PITCH TRACK TARGET DEVIATION IN NATURAL SINGING

David Gerhard
Department of Computer Science, Department of Music
University of Regina
Regina, SK CANADA S4S 0A2
david.gerhard@uregina.ca

ABSTRACT

Unlike fixed-pitch instruments such as the piano, human singing can stray from a target pitch by as much as a semitone while still being perceived as a single fixed note. This paper presents a study of the difference between target pitch and actualized pitch in natural singing. A set of 50 subjects singing the same melody and lyric is used to compare utterance styles. An algorithm for alignment of idealized template pitch tracks to measured frequency tracks is presented. Specific examples are discussed, and generalizations are made with respect to the types of deviations typical in human singing. Demographics, including the skill of the singer, are presented and discussed in the context of the pitch track deviation from the ideal.

Keywords: Singing, melody alignment, ornamentation, pitch track, vibrato.

1 INTRODUCTION

Musical query systems are designed with the expectation that the singer will know they are making a query, and therefore consciously or subconsciously regularize their singing, reducing the impact of ornamentation like *vibrato* and *rubato* which have the tendency to make melody recognition very difficult. This paper is concerned with so-called "natural" singing, where the singer is not specifically attempting to develop a query. Even when these ornamentations are so extreme that the target melody may be unrecognizable by automated methods, human perception is capable of regularizing the pitch and timing to identify the melody.

Human perception of singing is very forgiving, considering that the pitch track of an average non-expert singer is far from the ideal sequence of pitches intended by the singer or heard buy the listener. Trained singers achieve target pitches much more rapidly and accurately than untrained singers, and the vibrato is more easily modeled by an idealized sinusoid, however, the amplitude of such oscillation is greatly increased, sometimes to well beyond a semitone. Rossignol et al. (1999) have shown how to detect and model vibrato in the context of musical pitch, but when the musician is not expert at controlling the vibrato, the resulting waveform is not well-formed and while it still sounds vibrato-like, it no longer fits with a reasonable vibrato model. Section 5.2 presents examples of utterances made by trained and untrained singers, with an examination of the pitch deviation for each.

Current query-by-humming systems have several advantages, perhaps subconscious, over the average human listener in this regard. In the extreme, many systems are designed to operate on idealized pitch tracks, either by using fixed-pitch instruments or variable-pitched instruments played by experts. Also, the users of these systems often know that they are making a musical query, and so sometimes try to make their musical utterances as smooth, rhythmic and "correct" as possible. Amateur singers, when just singing for fun or without intent, often generate less accurate pitch tracks, aiming high for a note and compensating later, or accidentally switching keys in the middle of the song. Even with this "messy" input, as with so many other cognitive tasks, the untrained listener can easily and accurately identify familiar melodies.

It is the pitch track of the utterance which seems to hold the majority of melodic information (Weyde, 2004), and it is the human brain which seems to be able to filter out ornamentation, errors and state changes and "lock-in" to the intended or target set of pitches. This paper will examine some of these errors, and show some of the typical deviations from the intended or target pitch sequence.

2 A NATIONAL ANTHEM AND A RIVER CRUISE

The data used for this study was initially collected to study the differences between speaking and singing (Gerhard, 2002). 50 subjects were prompted to speak and sing various lyrics. The utterances thus extracted included intermediate vocalizations like poetry and rap music, as well as spoken phrases and sung lyrics. This study used two pairs of utterances from each subject:

1. Please sing the phrase "Row, row, row your boat, gently down the stream."

2. Please speak the phrase "Row, row, row your boat, gently down the stream."

3. Please sing the phrase "O Canada, our home and native land."

4. Please speak the phrase "O Canada, our home and native land."

Having each subject answer all of these prompts ensures that all variables are controlled for except the differences between speaking and singing.

The subjects ranged in experience from self-confessed novices with little or no musical background or public speaking experience, to professional radio voices and trained opera singers. No pitches were given, and subjects sang in a number of keys. Some subjects sang in more than one key, and some, unintentionally, sang in multiple keys across a single melody. Most subjects were already familiar with the songs, and for those that were not familiar, and example was sung for them. Most subjects sung the "expected" tune, although one decided to "rap" our national anthem. The samples were recorded with consistent equipment (speech-recognition microphone and digital recording software) at 44.1 kHz. This paper contains results relating to Prompt 3, the first line of the Canadian national anthem, sung.

