Working Paper Series
ISSN 1170-487X

Melody transcription for interactive applications

by: **Rodger J. McNab & Lloyd A. Smith**

Working Paper 96/32
December 1996

Department of Computer Science
The University of Waikato
Private Bag 3105
Hamilton, New Zealand

Melody transcription for interactive applications

Rodger J. McNab and Lloyd A. Smith
{rjmcnab,las}@cs.waikato.ac.nz

Department of Computer Science
University of Waikato,
Private Bag 3105
Hamilton, New Zealand

Abstract

A melody transcription system has been developed to support interactive music applications. The system accepts monophonic voice input ranging from F2 (87 Hz) to G5 (784 Hz) and tracks the frequency, displaying the result in common music notation. Notes are segmented using adaptive thresholds operating on the signal's amplitude; users are required to separate notes using a stop consonant. The frequency resolution of the system is ±4 cents. Frequencies are internally represented by their distance in cents above MIDI note 0 (8.176 Hz); this allows accurate musical pitch labeling when a note is slightly sharp or flat, and supports a simple method of dynamically adapting the system's tuning to the user's singing. The system was evaluated by transcribing 100 recorded melodies—10 tunes, each sung by 5 male and 5 female singers—comprising approximately 5000 notes. The test data was transcribed in 2.8% of recorded time. Transcription error was 11.4%, with incorrect note segmentation accounting for virtually all errors. Error rate was highly dependent on the singer, with one group of four singers having error rates ranging from 3% to 5%; error over the remaining 6 singers ranged from 11% to 23%.

Introduction

Music transcription systems have the potential to be useful in a number of applications—transcribing folk songs, for example, from recorded archives (Askenfelt, 1975), or for providing real time accompaniment for a performer (Vantomme, 1995). Until recently, however, neither the signal processing power nor the sound input

capability necessary to make music transcription generally accessible has been available on low cost computer systems.

Moorer (1977) was the first to describe a complete music transcription system. His system transcribed two-voice input which conformed to a number of restrictions—only melodic instruments, without vibrato, could be used, frequencies were required to stay within the diatonic scale, and no note could be played which was a harmonic of a simultaneously sounding note. These restrictions exclude instruments such as gongs and bells and the human voice. Furthermore, voices were not allowed to cross, nor tempo to vary. Rhythms were represented in terms of a 'fundamental duration' discovered through the use of a histogram. The system was tested using synthesised violin duets.

Piszczalski and Galler (1977, 1979a, 1979b) developed a monophonic transcription system based on spectral analysis using a 32 ms FFT. Frequencies were identified by finding partials and using them in a manner similar to the histogram method described by Schroeder (1968). Notes were segmented based on amplitude, and musical pitch was assigned by averaging the frequencies over the duration of a note to approximate its perceived pitch.

The Visa project (Askenfelt, 1978) produced a system intended to transcribe folk melodies from field recordings. An analog pitch tracker produced a frequency track that was digitally filtered to remove errors. Because folk musicians often do not use equal tempered tuning, the system determined the scale by creating a histogram of all frequencies in the song and allowing a human operator to position the scale's frequency boundaries. The system segmented notes by examining the pitch track, with each note lasting as long as the frequency remained within the note's boundaries. To assign rhythm, the operator estimated the duration of a quarter note and positioned measure boundaries.

In recent years, little has been published regarding music transcription systems as a whole, with work focusing either on frequency identification (Kuhn, 1990; Brown, 1992) or on polyphonic source separation (Chafe *et al.*, 1985; Vercoe and Cumming,

1988; Wang, 1994). This paper describes a music transcription system designed to accept monophonic voice input; the purpose of the system is to support interactive applications. Two applications have been prototyped using the melody transcription front end. One is a sight-singing tutor—a system that displays a test melody then transcribes and evaluates the user's attempt to sing the melody (Smith and McNab, 1996). The other application is a system that uses acoustic input to retrieve melodies from a database of 9500 folk tunes (McNab *et al.,* 1996).

The paper is organised as follows. Section I describes the transcription system, discussing segmentation of notes from the acoustic stream, identification of note frequencies, and assignment musical pitch and rhythm labels. Section II describes an evaluation of the system and discusses results of the evaluation. Section III summarizes and presents conclusions.

