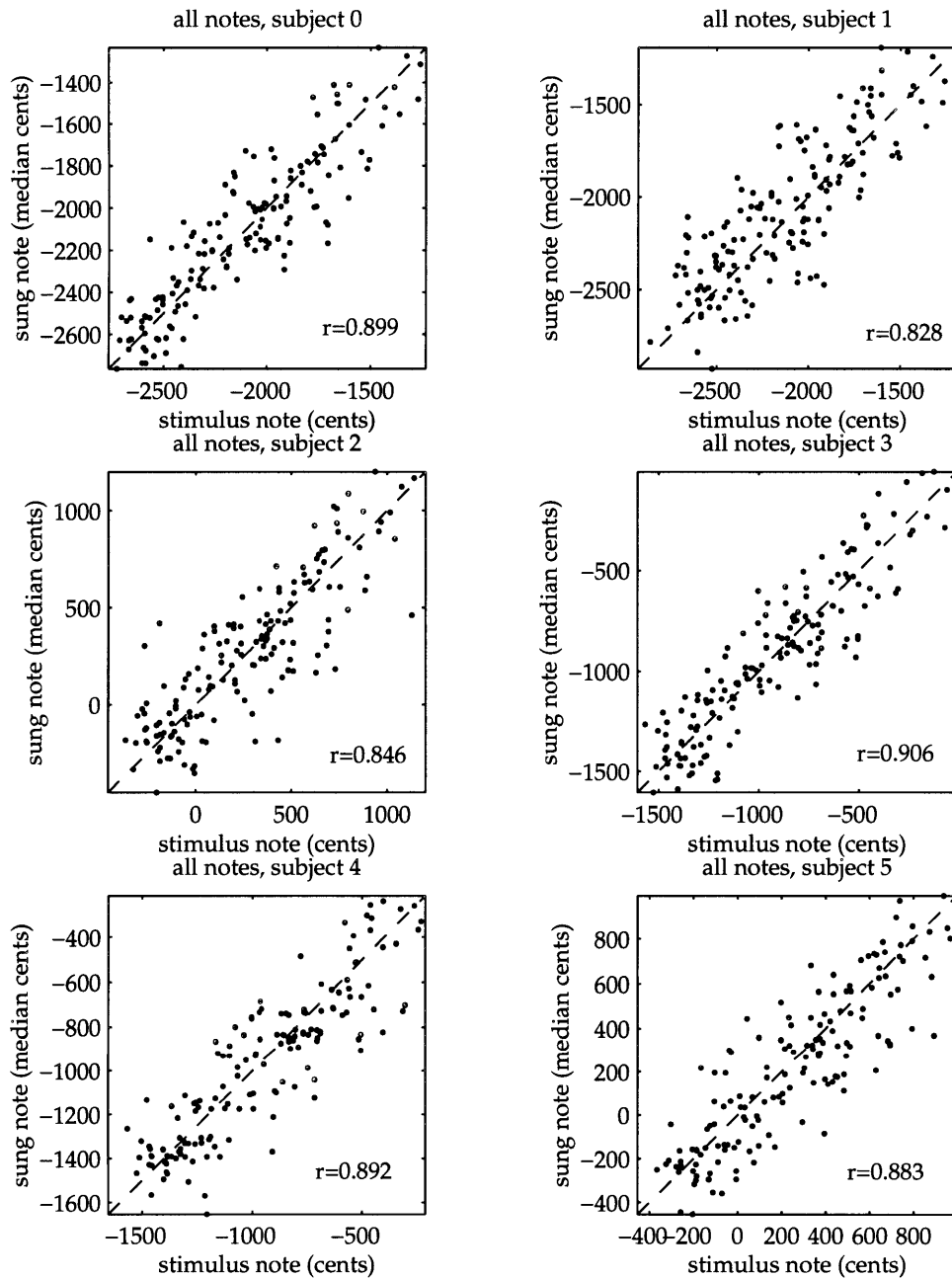


FIGURE 11. Correlation between absolute stimulus pitch and absolute sung pitch. It is worth noting that the pitches do not deviate from the ideal line at the extremes of the singers' ranges. There is no discretization on the x-axis because of the random pitch perturbation for each trial.

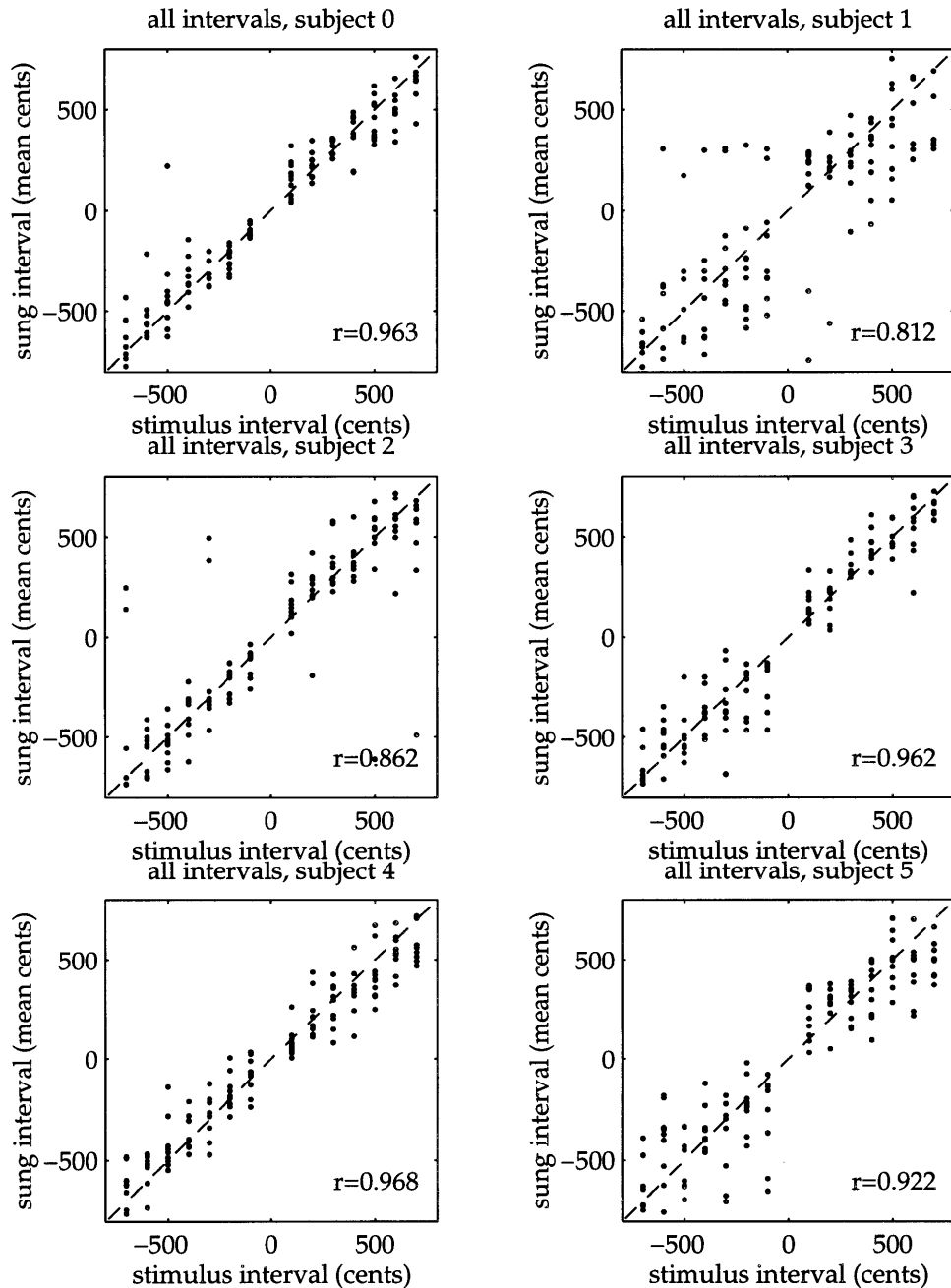


FIGURE 12. Correlation between stimulus and sung interval. The data on the x-axis is discretized because subjects were presented with only chromatic intervals.

5.2 The benefits of intervals

Using interval as a basis for our comparisons confers several benefits over using absolute pitch. A significant one is key independence. Since relative intervals exist without reference to a key, they allow for much more flexible input to, for example, an index of melodies. For absolute pitches to do the same, they must be manipulated, such as subtracting the mean pitch for the trial or making the pitches relative to the first note. These two alternate pitch methods not only throw out the information that makes them unique from an interval representation, but in the second case, can be completely represented by the interval scheme. (i.e. if pitch2 through pitch5 have pitch1 subtracted from them, then pitch2 is equal to interval1, pitch3 is equal to interval1+interval2, and so on.)

Further results below, such as the consistency of variances (in Section 5.4) and the eigenvectors of the covariance matrices (in Section 6.3), add support to our use of intervals in the analysis.

5.3 Accuracy of interval size

Now that we have decided to focus our investigation on intervals, what trends do we expect to find in the data? How does the sung interval interact with the stimulus interval within this stream of notes? One of our first expectations as explained in Section 1.2 was that extreme intervals were to be “squashed”: large intervals would be sung smaller than they really are. The effect, shown by the negative slope in Figure 13, is slight, but more pronounced in the odd-numbered subjects.

The discontinuity in the mean error curve, combined with the negative slope of both segments, supports a strong division between up- and down- intervals. That is, there is a strong grouping among the intervals going up, and the intervals going down. Interestingly, this lends credence to the use of the various sign-based contour representations. We do not doubt the general usefulness of that approach, but, even with the worst singers, there is more to be learned from their data than a simple binary division.