3 FEATURE EXTRACTION

The signal processing analysis used for this study concentrated on the pitch track of the utterance. Other features, such mel-frequency cepstral coefficients, were considered, but the specific pitch value and change over time is of particular interest in this work, so a direct estimate of the fundamental frequency of the signal was extracted and tracked over the course of the utterance. Not all human vocal utterance is periodic, however, so the segments of the utterance with pitch (the *voiced* segments) must also be identified.

3.1 Frequency estimation

The YIN frequency estimator (de Cheveigné and Kawahara, 2002) was used to do the initial pitch extraction. Several readily available frequency estimation algorithms were examined and evaluated including the algorithms available in Colea (Loizou, 2003), and YIN was shown to fit the purpose well. It responds well to human vocal sounds, and provides a measure of confidence which can be used as a detector of pitched segments. Each utterance was analyzed using YIN and the confidence measure was used to provide an initial segmentation.

3.2 Segment detection: energy and zero-crossing rate

The measure of confidence from the YIN pitch estimations was combined with a zero-crossing rate fricative detector and a thresholded RMS energy calculation to produce the initial set of note boundaries. The zero-crossing rate is a good estimator of the spectral centroid, (Kedem, 1986) and as such can be used to identify voiced segments of the utterance. The clip is divided into windows of 512 samples each, and the zero-crossing rate is measured for each window. If the zero-crossing rate is above a previously set threshold, the window was considered not to be voiced, and the segment is split at that point.

As a further attempt to identify possible note boundaries, the overall energy of the signal was calculated at each frame (as in the zero-crossing rate) and when the energy dropped below a pre-defined threshold for a set period of time, the segment was split at that point. To avoid noise at the threshold boundary, a hysteresis-like algorithm was employed, with a pair of thresholds. The energy would have to cross the lower threshold in the negative direction to indicate the end of a segment, and cross the higher threshold in the positive direction to indicate the beginning of a new segment.

3.3 Discussion

One difficulty with this procedure is that when notes change without a vocal stop or fricative, the note boundary is not identified. Other standard methods of detecting note boundaries include pitch track discontinuities, spectral envelope discontinuities, and changes in filter-bank energy levels. Unfortunately, none of these techniques are successful all of the time with human singing. Successive notes can produce identical features, especially if the singer is singing a series of notes on a single syllable (as is the case when humming). Singers usually bend pitch from one note to the next rather than making a discrete jump, especially if there is no breaking stop or fricative. If the notes are far apart, a threshold can be set such that the differential of the pitch track rising above this threshold indicates a note boundary. Singers with even moderate levels of vibrato can easily exceed half a semitone, so a threshold set high enough to avoid being triggered by vibrato may miss a valid semitone note transition.

4 ALIGNMENT ALGORITHM

Because the target tune is known *a priori* in this case, the alignment algorithm simply finds the best match between the extracted pitch track and the ideal, or target pitch track. The sequence of steps in this alignment algorithm is:

1. Identify note boundaries in the frequency estimate of the utterance under consideration.

2. Quantize to a single pitch for each segment.

3. Convert absolute frequency estimates (Hz) to relative frequency (cents).

4. Align the segments to the target pitches of the known melody

Because the target pitch sequence and rhythmic structure is known, a "best fit" can be achieved. For this procedure to be useful in a query-by-humming system, the first three steps are common to all matching tasks and can therefore be performed once on the incoming signal. This procedure works best, however, when the number of pitched segments from the estimation and the target are

the same. The segmentation problem (breaking the signal into pitched segments) is quite difficult for natural human singing. It should be noted at this point that the alignment algorithm presented here was intended only to allow analysis of the deviation from target of human singing.

