Proc. 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York, Oct. 17-20, 1999

VOCAL INTERFACES TO MUSICAL MATERIAL

Mark Kahrs[†]

Multimedia Group
CAIP Center
P.O.Box 1390
Piscataway, NJ 08855-1390 USA
kahrs@caip.rutgers.edu

ABSTRACT

As the World Wide Web and the Internet becomes the dominant form of information distribution, consideration must be given to the indexing of musical material including themes, melodies, rhythm tracks and so forth. This paper describes the implementation of an algorithm for locating song titles from the vocal input of amateur singers. The prototype algorithm exceeds 90% accuracy for 9 different singers.

In memory of John Choi

1. INTRODUCTION

As the World Wide Web (WWW, the Web) becomes the dominant form of information distribution, consideration must be given to the access to musical material indexed by themes, melodies, rhythm tracks and so forth. The development of Digital Libraries also presents new opportunities for searching and retrieval via acoustic input.

This paper describes the experimental results in the design and implementation of an algorithm for locating song titles from the vocal input of amateur singers. As expected, amateurs lack precise control of pitch and timing, therefore any successful search algorithm must accommodate inaccuracies in both dimensions. An experimental system is described that matches an acoustic vocal input against a library of monophonic musical scores and returns a list of matched results.

Past research into this topic has taken one of two paths: either to fully analyze the input for musical content [1, 2] or to use melodic contours (McNab, et al.[3, 4, 5]). In the first case, the complexity of the analysis procedure is an overwhelming obstacle; finding key signatures in amateur singers is *not* easy (or even plausible in some cases). In the second case, contours can formed by up and down intervals and then matched against the database. Any matching procedure must be forgiving, since human performance is variable in pitch, timing and timbre. Using coursely quantized intervals (up and down were used by McNab, et al.) is one such method. We elected to use a different method of contour calculation and in turn, expect more from the matching procedure. The block diagram of the system is shown below in Figure 1.

The input is recorded and then fed to a pitch detector. The input is also given to a frequency domain based note on and off analyzer to derive a pitch mean. The note is then compared against a reference target and the result is reported to the user. Each input must be compared against a subset of the reference pitches for all the melodies in the library. Finally, a list of matches is presented to the user.

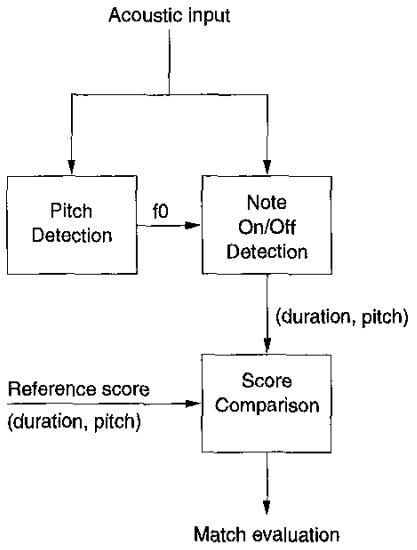

Figure 1: System block diagram

In the next five sections, the insides of each of these blocks will be detailed. To begin, the acoustic properties of singing will be briefly reviewed. Next, the details of the signal processing algorithms will be given. Searching the database will be examined next and then the results of the experimental trials will be presented. Finally, the conclusion presents the failures and successes of the approach shown in Figure 1.

2. ACOUSTICS OF SINGING

Many researchers have examined the production of the singing voice from both the acoustic and physiological point of view. There are many aspects to the production of singing including breathing, the vocal source and complex articulation. An overview can be found in the book by Sundberg[6].

Vocal pitch is different for singing (as opposed to speech): the

[†]Work performed as a consultant to AT&T Labs – Research, Shannon Laboratory, 180 Park Avenue, Florham Park, NJ 07932-0971

singer is expected to maintain a constant pitch over the duration of the note as well as effect the transition to a new pitch within the small temporal boundary of the inter-note spacing.

Formant structure is different in speaking and singing – perhaps most notable is the existence of the "singer's formant"[7], which is introduced by the lowering of the singer's larynx. This results in a widening of the third formant and an increase of output around 2.5 to 3 kHz and manifests in a timbral change.

3. SIGNAL PROCESSING

Signal processing is performed using the digitized acoustic input, with the purpose of deriving a "piano roll" type score (i.e., pitches and durations) that is compared with the reference score. Pitch detection is the first phase of analysis, followed by a segmentation (note on/off time) detection algorithm.