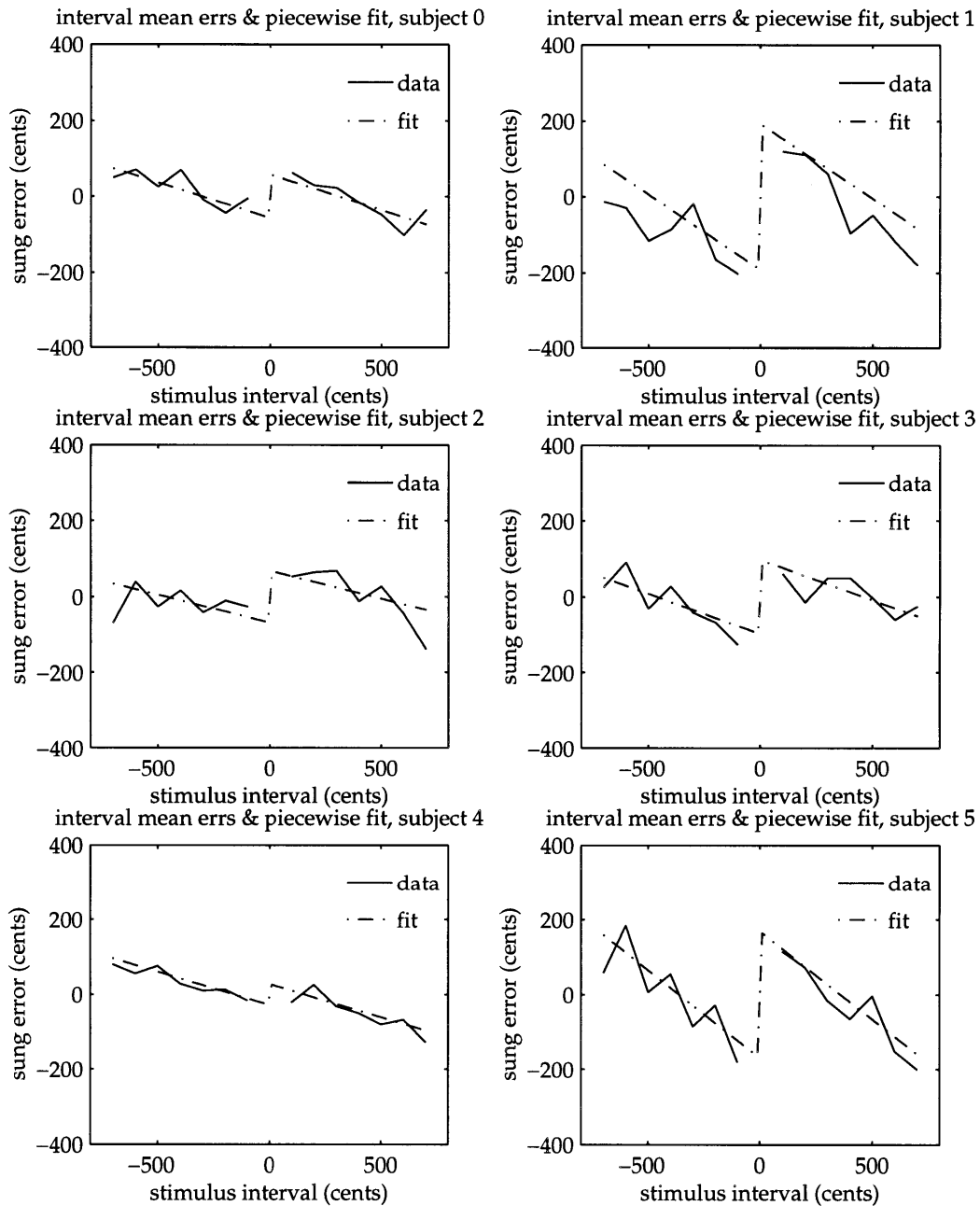


FIGURE 13. Deviation of mean interval size from given interval. An ideal result, where the mean interval size would match the given interval, would be a constant at zero. The consistent negative slope demonstrates some interval “squashing”. The discontinuity at zero signifies a categorization between ascending and descending intervals.

Curiously, the best linear fit to these mean errors is a constant, zero. Also, if one only examines the absolute value of the interval, as in Figure 15, the mean stays constant, with the effects of the non-linearity cancelled. Therefore, for now, we will be satisfied with using the sung interval as a good approximation to the intended interval. We will explore the effects of the piecewise-linear model of error noted here in our melody recognition model described in Section 6.3.

5.3.1 Variances of intervals

Although the means of the errors exhibit interesting behaviors, the interactions of the variances for the intervals are far more revealing.

Displaying information similar to Figure 13, Figure 15 displays the errors by absolute value of the interval. The size of the interval, whether it is up or down, does not affect the standard deviation of the error. That is, there is no significant difference between the standard deviation and a constant for a given subject, with the notable exception of subject 2, due to extreme outliers. The six plots in Figure 15 are summarized in Figure 14.

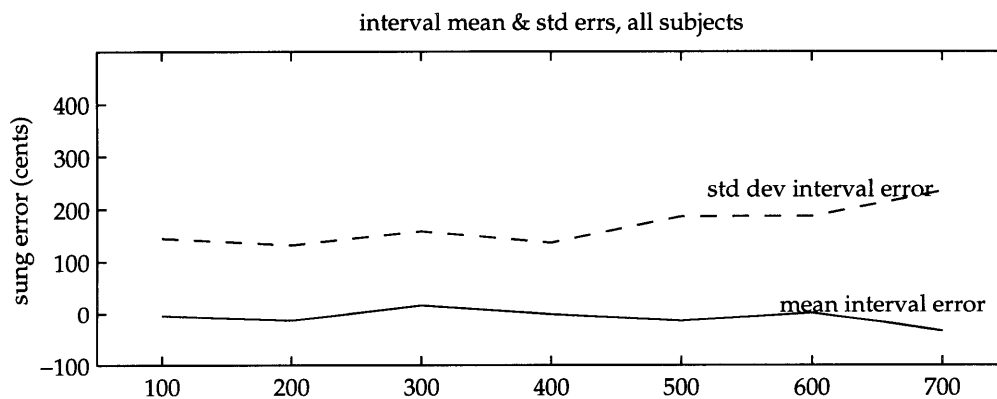


FIGURE 14. Summary of mean and standard deviation of errors across all subjects. See Figure 15 for individual subject data. The mean is zero, and the standard deviation has a slight positive slope.

The fact that the variances stay constant for a given subject leads us to the surprising conclusion that “all errors are created equal.” This result runs counter to our initial expectation that error would increase significantly with interval size: subjects make the

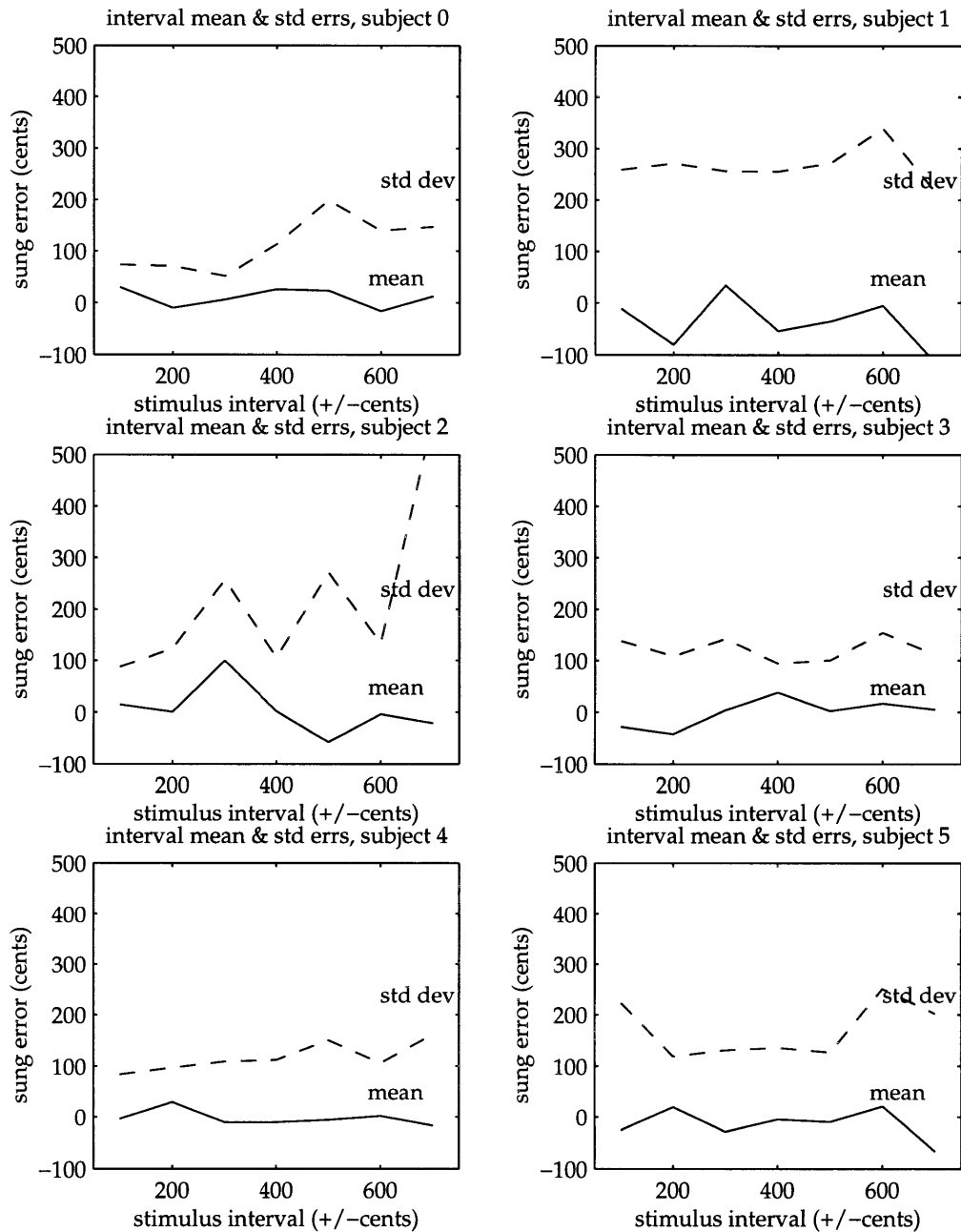


FIGURE 15. Mean and standard deviation of errors by absolute value of stimulus interval. The mean error flattens out to zero when collapsing interval size, in contrast to Figure 13. With the exception of subject 2, the standard deviation for each subject remains essentially constant.

same absolute range of interval errors, rather than a percentage of a given interval size. Therefore, for a single interval, the error size does not depend on the interval size. The mean of the sung interval can be modeled by the equation:

$$\text{Sung} = M \cdot \text{Given} + b \cdot \text{sgn}(\text{Given})$$

But for most of our purposes, we assume that the mean of the given minus sung is zero, as shown in Figure 14.