I. MELODY TRANSCRIPTION

A. Preliminary Processing

The melody transcription system is implemented on a Power Macintosh 8500/120, and uses the built-in sound I/O of that machine. The input acoustic signal is sampled at 22 kHz and quantized to an eight bit linear scale; the entire signal is recorded before performing further processing. The signal is then passed through a low pass digital filter with a cutoff frequency of 1000 Hz. stopband attenuation of 14 dB and passband ripple of 2 dB. The filter is implemented as a linear phase FIR filter having nine coefficients. The filtered signal is used for all further processing.

B. Note Segmentation

The purpose of note segmentation is to identify each note's onset and offset boundaries within the filtered acoustic signal. In order to allow segmentation on the signal's amplitude, we ask the user to sing using the syllable *da,* thus separating notes by the short drop in amplitude caused by the stop consonant. The representation used by the segmentation procedure is the RMS power of the signal, calculated using overlapped 10 ms time frames, with a new frame starting every 5 ms. In order to

accommodate noise in the signal, as well as differing recording conditions, two adaptive thresholds are used, with a note onset recorded when the power exceeds the higher threshold and a note offset recorded when the power drops below the lower threshold. A segment is ignored if it is not at least one third the duration of the shortest notated note according to the tempo, both of which are set by the user. With a sixteenth being the shortest notated note, and a tempo of 120 beats per minute, for example, any segment shorter than 42 ms is discarded.

The segmentation process is illustrated by Figure 1. Thresholds, shown in the figure by horizontal lines, are based on a second-order RMS power obtained by calculating the RMS of the RMS frame values over the entire buffer. The thresholds were set, through experimentation, at 35% and 55% of the second-order RMS value.

C. Frequency Identification

A reasonable range of frequencies for voice input is defined by the musical staff, ranging from F2 (87 Hz) to G5 (784 Hz), and the system is designed to accept frequencies in that range. While higher and lower frequencies are possible, we are not, at this point, considering applications likely to make use of those frequencies.

The frequency of the signal is tracked using the Gold-Rabiner algorithm (Gold and Rabiner, 1969), a time domain technique that uses both the peakedness and the regularity of the signal to determine frequency. We chose the Gold-Rabiner algorithm because it is well documented and well understood, and it is robust if the structure of the signal is not distorted (Hess, 1983). Furthermore, it is not our intention to perform research or development in frequency identification; if the performance of the pitch tracker is insufficient for a given application, it can be replaced by a more suitable algorithm.

The pitch tracker is implemented as described by Gold and Rabiner (1969), except that, because the algorithm was designed for speech, it was necessary to make two minor changes in order to track a wider range of frequencies. First, it was necessary to modify calculation of the variable *blanking time*—the time following a major peak during which no other peaks are accepted—so that shorter blanking times are calculated

and, thus, the shorter pitch periods of higher frequencies can be tracked. Second, it was necessary to widen the window width used to choose the correct estimate from the competing six parallel frequency estimators.

Because the Gold-Rabiner algorithm is a time domain algorithm operating on a sampled signal, identification of a pitch period's onset and offset can each be up to half a sample period away from its true position in the analog signal—so the estimate of the length of any given pitch period can be up to one sample off. At low frequencies, one sample period is a small fraction of the pitch period length, and the error is negligible. At higher frequencies, however, the error can be considerable. At 1000 Hz, for example, with a sampling rate of 22 kHz, an error of one sample per pitch period amounts to almost 5%, or nearly a semitone. There are several ways to overcome this problem. Hess (1983) suggests upsampling around the peaks to obtain the required accuracy. This results in a great deal of computation at high frequencies and very little at low frequencies. Linear or quadratic interpolation, using samples surrounding the peak, can also increase accuracy (Kuhn, 1990; Brown and Zhang, 1991). We chose the alternative of averaging pitch estimates over fixed length time frames. This solution has several advantages: it is easy to implement, it is fast to compute, it reduces the data rate, and, because the sampling error depends on the length of the frame, it gives a perceptually constant error rate. Our system uses a time frame of 20 ms, thus reducing the error to 0.23%, or ±4 cents, which approximates human frequency resolution above 1000 Hz (below 1000 Hz, human frequency resolution is less acute) (Backus, 1969). Not all frames in the transcription system are 20 ms long, however—averaging stops when a value is encountered that is greater than 10% higher or lower than the running average of the frame. This is to keep large pitch tracking errors, such as octave errors, from influencing the frequency assigned to a frame. When a frame is complete, either by reaching the 20 ms duration mark or by running into a greater than 10% frequency difference, its average frequency is represented as its number of cents above MIDI note 0, or 8.176 Hz. This representation is for convenience in handling frames:

for reasons discussed below, it is also advantageous to represent the frequencies of notes in this way.