3.1. Pitch detection

As shown in Figure 1, the input waveform is first analyzed for a pitch track. There are many pitch algorithms to choose from (see the encyclopedic book by Hess[8]) but Talkin's algorithm[9] is fairly robust. Pitch detection is often problematic around the transition between vowels and consonants (in particular, between voiced and unvoiced segments) and this will cause problems in the pitch matching due to the introduction of extra notes in the singer derived score.

3.2. Segmentation

Detection of note on and off times (so-called "onset" and "offset" times) can be done in either the frequency or time domain. In the frequency domain, the pitch track is examined for sudden changes of frequency. In the time domain, the amplitude envelope is examined for silences or other changes of amplitude. As might be expected, singer performance will effect the time domain waveform. In particular, when notes are performed with *legato* or *portamento*, the amplitude track exhibits little to go on. On the other hand, the derivative of the pitch track (or any other frequency waveform) can be analyzed for positive or negative slopes.

Experimentally, it was found that the third formant was the cleanest to derive this time domain behavior. The input waveform was filtered using a 16 band ERB (Equivalent Rectangular Bandwidth) filter[10, 11] and then converted to decibels. The highest band was convolved with a 125 millisecond Hann window to smooth the input in preparation for taking the derivative. This is shown in Figure 2. Next, the two point discrete time derivative was calculated. This is thresholded against the mean of the derivative to find the first note. Likewise, going backward from the end of the recording, one can find the last note. Between these two time points, one can find the on and off peaks. These form the on and off points. Next, starting from the earliest on peak, one advances to the next stop peak discarding intervening start peaks. This can happen if there is a step in the output of the third formant filter.

At this stage, the on and off times are used to calculate the mean of the pitch from the input pitch track. This is a requirement since sustaining a constant pitch over vowels is not possible with most amateur singers.

Finally, the experimental "piano score" of pitch and time durations is ready to be compared with the reference score of the tune in question.

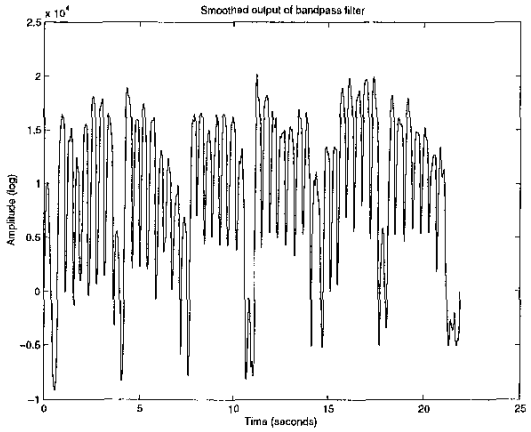

Figure 2: Bandpass filter output

4. COMPARING THE TUNES

In order to accommodate variable timing of the sung input as well as inaccurate pitches from the singer, Dynamic Time Warping (DTW) is used to match the notes. It has been used with great success in Automatic Speech Recognition[12] where it is typically used to compare cepstral coefficients. In Dynamic Time Warping, the time axis is distorted to minimize the error resulting from the comparison of the two sequences. Sakoe and Chiba[13] describe how to optimize DTW explicitly for the speech recognition task taking into account slope constraints over the time warp. Dynamic Time Warping depends critically on a distance function D given the two time series vectors X and Y. The classical view of the time normalized distance vector is given in Equation 1.

$$D(X,Y) = \frac{\sum_{k=1}^{K} d(c(k)) \cdot w(k)}{\sum_{k=1}^{K} w(k)} \quad (1)$$

In equation 1, $w(k)$ is the weighting function and $d(c(k))$ is the distance function. When $w(k) = 1$, then the weighting function is uniform over all time.

Dynamic Time Warping is a dynamic programming method that critically depends on a valid metric to match a test sequences against a reference sequence.

The first step is to put the two sequences on the same time scale. Accordingly, the reference score is truncated to the same number of notes as the test input (this assumes, of course, that the reference is longer than the test input). Second, the time scale is made the same for the two scores by rescaling the durations of the notes. This is required because DTW doesn't work when the time scale differences are too extreme. Next, the pitches are made zero mean by subtracting the mean from all the pitches. A zero mean is required so that the difference between the two note inputs will be zero in the identical case. The next step is to transpose the test input to the same range as the reference. This is easier than it may seem. Pick the lowest note in each score; the transposition of the test score is divided by the ratio of the two lowest notes. This means that comparing a score transposed by a constant interval with the original score will be seen as identical, as they should.