5.4 Accuracy of accumulated intervals

The interaction of intervals within a stream of notes is at least as important to our model as the accuracy of the individual intervals themselves. One can measure individual intervals, but the result says nothing of how humans put them together to form a musical phrase. This part of the examination of the subjects is critical to our understanding of the experiment.

We began by examining scatterplots of different groups of intervals, for example, comparing the first note to the last note of the phrase, the first note to the highest pitch in the trial, and the second note to the fourth note. Example plots and their correlations for one subject are shown in Figure 16. It soon became clear that the plots were practically identical for any pair of notes; the correlation between the given and sung interval for each of these plots was extremely close. Subjects were just as accurate for the first note to the second note as the first note to the fifth note.

This result was unexpected at first. We had expected errors to accumulate, at least to some extent. Errors would perhaps decrease for some salient reference point in the stream of a melody. For example, the first note and last note were expected to be quite important, as well as the highest and, to a lesser extent, lowest pitch in the group of five notes. These expectations were not supported by the data, however. In fact, if anything,

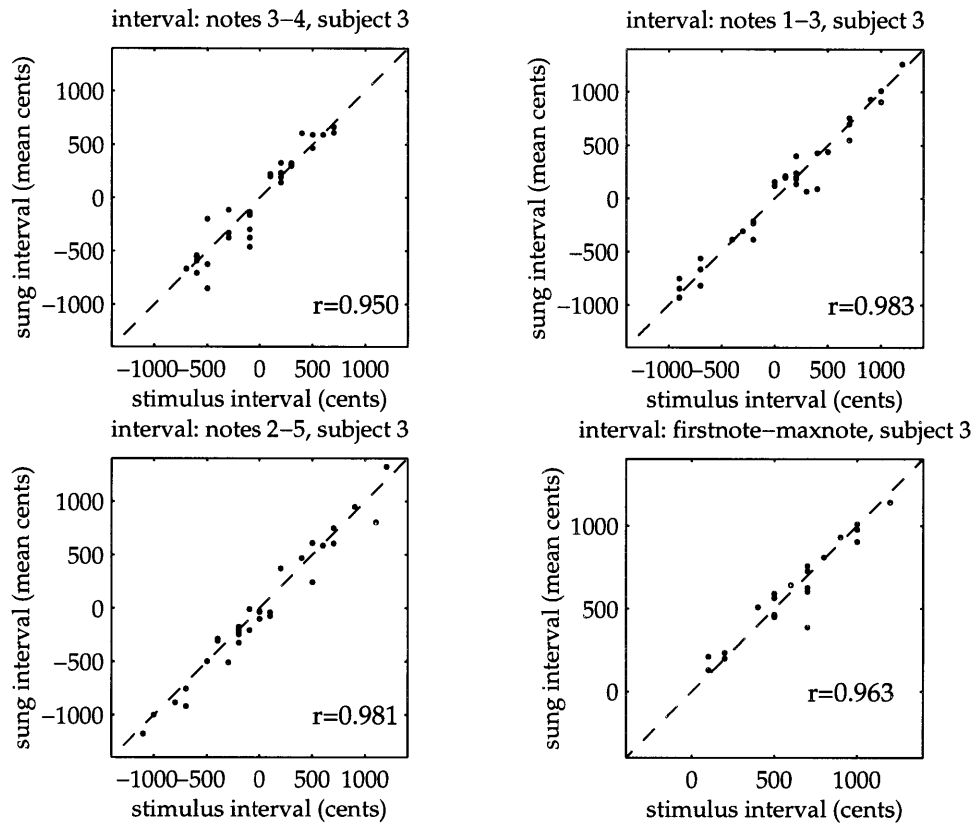


FIGURE 16. Typical relationships of correlations of intervals of various note-distances in the five-note phrases. These correlations are representative of those found for every subject.

in the case of the first note to highest note, the correlation of given interval to sung interval was lower than for the others.

Spurred on by this result, we more methodically examined how pitch errors interacted with distance in time; that is, what happened to the variance as more notes came between the two notes being compared. By looking at the variances in the absolute errors for the first note to second note, the first note to third note, and so on, we concluded that there was no difference between each case. Furthermore, when looking at all note-pairs, with pairs being separated by one, two, three, or four intervening intervals, there was no difference in variance, as seen in Figure 17.

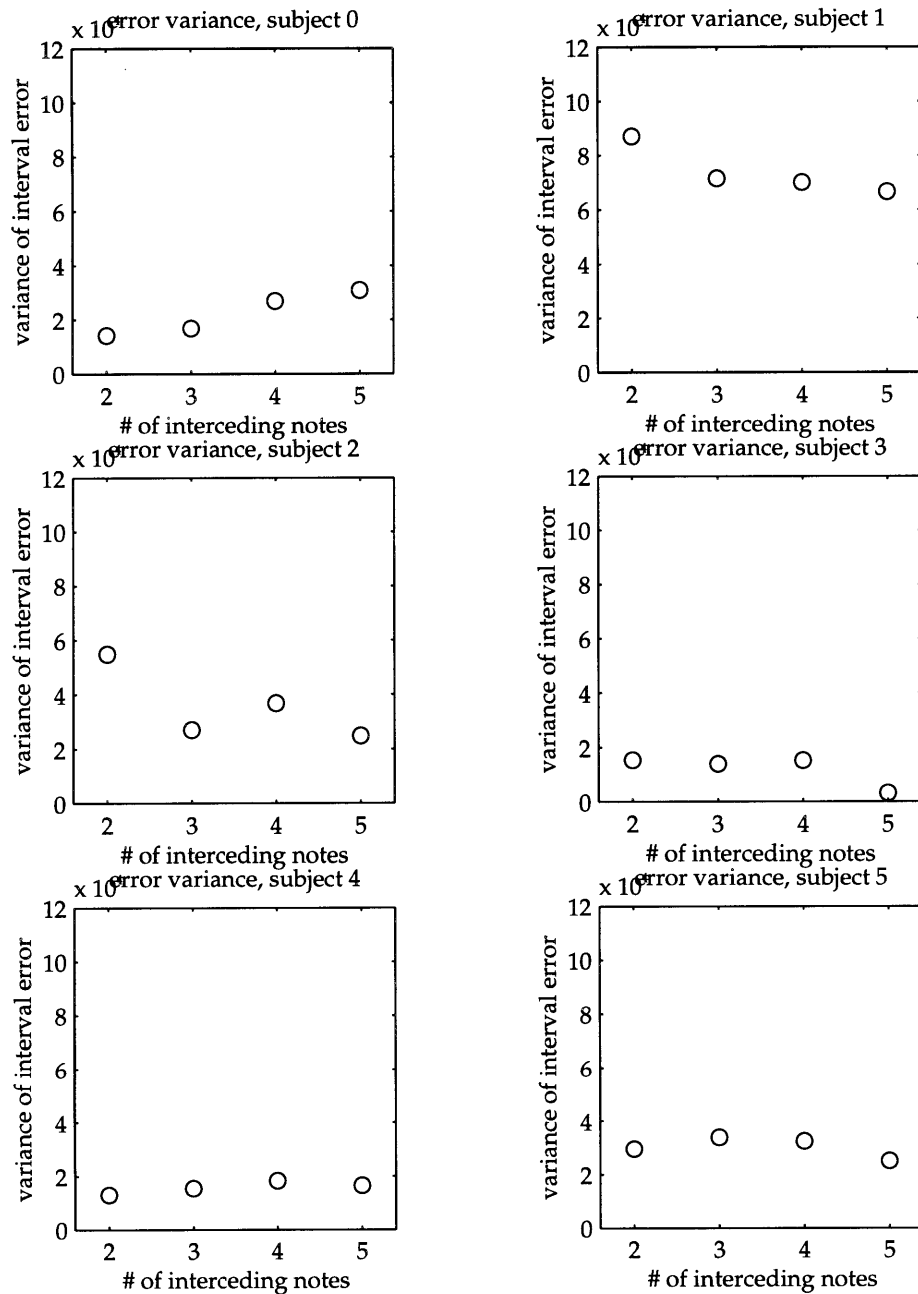


FIGURE 17. Variance of interval error versus number of notes in time. The x-axis reflects the increasing distance between notes in time, with as the 3rd-4th, 1st-3rd, 2nd-5th, and 1st-5th notes as examples of the increasing distance. Only subject 0 increases (doubles) in variance, but since these are variances, it would have to increase linearly for errors to accumulate.

That the variance of the errors should be so consistent was quite a surprise, and its implications are probably the most important results from the experiment. Errors do not accumulate from note to note. Rather, more often than not, subjects correct for their mistakes. If one interval is sharp, then the next is likely to be flat compared to what it would nominally be according to the stimulus. Individual notes are therefore isolated, so the error does not propagate significantly through the rest of the trials.

Another approach that supports this observation is to look at the correlation between the errors of the interval immediately preceding and the error of the interval immediately following a note. Figure 18 shows such a result. Notice that every subject shows a negative correlation, and that the correlations hover about -0.5. This result is exactly predicted by a model of constant variance.