Figure 2 shows the frequency track of the notes segmented in Figure 1.

C. Pitch/Rhythm Labeling

Once a note's onset and offset boundaries are known, and the frequencies of the frames making up the note are determined, it is necessary to assign the note a single representative frequency. This is done using a histogram with overlapping bins. Each bin spans the width of a semitone (100 cents), with bins increasing in frequency by 5 cents at a time. Because frames are of varying lengths, each bin represents the number of samples falling within frames determined to be of the encompassed frequencies. Once the highest peak in the histogram has been found, all frames which lie within the winning bin are averaged to produce a single frequency value. Figure 3 shows the histogram corresponding to the fourth note, spanning time 3.0 to 3.9 seconds, in Figures 1 and 2. As can be seen by the frequency track in Figure 2, there are a number of octave errors in this note; the octave errors are also apparent in the histogram, but the frequency has been correctly identified, as 5918 cents above MIDI note 0, by averaging all frames falling between 5865 and 5965 cents. Representing all notes in this way makes it easy to assign musical pitch labels: on the equal tempered scale, semitones fall at intervals of 100 cents, so C4, or middle C, is 6000 cents, while A4, or concert A, is 6900 cents. This scheme accommodates alternate tunings, such as Pythagorean or just, by simply changing the relationship between cents value and musical pitch label; it can also readily represent nonWestern or experimental musical scales.

A further convenience of the relative-cents representation is that it can adapt to the user's own tuning. In some applications, such as a system that allows a search of music databases queried by sung input (McNab *et al.*, 1996), it is appropriate for the system to begin by assuming the user is singing to the equal tempered scale, but then to adjust the scale during transcription. This is easily done by using a constantly changing offset, illustrated by Table I. Here the singer has sung the first five notes of *Mary Had a Little Lamb*. The system begins by assuming the singer uses an equal tempered scale

tuned to A-440, and the offset starts at 0. The first note is closest to E4, and is identified as such, but it is 30 cents flat on the A-440 equal tempered scale, so the offset receives the value 30. The second note, when the offset is added, is closest to D4, but is 10 cents sharp (with the offset added), so 10 is subtracted from the offset. The interval between the fourth and fifth notes is 180 cents, so it would likely be perceived as a whole tone. If fixed tuning were used, this note would be labeled as D#4, 6300 cents above MIDI 0. For applications in which fixed tuning is appropriate, such as singing tuition, the offset is fixed at 0.

The above discussion has focused on assigning pitch labels. Determining intended rhythms from performed note durations is a difficult problem that is receiving a great deal of attention from music researchers (Widmer, 1995). Blostein and Haken (1990) describe a template matching procedure for determining keyboard rhythms from MIDI input. Rosenthal (1992) attacks the same problem using a hierarchical analysis method inspired by the generative model of Lehrdahl and Jackendoff (1983). Sundberg, Friberg and Fryden (1991) and Berndtsson (1996) follow an analysis by synthesis approach, synthesizing musical performances then analyzing them to determine the factors leading to natural and expressive performance. While we hope to be guided by such research in developing more sophisticated methods of assigning rhythms in future versions of our transcription system, the system currently takes the expedient route followed by previous transcription systems, quantizing each note to the nearest allowable rhythm, based on its duration.

Figure 4 shows the transcription resulting from the segmentation of Figure 1 and the frequency track of Figure 2.

II. Evaluation

This section describes an experimental evaluation of the melody transcription system. The experiment was designed to simulate use of the system in transcribing monophonic recordings; this is an important potential application for melody transcription because of the thousands of field recordings of folk songs held in the

Library of Congress and other collections (Goodrum and Dalrymple, 1982). There was one major departure from the transcription-of-field-recordings paradigm—people were asked to record two versions of each song, one using the words, and the other using the syllable *da*. The use of *da* allows the system to segment notes by amplitude; words were recorded to provide data for future development and evaluation of more sophisticated segmentation methods.

A. Method

1. Subjects

Ten people, five male and five female, were recorded, each singing 11 Christmas songs. Christmas songs were chosen on the assumption that they would be well known to the subjects and that there would be little variation in the versions of the tunes sung. All subjects had some experience playing a musical instrument, with only one having no formal training. Two subjects had degrees in music and extensive singing experience, three had a great deal of singing experience in amateur choirs, two had a small amount of singing experience, and the remaining three had little or no singing experience. Two of the subjects had experience with the transcription system.