A metric function was designed to calculate the distance between the test and reference notes that accommodates the structure

of the equal tempered musical scale.

$$||X - Y|| = \frac{\log(X) - \log(Y)}{\log(\sqrt[12]{2})} \quad (2)$$

The sum of the errors calculated over the time stretching procedure (see equation 1) is the end result of the time-warping phase; lower scores are better matches. All of the reference library scores are compared against the pitch track of the singer; the lowest score is the "winner".

Another sequence comparison method derived by Mongeau and Sankoff[14] from Knuth-Morris-Pratt string matching was employed by McNab, et al.[3]. McNab, et al. can use this method because they coursely quantize the pitches to "Up" and "Down".

5. EXPERIMENTAL RESULTS

A group of nine willing volunteers sung scales and then well known songs using the phoneme /ma/ in place of the words. This simplifies the analysis considerably since only one consonant and one vowel are used. The scale amply demonstrated the inability of untrained singers to maintain semitone pitch intervals or, in some cases, sing an octave.

5.1. Recording

The nine singers sang a number of songs including "Happy Birthday" and "Twinkle-Twinkle Little Star". These two songs were used in the experiments. They were recorded at 48 kilosamples per second and downsampled to 8 kHz with 16 bit samples. In the next section, a real example is processed from beginning to end.

5.2. Example processing

In Figure 3, the input waveform from a singer is given. The tune is "Twinkle-Twinkle Little Star".

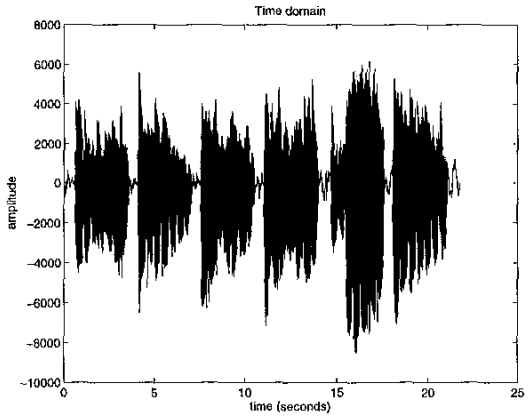

Figure 3: Time domain input

The result of the pitch detector is shown in Figure 4.

Next, the reference score is truncated to be the same number of notes as the input. Then, the time scales are made identical and the input is transposed. The result is shown in Figure 5.

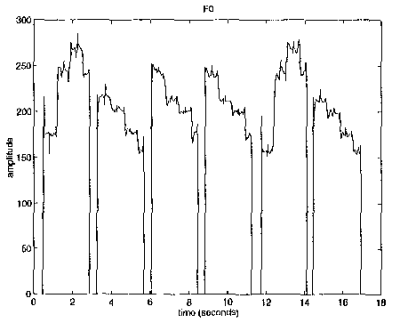

Figure 4: Pitch detector output

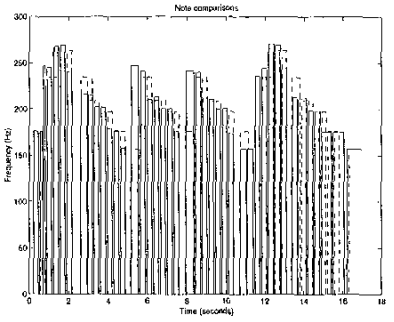

Figure 5: Time rescaled and transposed notes

The result of applying DTW can be found in Figure 6. This shows that the comparison against the song reference is nearly a diagonal after the time warping. The mean error was 0.0285. When compared against a different tune (Figure 7), the resulting error mean was 0.0934. This difference, though small, is sufficient to distinguish the two in all but one case.

5.3. Results

Errors result from the detection of pitch on and off times as well as from pitch detection errors. Pitch detection is easiest during vowels, but the transitions from vowel to consonant can cause problems. In spite of these inaccuracies, the prototype algorithm exceeds 95% accuracy for nine different singers for a simple tune library. The example shown above illustrates that mean errors, even for dissimilar songs, can result in a close miss. A better metric is needed to clearly illustrate song differences.