If we take two adjacent intervals X and Y , and examine the interval $(X+Y)$, our constant-variance model predicts that the variances of X , Y , and $(X+Y)$ are equal. So:

$$\text{Var}(X) = \text{Var}(X + Y) = \text{Var}(Y)$$

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \cdot \text{Cov}(X, Y)$$

$$\text{Var}(X) = -2 \cdot \text{Cov}(X, Y)$$

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) + \text{Var}(Y)}} = -\frac{1}{2} \quad (\text{Ross [1987]})$$

This value matches the experimental results nicely. The practical upshot of this result is that each interval can be trusted as well as any other. This result is a great boon for our representation discussed in the next chapter.

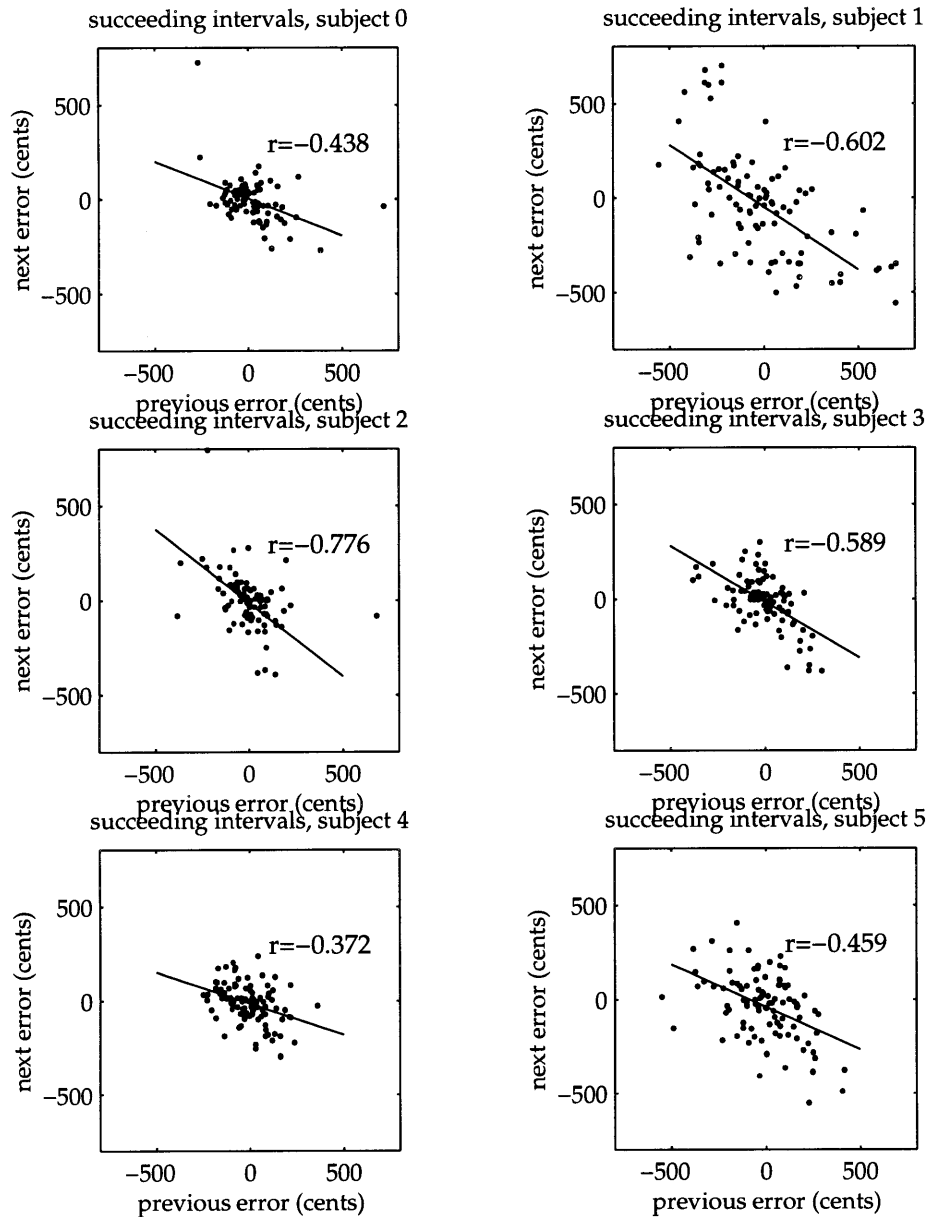


FIGURE 18. Scatterplots for the error of the interval immediately preceding a note versus the interval immediately following. There is a consistent correlation of approximately -0.5.

Representation and Model

Using the results gained from Chapter 5, we may now arrive at a model of melody expression and recognition. The most important fact to be taken from the experiment is that subjects correct for errors while singing. This fact is supported by the evidence that each interval in the series of notes is as accurate as any other. Combinations of intervals do not degrade performance, nor does interval size diminish accuracy. The only regular distortion lies in the mean error depending on interval size, the non-linearity seen in Figure 13 on page 40.

Conveniently, our experimental results suggest a very simple representation. If some of our expectations in chapter one had proven to be true, we would have had to adjust for several factors such as distortions of intervals and different confidences based on note distance. After our experiment, however, we can set forth a simple but effective representation of sung melody.

6.1 Basic representation

Our primary goals in seeking a representation of melody are compactness, expressiveness, and portability. We wish to be compact, so that transformation into our representation does not cause an explosion in data. Such an increase in data would be ineffective for large databases of melodies and would slow computation, especially for a database. We want to retain the expressive qualities of the human input. Simple quantizing would most likely suppress much musical intention, as well as potential expressiveness. Also, this representation is not exclusively tied to melody recognition. If we were to get a novel input such as a melody intended to be added to a database, we would like to be able to keep its expressive qualities in order to determine the underlying musical intention. We want portability in that we do not necessarily want a representation that relies too much on specific knowledge of a given user. The representation must be adaptable to many types of inputs and objects it seeks to match.

The basic representation to be used in our melody model is remarkably simple. We arrive at an interval representation as mentioned in Chapter 2. For n notes to be compared, the representation is a $(n-1)$ -dimensional vector of intervals. These intervals, as alluded to before, are the (signed) differences in pitch between adjacent notes, in order, in the melody. This “transformation” applies to both the human input to the system, and the existing data to be compared. Note that though simple, this representation is richer than the traditional, u/d/s view of contour.

For example, the representation for the stimulus trial shown in Figure 10 on page 33 would be the vector [700 500 200 -200]. The corresponding phrase sung by subject 0 would be [894.1 374.0 225.5 -268.4]. Although the vectors for the two renditions are different, they are close, and more expressive than [+ + + -] from the u/d/s case.

This simple representation meets our above requirements of compactness, expressiveness, and portability. It is justified by our experimental data. The chief information we learned is that for intervals, one interval error is essentially the same as another, and

therefore we cannot order intervals in terms of confidence. Sung intervals do not deviate from their nominal values in a regular way, so those cannot be transformed, either.

6.2 Basic model

To gauge the effectiveness of this model, we shall use it on a “toy” problem similar to a typical application. Much of the underlying motivation for this study is for indexing a database of melodies, attempting the same sort of recognition with a computer as with the cocktail pianist in our introduction. Thus we will use our experimental data in reverse. We already know the mapping of musical intention from the sung phrase to the stimulus phrase, so we can imagine that the stimuli create an index of melodies, and that the sung phrases attempt to index them.

In this toy system, we measure the distance between each sung phrase and each of the possible (thirty-two, in this case) stimulus phrases in our “database.” We use the simple (4-dimensional, here) euclidean distance as our distance metric. We minimize the distance to obtain a best match. This basic melody recognition model performs well, and demonstrates the characteristics of our representation.

We see the results of this first match-model in Figure 19. The surface has peaks where the given and sung phrases are closest. This results in most maxima being along the $x=y$ line.

The crucial data is more visible in Figure 20, which shows the best matches, as in a possible output from a melody database. For all but subject 1, whose singing ability is limited, the matches are quite good, with most trials being correctly identified as the best or second-best match. Some trials, such as sung trial #2, are consistently missed across subjects, and looking at the musical material, it is easy to see why (see the Appendix). Stimulus phrase #2 is extremely difficult to sing, and is strikingly similar to the first stimulus, with which it is most commonly confused. The performance of the matches