2. Recording Procedure

Subjects were recorded using a high quality portable analog tape recorder, a Sony Professional Walkman, model WM-D6C. Each subject was recorded separately, at a convenient place and time.

Before recording, subjects were instructed to sing as much of each song as possible, starting at the most natural place, to keep a constant tempo, to restart any song, if necessary, and to hold the microphone as still as possible to minimize noise and to keep the signal strength constant. A recording level was then set while the subject sang a song of his or her choice, using the syllable *da*.

Each song was recorded first using *da*, then using the words. Songs were recorded in the following order: *Jingle Bells*, *Away in a Manger*, *We Wish You a Merry Christmas*, *Silent Night*, *Twelve Days of Christmas*, *O Come All Ye Faithful*, *Hark! The Herald Angels Sing*, *We Three Kings*, *Go Tell It On the Mountain*, *Joy to the*

World, and *Deck the Halls*. For *Twelve Days of Christmas*, subjects were asked to sing only the first verse. Recording sessions lasted between 25 and 60 minutes.

Recordings were transferred to disk via line-in on a Power Macintosh 8500/120. Sound was sampled at 22 kHz and quantized to an eight bit linear scale. Songs that were aborted and subsequently restarted were not transferred, and as little silence as possible was transferred at the beginning and end of each song.

There were a total of 217 recorded songs; one subject did not know the tune or the words of *Go Tell It On the Mountain,* and knew only the tune of *Joy to the World.* The average duration of songs was 26 seconds, with the longest being 60 seconds and the shortest two seconds (one subject sang only the first phrase of *We Three Kings*).

3. Evaluation Procedure

Evaluation was carried out using the songs sung on the syllable *da.* Because *Go Tell It On the Mountain* was not sung by one subject, that song was not used in the evaluation. The remaining ten songs were used, for a total of 100 recorded songs, comprising over 5000 sung notes, with a duration of 45 minutes and 3 seconds.

Performance was evaluated at the note event level, prior to musical pitch and rhythm labeling; in other words, the question to be answered was: did the system correctly identify note boundaries and frequencies, as sung? Each note segmented by the system was inspected using a special purpose program based on the melody transcription module. The program allowed the operator to visually inspect segmentation points marked on graphs of amplitude or frequency, to manually reposition segmentation points, and to play synthesized segments and segments from the sampled file.

Segmentation errors fall into several categories: deletions, insertions, concatenations and truncations. Errors falling into each of these categories were tabulated, as well as correctly segmented notes. Only segments long enough to be accepted as notes were considered (a sung note truncated so severely it could not be accepted was a deletion). Depending on the tempo chosen by the singer, this could be as short as 40–50 ms.

A single sung note separated by the system into two notes was tabulated as two errors—a truncation and an insertion.

Two sung notes joined into one were tabulated as one correctly identified note and one concatenation; if more than two notes were involved, the first was counted correct and the rest were counted as concatenations.

Frequency identification errors were tabulated in three categories: octave above the correct frequency, octave below, and other incorrect frequency identifications.

The speed performance of transcription was also evaluated, using a dedicated Power Macinstosh 8500 with a clock speed of 120 MHz. Timing was carried out using the system's internal clock, which has a resolution of 17 ms.

B. Results

Table II summarises the test results, showing error rates for each error category, as well as the overall score. In calculating error percentages, segmentation categories were divided by 5251, the total number of sung notes, while frequency categories used 4838, the total number of segmented notes. Virtually all the errors arise from incorrect segmentation of the acoustic signal. Of the 376 concatenation errors, 294—almost half the total number of errors—are concatenations of notes shorter than quarter notes. There were only four frequency identification errors, and all four were octave errors; three times an octave below the correct frequency was identified, and once an octave above the correct frequency was chosen.

Table III shows the error rate for each song, ranging from almost 8% for *Deck the Halls* to 14% for *Away in a Manger.*

Table IV shows the error rate for each subject. The subjects fall into two clearly defined groups, with subjects 1, 2, 3, 5, 8, and 10 having error rates ranging from 11% to 23%, and subjects 4, 6, 7, and 9 having error rates of 3% to 5%.

The highest frequency sung in the recordings was 768.3 Hz (G5, 35 cents flat), and the lowest was 85.8 Hz (F2, 30 cents flat); both were correctly identified by the transcription system.

lower percentage of the note's duration, or by setting the shortest notated rest to a value longer than a sixteenth (shortest notated note and shortest notated rest are separate user options). The subject reported that the system was useable and seemed to enjoy the experience. It is likely that other people in the high error group would be able to learn to use the system, although similar compromise concerning the singing syllable may be necessary, as well as coaching from an experienced user.