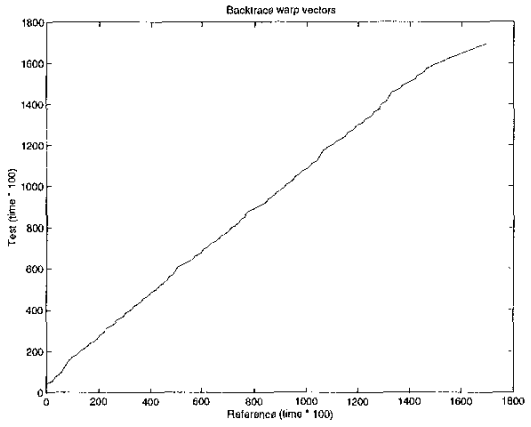

Figure 6: Backtrace vector from DTW for correct song

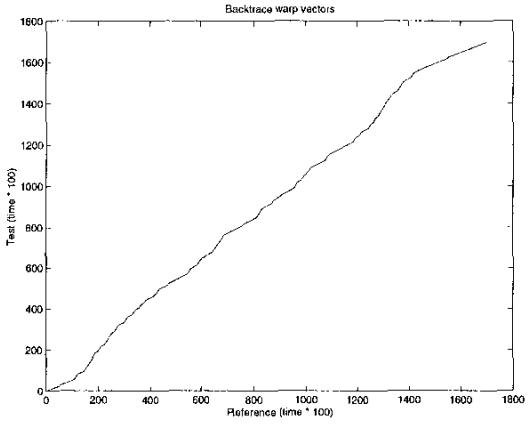

Figure 7: Backtrace vector from DTW for incorrect song

6. CONCLUSION

Systems like the one described here provide a first step toward automated access to musical scores and recordings. There are many further avenues of research including:

- The prototype system described here only works for a small number of scores. Large database searching techniques are needed to reduce search time.
- Expanding the range of the input to cover all vowels and consonants so that words can be used instead of simple phoneme, thereby increasing the naturalness of the input.
- Use of other acoustic inputs including humming, whistling and drumming

In spite of these limitations, the prototype system described here shows great promise for use in accessing web based musical databases. The basic methodology of timing identification and dynamic time warping could be used for other inputs including rhythmic inputs.

7. ACKNOWLEDGEMENTS

This research was initiated and supported by Steve Crandall. Julia Hirschberg backed him up and also provided detailed critical remarks on the processing of speech-like signals. Yannis Stylianou provided useful advice about Dynamic Time Warping. Niki Fallen helped Steve Crandall record the amateur singers.

8. REFERENCES

[1] Edward J. Coyle and Ilya Shmulevich, "System for machine recognition of music patterns," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. May 1998, vol. 6, pp. 3597–3600, IEEE.

[2] Ilya Shmulevich and Edward J. Coyle, "Establishing the tonal context for musical pattern recognition," in *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*. 1997, IEEE.

[3] R. J. McNab, L. A. Smith, I. H. Witten, C. L. Henderson, and S. J. Cunningham, "Towards the digital music library: Tune retrieval from acoustic input," in *Proceedings of the ACM International Conference on Digital Libraries*. March 1996, pp. 11–18, ACM.

[4] R. J. McNab, L. A. Smith, and I. H. Witten, "Signal processing for melody transcription," in *Proc. 19th Australasian Computer Science Conference*, January 1996.

[5] R. J. McNab, L. A. Smith, D. Bainbridge, and I. H. Witten, "The New Zealand Digital Library MELody inDEX," *D-Lib Magazine*, May 1997.

[6] J. Sundberg, *The science of the singing voice*, Northern Illinois University Press, 1987.

[7] J. Sundberg, "Formant structure and articulation of spoken and sung vowels," *Folia Phoniat.*, vol. 22, pp. 28–48, 1970.

[8] W. A. Hess, *Pitch Determination of Speech*, Springer Verlag, 1983.

[9] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier, 1995.

[10] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, pp. 103–138, 1990.

[11] Malcolm Slaney, "Auditory toolbox, version 2," Tech. Rep., Interval Research Corporation, 1998.

[12] L. Rabiner and B-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, February 1993.

[13] H. Sakoe and S. Chiba, *Dynamic Programming Algorithm Optimization for Spoken Word Recognition*, pp. 159–165, Morgan Kaufmann, 1990.

[14] M. Mongeau and D. Sankoff, "Comparison of musical sequences," *Computers and the Humanities*, vol. 24, pp. 161–175, 1990.