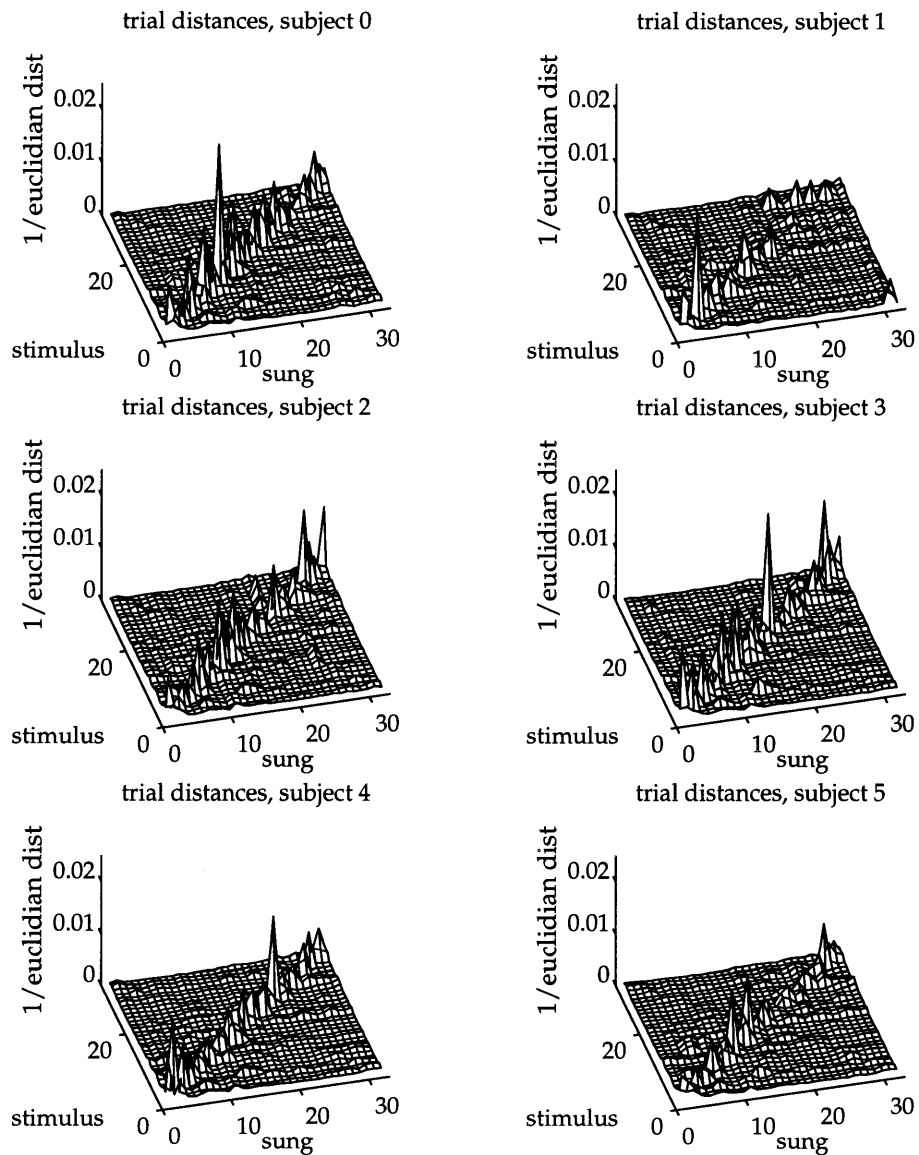


FIGURE 19. Inverse distances from sung trial to stimulus (indexed) trial. Basic model and representation. The sung trials are indexed by the number of the stimulus, so there should be a maximum ridge along the line, $x=y$.

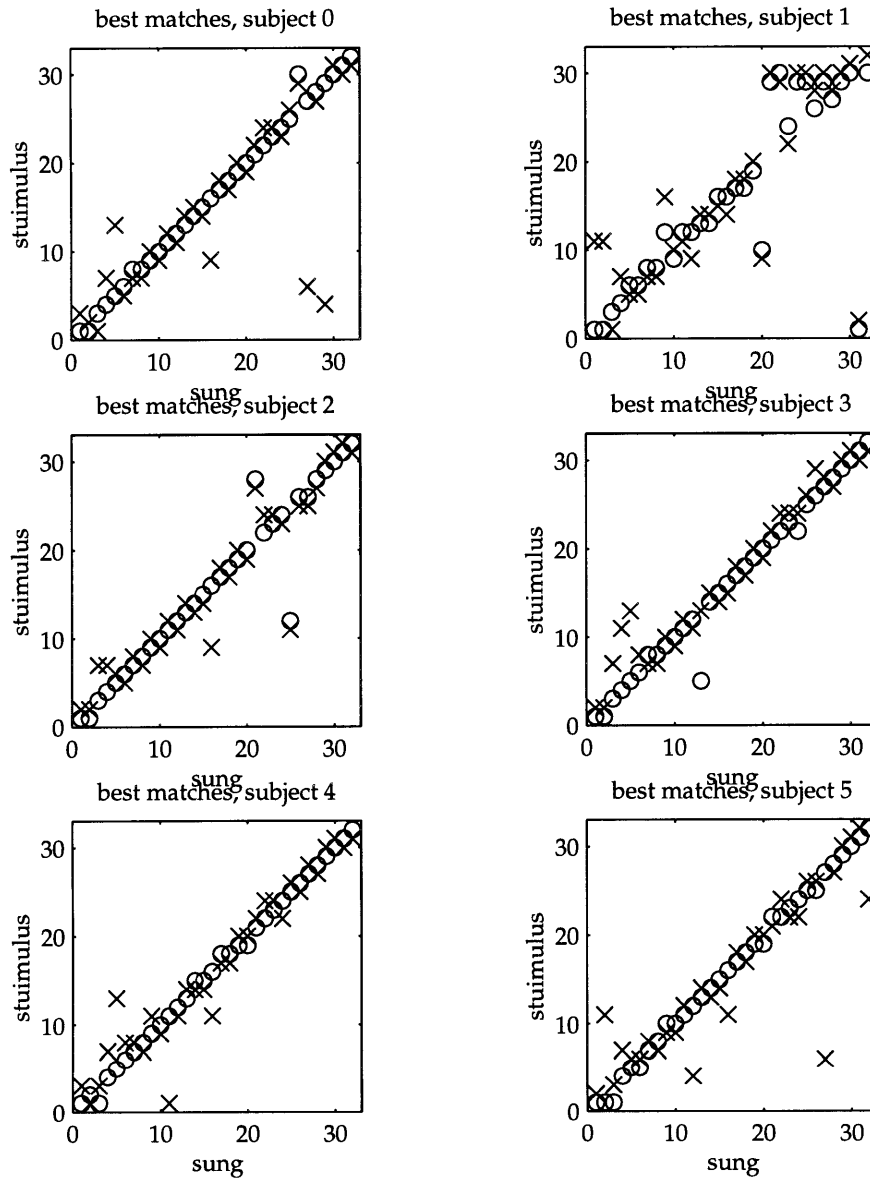


FIGURE 20. Best matches for basic model. For each sung trial, the closest intended (stimulus) trial is marked with a circle. The second closest trial is marked with an 'X'. Performance for most subjects is very good, with most points along the $x=y$ line. Note the clustering of best- and second-best matches in 2×2 cells. This is due to adjacent trials having similar (binary) contours. See the Appendix for details.

for this basic representation and its modified forms are summarized in Table 1 on page 60.

It is worth commenting on the performance of this system compared to others at this point. The Ghias et al. [1995] melody index system reviewed in Chapter 2 was based on the ternary contour definition. Since every stimulus trial in our experiment has a corresponding trial with the same binary contour, we are already, with this simple system, distinguishing between melodies indistinguishable in a sign-based contour system.

In this light, it is not surprising that we see first- and second-choice confusions in 2x2 cells in Figure 20. These cells have identical sign-based contours, and so the indexed melodies are very similar but can still be distinguished. This system is very promising for representing melodies.

6.3 Extensions to the basic representation

Our results suggest a number of basic improvements to our model. The first improvement we will attempt, however, is inspired by basic pattern recognition techniques. All of the improvements to the basic vector representation involve some sort of transformation of the input vector, and possibly the comparison vector as well. The distance and scoring metric of inverse euclidean distance remain the same in our model.

6.3.1 Covariance matrix

Therrien [1989] suggests that one way of optimizing this sort of recognition task is to transform the space in which the recognition takes place. We will optimize this interval space by examining the covariance matrix for all of the sung trial-vectors for a subject, and transforming the recognition space by multiplying by the eigenvectors of the said covariance matrix. This data-driven approach is designed to get as much information out of the available data as possible.

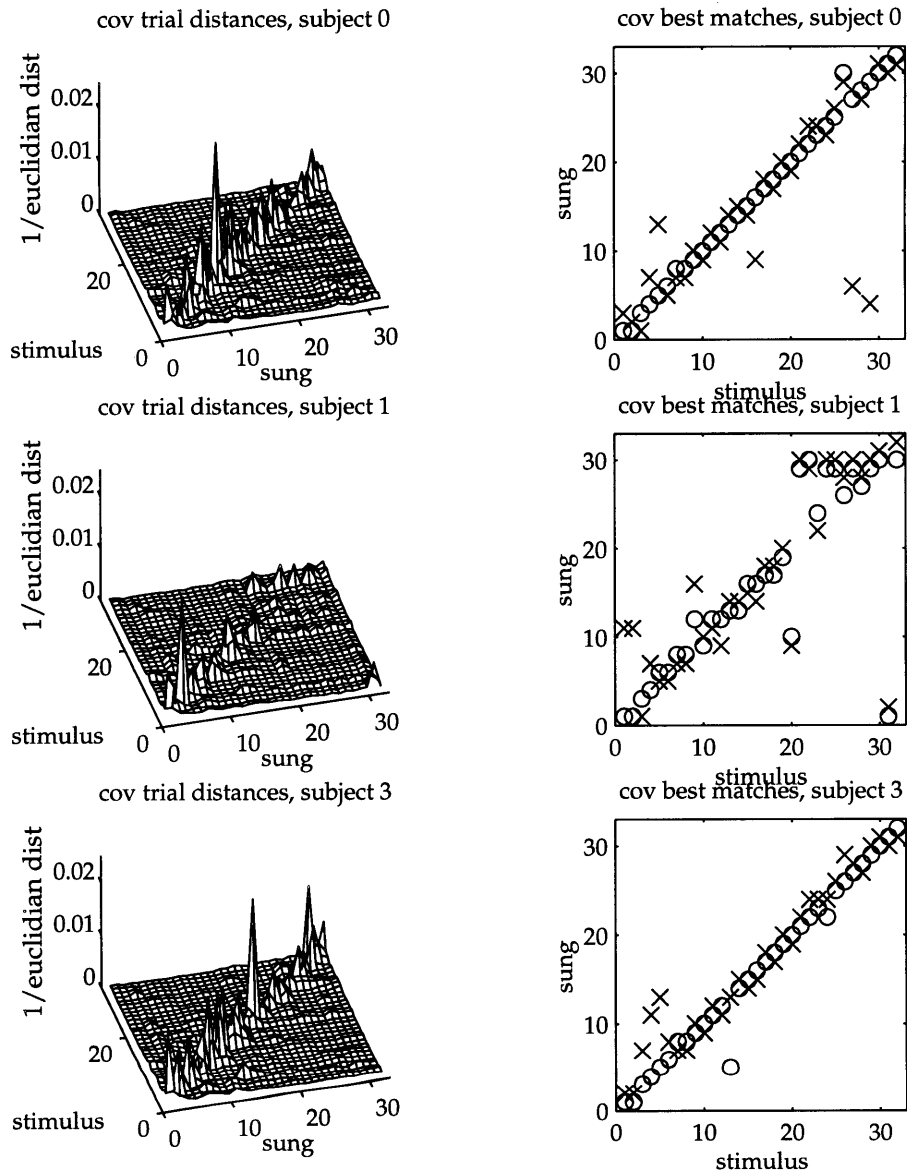


FIGURE 21. Distances and best matches of representative subjects with principal component-based optimization. Best matches are denoted by circles, and second-best by 'X's. Note there is no discernible difference between these plots and those in Figure 19 and Figure 20.