III. Conclusion

This paper describes a system that accepts monophonic voice input and transcribes it into common music notation. The system is designed to support interactive applications; it requires less than 3% of recorded time to transcribe acoustic input on a Power Macintosh 8500 with 120 MHz clock. Even on a system with a slower clock, this should be fast enough to support most applications.

The system could be improved in several ways. Real time performance is possible, with segmentation based on a short term or running average of the signal's power. Such operation would be necessary to support some applications, such as automatic accompaniment (Vantomme, 1995). The current method of operation is suitable, however, for many applications, such as the two which have been prototyped, a sight-singing tutor (Smith and McNab, 1996) and a music retrieval system (McNab *et al.*, 1996).

More important is to improve the system's note segmentation. It may be possible to improve the current segmentation procedure by modifying the representation—by using the first derivative of the signal's RMS amplitude, for example. It is preferable, however, to develop a segmentation method that allows the user to sing lyrics, solfege syllables, or other syllables, such as *la*. We have done preliminary experiments with a segmentation procedure based solely on frequency, but this method is not yet as reliable as segmentation based on amplitude. The current system achieves its frequency identification accuracy through the histogram voting procedure over previously segmented notes; in order to increase the reliability of segmentation based on

frequency, it may be necessary to replace the Gold-Rabiner pitch tracking algorithm with one that is more accurate at the individual frame level.

ACKNOWLEDGMENTS

The work reported here was supported by a University of Waikato Research Grant.

REFERENCES

Askenfelt, A. (1978) "Automatic notation of played music: the Visa project," Proc. International Association of Music Librarians Conference, Lisbon 109–121.

Backus, J. (1969) *The Acoustical Foundations of Music*, John Murray, London.

Berndtsson, G. (1996) "The KTH rule system for singing synthesis," Computer Music Journal **20**, 76–91.

Blostein, D. and Haken, L. (1990) "Template matching for rhythmic analysis of music keyboard input," Proc. 10th International Conference on Pattern Recognition, Atlantic City, NJ.

Brown, J. C. (1992) "Musical fundamental frequency tracking using a pattern recognition method," J. Acoust. Soc. Am. **92**, 1394–1402.

Brown, J. C. and Zhang, B. (1991) "Musical frequency tracking using the methods of conventional and 'narrowed' autocorrelation," J. Acoust. Soc. Am. **89**, 2346–2354.

Chafe, C., Jaffe, D., Kashima, K., Mont-Reynaud, B. and Smith, J. (1985) "Techniques for note identification in polyphonic music," Proc. International Computer Music Conference, 399–405.

Gold, B. and Rabiner, L. (1969) "Parallel processing techniques for estimating pitch periods of speech in the time domain," J. Acoust. Soc. Am. **46**, 442–448.

Goodrum, C. A. and Dalrymple, H. W. (1982) *Guide to the Library of Congress* (Library of Congress, Washington, D. C.).

Hess, W. (1983) *Pitch Determination of Speech Signals* (Springer-Verlag, New York).

Kuhn, W. B. (1990) "A real-time pitch recognition algorithm for music applications," *Computer Music Journal* **14**(3), 60–71.

Lehrdahl, F. and Jackendoff, R. (1983) *A Generative Theory of Tonal Music* (MIT Press, Cambridge, Massachusetts).

McNab, R. J., Smith, L. A., Witten, I. H., Henderson, C. L. and Cunningham, S. J. (1996) "Towards the digital music library: tune retrieval from acoustic input," Proc. ACM Digital Libraries 96, Bethesda, Maryland 11–18.

Moorer, J. A. (1977) "On the transcription of musical sound by computer," Computer Music Journal **1**(4), 32–38.

Piszczalski, M. and Galler, B. A. (1977) "Automatic music transcription," Computer Music Journal **1**(4), 24–31.

Piszczalski, M. and Galler, B. A. (1979a) "Computer analysis and transcription of performed music: a project report," Computers and the Humanities **13**, 195–206.

Piszczalski, M. and Galler, B. A. (1979b) "Predicting musical pitch from component frequency ratios," J. Acoust. Soc. Am. **66**, 719–720.

Rosenthal, D. (1992) "Emulation of human rhythm perception," Computer Music Journal **16**(1), 64–76.

Smith, L.A. and McNab, R.J. (1996) "A program to teach sight-singing," Proc. Technological Directions in Music Education, San Antonio, TX, 43–47.

Schroeder, M. R. (1968) "Period histogram and product spectrum: new methods for fundamental-frequency measurement," J. Acoust. Soc. Am. **43**, 829–834.

Sundberg, J. Friberg, A. and Fryden. L. (1991) "Common secrets of musicians and listeners: an analysis-by-synthesis study of musical performance," in *Representing Musical Structure*, ed. P. Howell, R. West and I. Cross (Academic Press, London), pp. 161–197.

Vantomme, J. D. (1995) "Score following by temporal pattern," Computer Music Journal **19**(3), 50–59.

Vercoe, B. and Cumming, D. (1988) "Connection machine tracking of polyphonic audio," Proc. International Computer Music Conference, 211–216.

Wang, A. L. (1994) "Instantaneous and frequency-warped signal processing techniques for auditory source separation," Ph.D. Thesis, Stanford University.

Widmer, G. (1995) "Modeling the rational basis of musical expression," Computer Music Journal 19(2), 76–96.

Cents relative to MIDI #0	Notated Value	Offset
–	–	0
6370	6400 (E4)	30
6180	6200 (D4)	20
5995	6000 (C4)	5
6160	6200 (D4)	40
6340	6400 (E4)	60

Table I. Determining musical pitch with a changing offset.

Error Category	Number	% Error
Deleted Notes	76	1.45
Inserted Notes	39	0.74
Concatenated Notes	376	7.16
Truncated Notes	102	1.94
Octave High	1	0.02
Octave Low	3	0.06
Incorrect Frequency	0	0.00
Total	597	11.37%

Table II. Transcription accuracy.

Song Title	Avg. No. Errors	Avg. No. Notes	% Error
Deck the Halls	48	615	7.80%
Hark! The herald Angels Sing	58	662	8.76%
We Three Kings	49	507	9.66%
Silent Night	51	450	11.33%
Twelve Days of Christmas	30	257	11.67%
O Come All Ye Faithful	71	569	12.48%
Jingle Bells	84	649	12.94%
We Wish You a Merry Christmas	62	479	12.94%
Joy to the World	73	560	13.04%
Away in a Manger	71	503	14.12%

Table III. Average error for each song.

Subject	No. Errors	No. Notes	% Error
1	116	598	19.40%
2	126	542	23.25%
3	63	568	11.09%
4	27	591	4.57%
5	60	489	12.27%
6	14	542	2.58%
7	16	542	2.95%
8	74	420	17.62%
9	30	583	5.15%
10	71	376	18.88%

Table IV. Error for each subject.

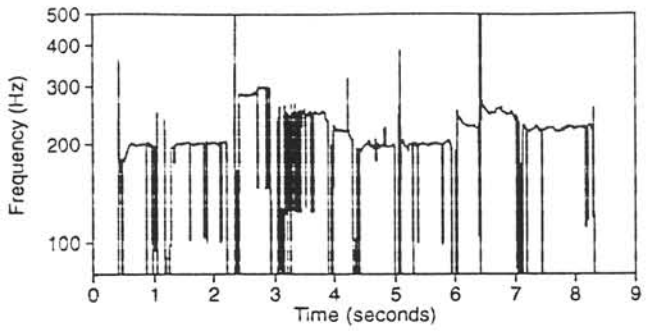

Figure 2. Frequency track of notes segmented in Figure 2.

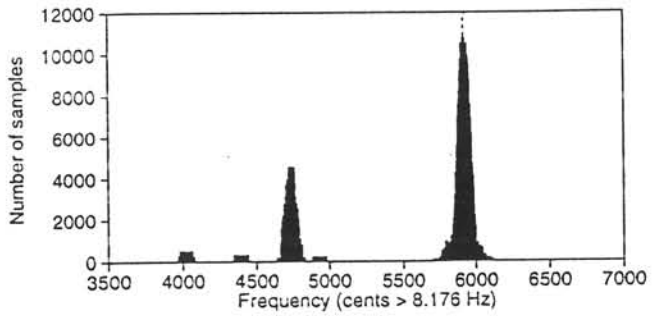

Figure 3. Using a histogram to determine frequency.

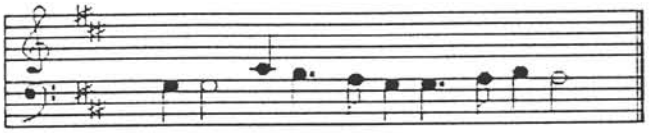

Figure 4. Transcribed notes.