This principal component analysis was not effective on our data set. Although the transformation through the eigenvectors did change the representation space, it did not significantly change the distances, and therefore the ordering was not changed. The first and second best matches are exactly identical to the matches in our original, basic representation. This pattern recognition technique has gained us nothing, which suggests that the basic interval space is close to optimal.

As a side note to our covariance matrix discussion, we briefly examined the possibility of using some variation on an absolute-pitch representation. After zero-meaning the vectors of five patchiest gain key independence) and taking the eigenvectors of the covariance matrix, we noted that one of the five eigenvalues for each of the subjects went to zero. That is, the dimensionality of these five-note vectors was actually of rank 4, because we removed the degree of freedom contained by the means. There was no gain in information by moving to a five-note representation (once it was made useful) from a four-interval representation; our use of intervals was further justified.

6.3.2 Piecewise-linear accommodation to subjects

The next modification to our representation was to return to the observed non-linearity in the means of the errors according to interval size, as illustrated in Figure 13 on page 40. Since we are able to characterize the distortion in interval size fairly accurately for each subject, we should be able to use this information to transform the distorted utterance back to the predicted intention easily. We fit the data into the discontinuous model:

$$\text{Sung} = M \cdot \text{Given} + b \cdot \text{sgn}(\text{Given})$$

and treat the modified sung data as our new pitch vector.

This piecewise-linear technique was slightly more successful than the above pattern recognition technique. For subject 1, who had a very marked effect of interval distortion, this accommodation had a slight performance increase in best matches. The effects

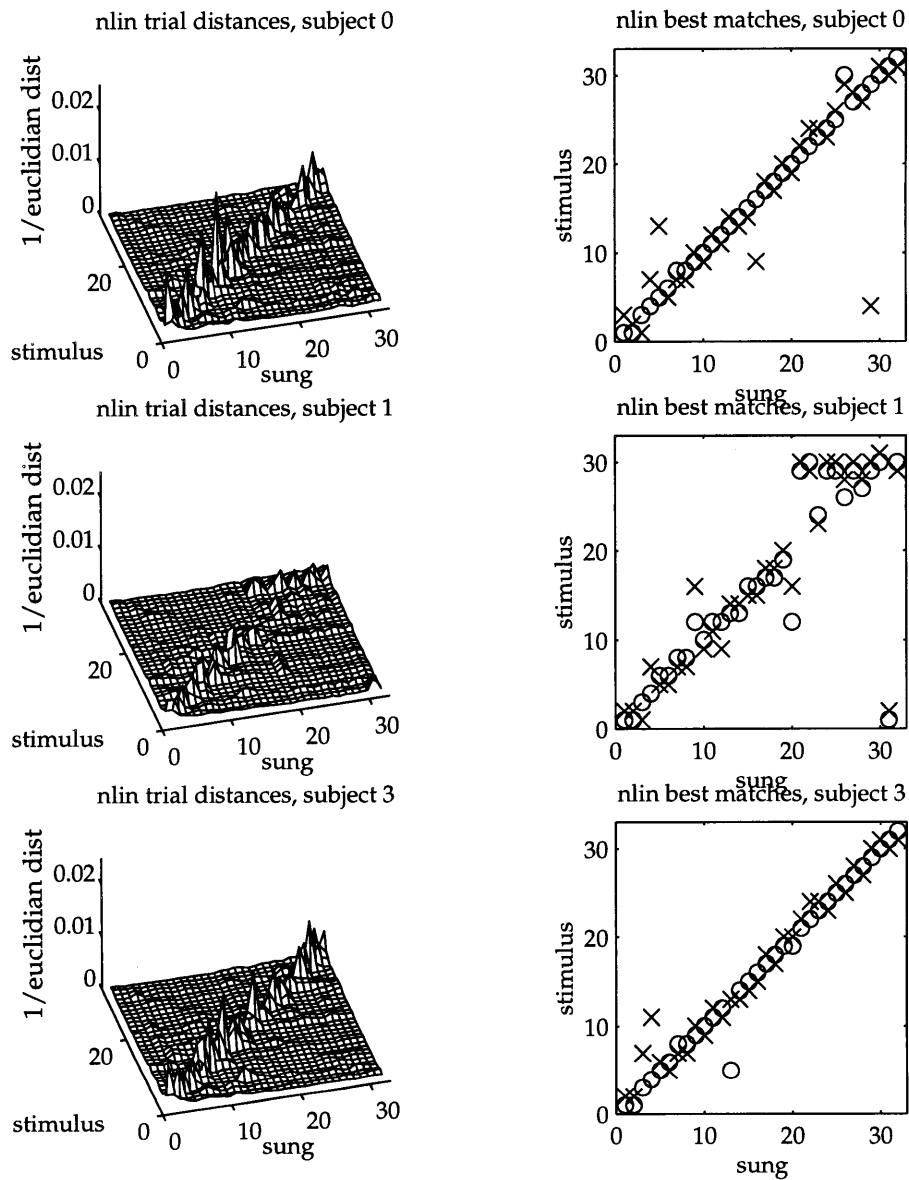


FIGURE 22. Inverse distances and best matches for the non-linear accommodation technique. Subject 1 has slightly improved performance over the basic model.

can be seen in Figure 22. The performance did not change for subjects 0 and 3, although there was clearly an observable change in ordering. This method could be effective in adjusting for particularly untrained singers, but it requires some prior knowledge of the singer's habits.

6.3.3 Ten-dimensional expansion

Our final modification to the basic model came directly from the observation that “all errors are created equal”. If that statement were true, would not adding more data with negatively correlated errors help? Although we could not add any information *per se* to the representation, we could attempt to utilize all the information in the data we did have. From our experiment, we discovered that more distant stimulus intervals (such as the second note to the last note) were as correlated with the sung intervals as more closely adjacent intervals were. If these more distant figures could be worked into the representation, we might be able to see an improvement in matching sung phrases with intended phrases.

In effect, we added all of the possible note-pairs to our adjacent-interval representation by transforming the sung and intended vectors as follows:

$$\begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} y_1 & y_2 & y_3 & y_4 & y_5 & y_6 & y_7 & y_8 & y_9 & y_{10} \end{bmatrix}$$

By comparing the new 10-dimensional sung and intended vectors to each other and taking the euclidean distance in this expanded space, we managed the best improvement in matching ability over the original representation. The matching for subject 0 decreased in accuracy slightly, but the other two we examined for this summary improved noticeably. The degraded performance of matching for subject 0 may be due to his increasing variance as intervals accumulated: this expansion relies on the improvements afforded by constant variance with note distance, and his was the only data not consistent with those results. This expansion to a higher dimensional space was a surprise success, as there is no information added to the system. It does take full advantage of our result that most subjects do not “drift” when singing a string of intervals.

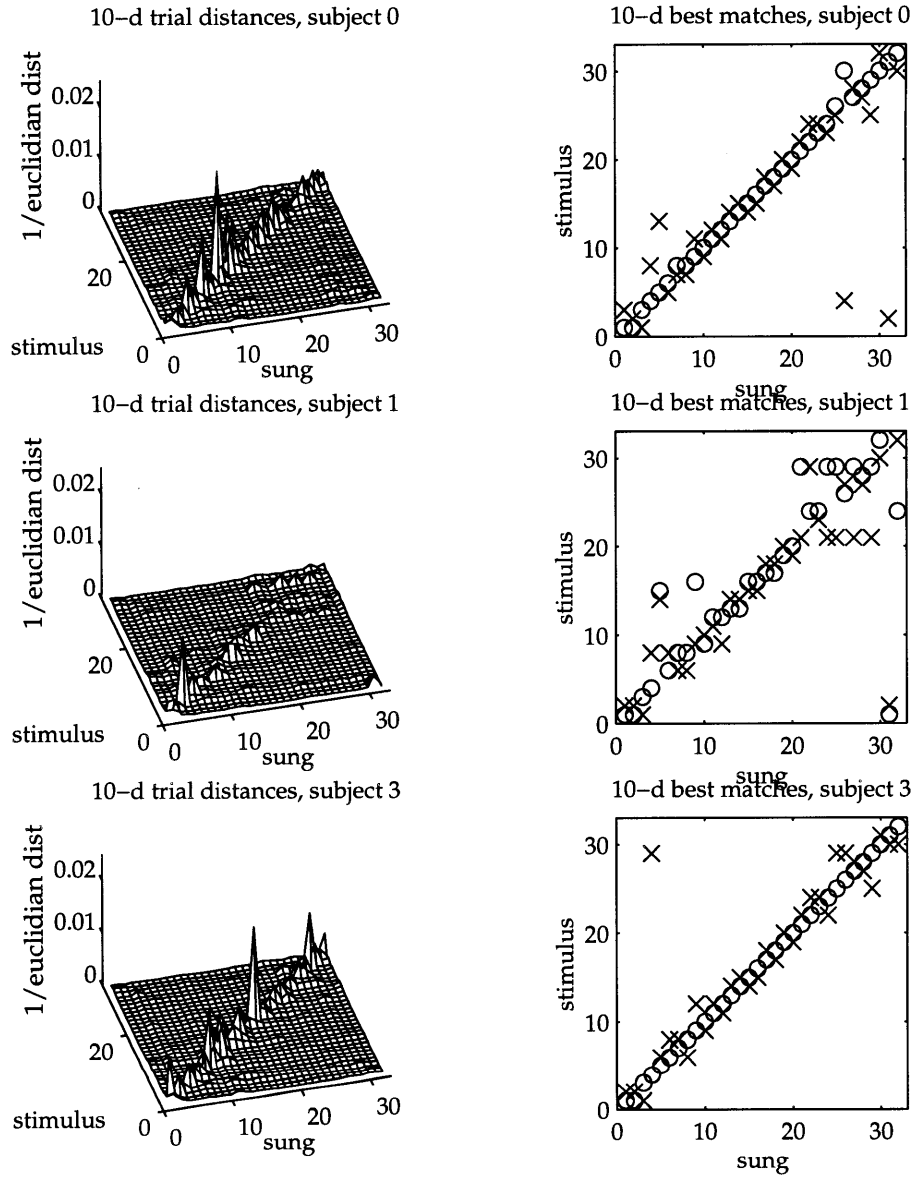


FIGURE 23. Inverse distance and best matches for ten-dimensional expansion of interval data. Subjects 1 and 3 improve over the basic model, while subject 0 declines slightly in accuracy.

6.3.4 Sign-based contour

For the sake of comparison, we ran our data through a sign-based contour system. We encoded the sign of the intervals as -1 and +1, and took the inner product of the resulting sung and index vectors as the score of the match. Results are represented in Figure 24. This representation, as we mentioned above, is limited in performance by the number of stimuli: there are 32 stimuli for 16 possible contours. We do see, as in Figure 25, that performance vastly increased when we allow second-best matches to be counted as “hits”.

6.3.5 Summary

The results for the four different representations and an implementation of u/d/s contour are summarized in Table 1 and Figure 25. The covariance-based optimization does not change performance at all from the basic representation. The piecewise-linear individual subject accommodation makes a slight improvement for the subject with the worst performance. The full all-interval expansion makes more of a performance gain for subjects 1 and 3, at the cost of slight degradation in subject 0.

TABLE 1. Comparison of performance of five models, percent correct.

Sub	Basic Model		Cov Optimization		Accommodation		Full expansion		U/D/S	
	1st	1st/ 2nd	1st	1st/ 2nd	1st	1st/ 2nd	1st	1st/ 2nd	1st	1st/ 2nd
0	90.6	96.9	90.6	96.9	90.6	96.9	87.5	96.9	50.0	96.9
1	40.6	68.8	40.6	68.8	43.8	71.9	43.8	78.1	34.4	71.9
2	87.5	90.6	87.5	90.6	87.5	90.6	87.5	90.6	40.6	90.6
3	87.5	100.0	87.5	100.0	87.5	100.0	96.9	100.0	50.0	100.0
4	87.5	100.0	87.5	100.0	84.4	96.9	87.5	100.0	43.8	90.6
5	78.1	96.9	78.1	96.9	71.9	100.0	81.3	93.8	50.0	100.0

We conclude that our basic interval representation is the safest for most applications. It requires no previous knowledge of the singer, and performs across all situations. If

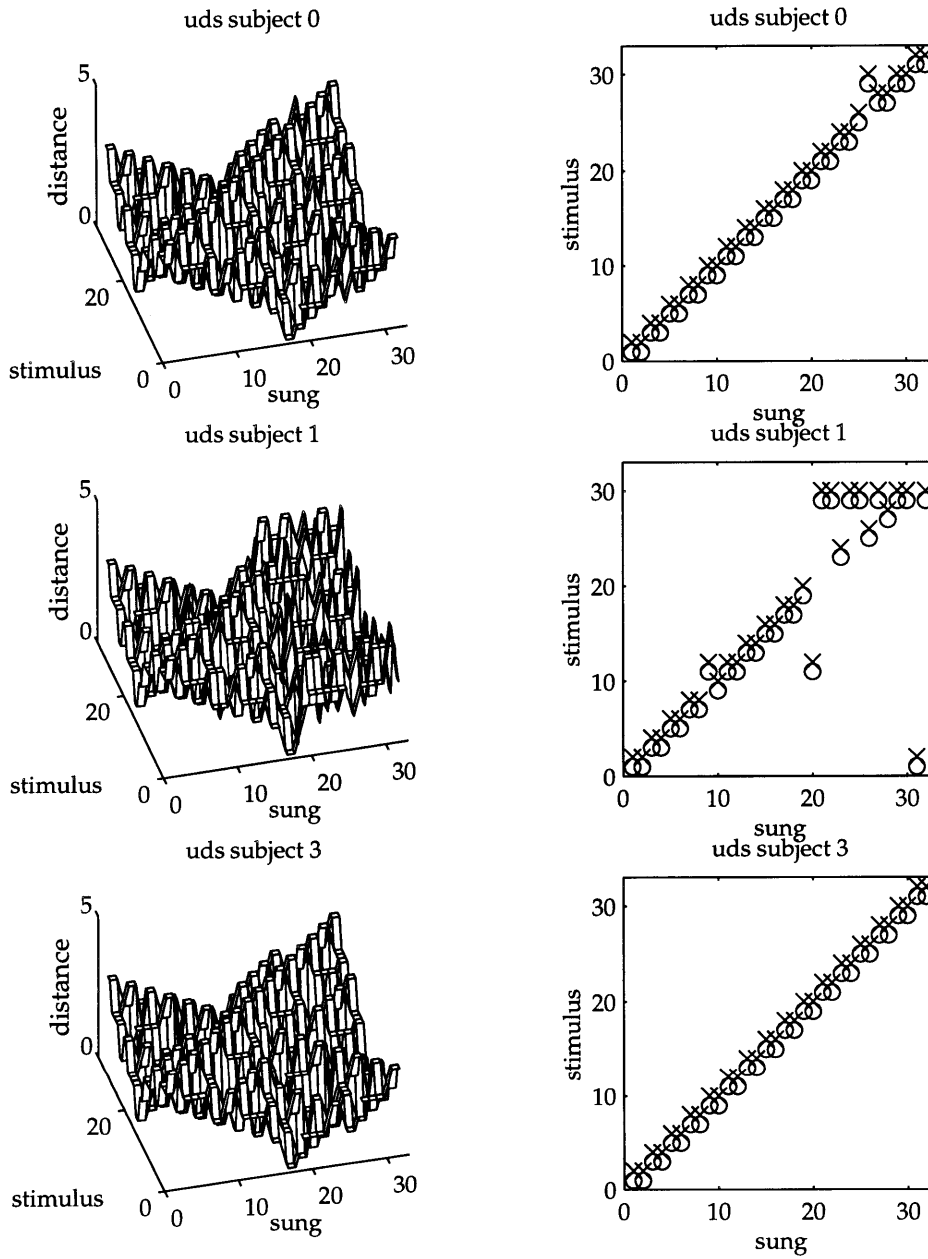


FIGURE 24. Performance of the sign-based contour representation. The distance metric is simply the dot product between the signed contours. Note the confusions of all contours with identical signs.

knowledge about a singer or the ability to train on a subject exist, and the subject is likely to be a poor singer, then accommodation to the subject's particular type of errors could possibly raise the matching ability of the system. The full-interval expansion is also a strong possibility for any type of singer, especially for melody recognition tasks. The only drawback is a less musical interpretation for every element in the vector. A sign-based contour representation is not powerful enough for a case with this few notes. Intervals, even if inaccurate, offer better performance for indexing in the cases we have examined.

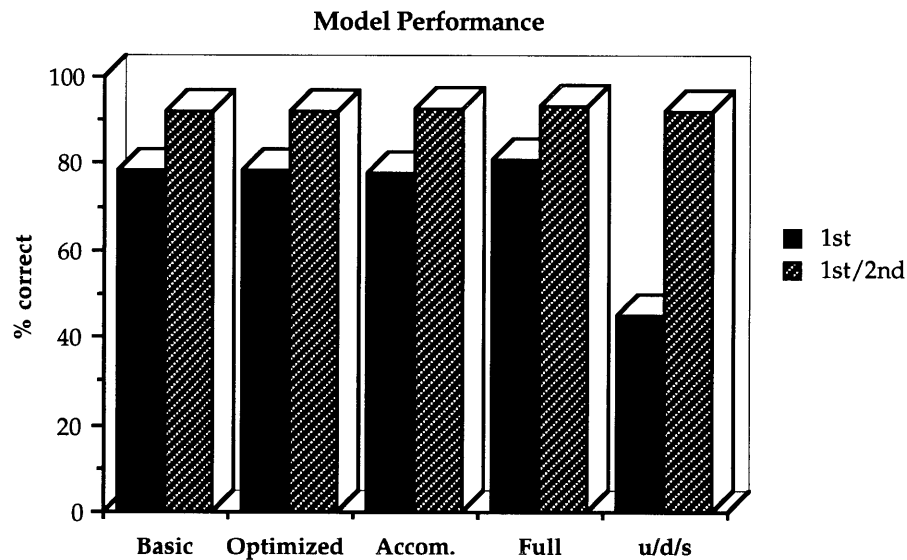


FIGURE 25. Summary of the performances of the five models of melody indexing. The means across all subjects are shown. The modifications to our interval representation have a small effect; the full expansion is the most promising. The performance of u/d/s is limited by the identical contours in our toy database.

In chapters one and two, we examined the motivations behind our study. We imagined a situation in which the study would be relevant, and reviewed both our expectations and the examinations of others into melodic contour. In chapter three, we reviewed the procedure of our experiment. In chapter four, we prepared the data for the statistical analysis of chapter five, which yielded elegant results. Chapter six outlined a potential means of using the experimental results. We will now evaluate our model, detail possible applications, and suggest future directions of research.

7.1 Model benefits and shortcomings

Our simple representation for melody has a number of features to recommend it; likewise, it has a number of limitations we will discuss. Among the benefits of the model are its compactness, ease of computation, motivation from actual human performance,

and adaptability to a variety of situations. Its drawbacks are tied to its simplicity: a number of features of melodies are not taken into account in the model.

7.1.1 Benefits

The simplicity of the model makes the contour representation trivial to compute. Once the input is pitch-tracked and segmented, we obtain pitch estimates for each note, and take the difference between them. Key independence, the fact that a given melody can be sung at any pitch height, is inherent to using intervals in our representation. The compact nature of the representation also eliminates storage as an issue. The general nature of the melody description allows it to be adapted to other situations than melody recognition, as we will discuss later in this chapter.

The simplicity of the model is its chief positive quality, but we should not forget its inherent richness. It allows more information to be conveyed than the previous binary and ternary sign-based models. Such models achieve their robustness through extreme simplicity, but lack the expressiveness of our representation. For example, a sign-based contour scheme would miss the differences in the melodies shown in Figure 6 on page 20, but ours would not. For less-contrived examples, sign-based contour representations typically require at least ten notes as input in order to distinguish between contours. Since we take interval size into account, we can describe far more melodies in far fewer notes.

The fear that pushes sign-based contours to seek robustness through simplicity of input is that average singers cannot accurately reproduce intended intervals well enough to get any information other than “up or down.” We demonstrated otherwise. Subjects make a variety of errors when replicating melodies, but there is sufficient information in their utterances to index melodies more compactly and accurately.

7.1.2 Limitations

Since our melody representation is simple, it ignores several features in a given melody that would be instantly recognized by a human listener. The primary feature of this

type is rhythm. As we mentioned in Chapter 2, contour is tied to a metaphor of motion. Human listeners not only hear “which way” (as accounted for in sign-based notation) and “how far” (as included in our representation), but “how fast,” which is rhythm’s prime contribution to the notion of contour. An interesting future research direction would be to incorporate note durations into a coherent representation of melodic contour.

Our representation, as any representation, also has a bias as to what sorts of melodies are similar. Because the metric for this representation is distance in pitch height, melodies will be judged to be similar if they go up and down by approximately the same distances at the same time. The system is blind to modality or any contextual clues, so tunes which sound different to listeners due to the type of scale used may be very similar in the representation.

System performance is limited by the quality of the pitch tracker and note segmentation. Any similar system has the same constraints. Sign-based contour representations allow for more leeway in terms of pitch, since they only require the sign. Our representation is more susceptible to pitch tracker errors. However, this drawback is not of great concern because using the median for note-level estimates gains stability, and singers are likely to make larger errors than those contributed by the pitch track. Note segmentation is a greater concern, because we rely entirely on note representations. If we require singers to clearly articulate note divisions, accurate segmentation can be achieved.

In its current state, our system has assumed melodies and inputs to be of equal lengths. This situation is seldom true for real-world applications. This limitation can be compensated by setting the n -dimensional vector to be the maximum length expected index. For all shorter indices and inputs, the end of the vector can be padded with zeros. This gives a neutral value that would minimally affect comparisons with longer vectors.

7.2 Applications

The most widely discussed application for our representation is query-by-humming. This emphasis is very understandable, as we are approaching a time when indexing multimedia databases by content is becoming both feasible and desirable. If we can develop a simple notation that summarizes a much richer representation of a song, whether it be a MIDI karaoke arrangement or a digital recording, we will have much more flexibility in manipulating and finding the content.

Another possible application is as an active assistant for a composer. With a proper musical knowledge/constraint system in place, our model allows for novel inputs. The program could be forgiving of singing errors if it embodied enough information about music and the singer to adjust for the errors. Simple knowledge would incorporate scales, and some sort of tonality-based expectation of starting and ending-notes. Such a system would be difficult to implement well, but our representation is well-suited to the possibility. The benefits of such a system, if only as a memory aid, could be tremendous to a composer.

Barry Vercoe's synthetic listener/performer [1984] could be extended with a similar sort of system to make for a more responsive synthetic collaborator. One can imagine an improvising accompanist which not only follows a singer's tempo, but listens and incorporates melodic fragments from the singer's performance in the accompaniment. Such a stunt is only possible if the singer's pitch can be correlated to a quantized score, and deviations from the score are labelled as "features" to be embellished.

7.3 Future research

Many aspects of this model suggest more that can be done in this field of research. Incorporating time into a melody model would be interesting, and potentially quite informative. This enhancement to the model would move us away from relying on pitch events, and towards being time-based. Some flexibility in the time domain must

be retained, just as we retained flexibility in the pitch domain by depending on relative pitch: singers are as likely to vary the tempo of a song as its absolute pitch.

Another direction, suggested by the above applications, is to develop a melody knowledge model that incorporates our representation as input. We believe our current melody model is neutral enough to fit into existing hierarchical melody models, such as that by Feroe and Deutsch [1981], and perhaps Lerdahl and Jackendoff [1984]. Knowledge about melodic structure would make deciphering unique inputs much easier than those accommodated now. We envision a sung interval-distance contour as an approximation to a region in our melody space, which would be pruned by a melody expectation system. Likely candidates would then be matched to the input to make a guess at musical intention. This unconstrained-input transcription has a kinship to other work at the MIT Media Laboratory (Martin [1996]).

7.4 The end

Inspired by the ease with which listeners are able to retrieve musical intention from an impoverished input, we have developed a new representation for use in melody recognition and other systems. This approximate-interval representation was based on experimental results that characterized singers' natural behavior in expressing melodies. We demonstrate this pitch-rich scheme to be superior to sign-only contour when comparing short melodies.

Acknowledgments

This document would not exist without the immeasurable support of my friends and co-workers. The Machine Listening Group, namely Dan Ellis, Bill Gardner, Michael Casey, Keith Martin, Eric Scheirer, and Paris Smaragdis, have all been wonderful teachers, friends, and co-conspirators. Keith, especially, has been a great officemate with good taste in music. Judy Brown has been a tremendous help with her advice on pitch and pitch tracking, providing the pitch tracker used for the data analysis. Masako Yamada saved weeks of work by using the pitch tracker on the data.

My advisors and readers, Whitman Richards, Barry Vercoe, and Peter Child, all shaped this thesis through their positive and critical comments. Thanks especially to Whit, whose faith and encouragement sustained me through this research. Elena Ruehr and Ed Cohen deserve recognition for our initial conversations which illustrated what a slippery topic this is to handle. Very special thanks to Kathryn Vaughn, who started me in music cognition, and who has, as a result, influenced this research from the outset.

The Media Lab staff, most notably Betty Lou McClanahan, Santina Tonelli, Karen Modrak, and Linda Peterson, have all been friendly and helpful for all of the time I've spent working at the Media Lab.

Thanks to all of my subjects who generously contributed their time and voices for this study.

Amie Strong, Anne Dudfield, and Erika Abbas have been the closest of friends offering support, advice, and love freely through this entire process. Thank you.

Thanks to all of my friends who have provided on-line and off-line support, and more importantly, much-needed distractions.

Thanks to Mike Keneally and Jane Siberry, whose unique musics and philosophies sustained me through the writing of this document. Hat.

Thanks Mom. And all of my family. I would not be doing this without your encouragement.

Experimental Stimuli

On the following pages we show the thirty-two different trial stimuli used in the experiment described in Chapter 3 and in the test of the model presented in Chapter 6. The phrases were written to include an equal distribution of the fourteen chromatic intervals between a descending perfect fifth and an ascending perfect fifth (-7 semitones to +7 semitones), excluding an interval of a unison (0 semitones). Since there were four intervals per phrase and 32 phrases, there were approximately nine of each of the fourteen required intervals among all trials. We used two each of the possible ($2^4 =$) 16 binary contours, which are listed at the beginning of each pair of phrases. Figure 26 and Figure 27 follow.

Figure 26 displays eight musical staves, each representing a stimulus phrase. The phrases are labeled with contour patterns above them:

- Phrase 1: + + + +
- Phrase 2: + + + -
- Phrase 3: + + - +
- Phrase 4: + + - -
- Phrase 5: + - + +
- Phrase 6: + - + -
- Phrase 7: + - - +
- Phrase 8: + - - -

Each staff shows a sequence of notes and rests on a five-line staff, with a double bar line indicating a phrase boundary. The notes are primarily quarter notes, with some eighth notes and rests. The contour labels indicate the relative pitch movement (up or down) between consecutive notes.

FIGURE 26. Stimulus phrases 1-16.



FIGURE 27. Stimulus phrases 17-32.

Bibliography

- Brown, J. C.(1992). "Musical fundamental frequency tracking using a pattern recognition method," *Journal of the Acoustic Society of America*, Vol. 92, No. 3, 1394-1402.
- Brown, J. C., Vaughn, K. V. (1996), "Pitch center of frequency modulated musical sounds," accepted by JASA contingent on revisions.
- Cole, H. (1974). *Sounds and Signs: Aspects of Musical Notation*. Oxford University Press.
- Deutsch, D., Feroe, J. (1981). "The internal representation of pitch sequences in tonal music," *Psychological Review*, Vol. 88, No. 6, 503-522.
- Dowling, W. J. (1986). "Context effects on melody recognition: scale-step versus interval representations," *Music Perception* Vol. 3, No. 3, 281-296.
- Edworthy, J. (1985). "Interval and contour in melody processing," *Music Perception*, Vol. 2, No. 3, 375-388.

- Ghias, A., Logan, J., Chamberlain, D., Smith, B. C. (1995). "Query by humming—musical information retrieval in an audio database," ACM Multimedia '95 San Francisco.
<<http://www.cs.cornell.edu/Info/People/ghias/publications/query-by-humming.html>>
- Gjerdingen, R. O. (1994). "Apparent motion in music?" Music Perception Vol. 4, No. 4, 335-370.
- Kageyama, T., Mochizuki, K., Takashima, Y. (1993). "Melody retrieval with humming," ICMC '93 Tokyo proceedings, 349-351.
- Kaufmann, W. (1975), *Tibetan Buddhist Chant*. Indiana University Press.
- Lerdahl, F., Jackendoff, R. (1984). "An overview of hierarchical structure in music," Music Perception Vol. 1, No. 2, 229-252.
- Martin, K. (1996). "A blackboard system for automatic transcription of simple polyphonic music," Perceptual Computing Technical Report #385.
- Ross, S. M. (1987). *Introduction to Probability and Statistics for Engineers and Scientists*. John Wiley and Sons.
- Therrien, C. W. (1989). *Decision Estimation and Classification: An introduction to pattern recognition and related topics*. John Wiley and Sons.
- Vercoe, B. (1984). "The synthetic performer in the context of live performance," ICMC '84 Paris Proceedings, 199-200.

3761-60