4.1 Note boundary identification

As indicated above, the first estimate of the note boundaries is found using a combination of the confidence measure of YIN, the zero-crossing rate and the energy. This produces reasonable results but occasionally leaves pitch segments which should be separated into a series of notes. If these segments are not separated, the pitch quantization will be unsuccessful, since contributions from more than one note will produce erroneous results.

Having a target melody gives the algorithm a target for the number of pitch segments to expect. If the number of segments is significantly smaller than that, some segmentation must be done. The slope of the pitch is used to identify the next reasonable segmentation site. A pair of parameters are used to find this site: BoundaryLength and BoundaryThresh. The pitch slope must remain above BoundaryThresh for the duration BoundaryLength in order to identify a segmentation site. BoundaryLength is dynamically adjusted to account for different singing styles.

This method is quick and produces reasonable results for study, but is not completely robust. SInging style influences these results greatly, and as will be discussed in Section 5, singers tend to *glissando* or glide from one pitch to the next, reducing the ability of the algorithm to find a reasonable segmentation site. Frustratingly, this does not seem to affect human perception of the same melody—people can recognize a melody whether or not the singer is gliding from one pitch to another or jumping as briefly as possible . Lyrics help the recognition, but even without lyrics we humans can recognize a tune which deviates in segment pitches as well as notes and rhythm.

A procedure that has not been implemented in this system is re-combination. It would be useful to be able to join two segments which seem to be the same pitch or belong to the same note. the difficulty with this is that two segments with two similar pitches could equally be a single note erroneously split or a repeated note. Note onset and offset characteristics have the potential to help solve this problem.

4.2 Pitch quantization

Once the pitch track has been split into the appropriate number of segments, each segment is assigned a pitch which represents the entire segment. There are a number of ways to assign the overall pitch of the segment, the simplest of which is to calculate the mean pitch of the segment. With a segment containing idealized vibrato, the mean pitch will be at the center of the oscillation and correspond well to the perceived pitch of the segment. Unfortunately, pitch track segments often depart from the idealized vibrato at the beginning or end of the segment, indicating a transition to another note. The median may be a more appropriate measure in this case.

4.3 Frequency conversion

The target melody is constructed in terms of the number of cents from the melodic root note of the key. In the "O Canada" melody, the root occurs at the third note in the sequence. All other notes are indicated in cents from the root, and so the first note in the melody, a major third up from the root, is indicated at 400 cents. Of course, any note can be used as the base for this representation, and indeed in melodies where the key root is not present, another note will have to suffice. Since the cent scale is a relative scale, the starting note is unimportant - all semitone intervals are 100 cents (assuming equal temperament). Equation 1 shows the conversion from hertz to cents:

$$C_i = 1200 \times \log_2 \left(\frac{f_i}{f_0} \right) \tag{1}$$

where C_i is the relative pitch of the note in cents, f_i is the frequency of the note in hertz, and f_0 is the frequency in hertz of the base note.

The extracted melody is likewise converted to cents. Since we know in advance that the melody contains the root note and that it is the lowest note in the melody, we can use the lowest segment-quantized pitch as the "root" of the frequency track. Again, the choice of the base frequency is arbitrary, and could depend on any segment or an average of all segments, however, it is important to pick a root note such that it is possible to align the candidate track with the target track.

4.4 Segment alignment

At this point in the algorithm there are two distinct sets of pitch segments (candidate and target), and the task is to align them to the best fit. The rhythmic fit is approximated first by aligning the beginnings of the first and last notes. This was originally intended as an initial condition to an iterative rhythmic alignment process, but the rhythmic alignment was not implemented, primarily because the initial alignment was found to be sufficient for our purposes. This is likely because rhythmic target deviation is typically very small compared to pitch target deviation. Future implementations of this system would include an algorithm to fine-tune the beginnings and endings of the target melody, and to evaluate the relative rhythm error. Aligning the beginning of the last segment gave better results than aligning the end of the last segment because note offset accuracy is much less important for melody identification than onset accuracy, and singers tend to cut off final notes in a phrase earlier than other notes.

Aligning the segments by pitch consisted of calculating the ratio between each pair of segments, and averaging these ratios across all segments, weighted by the segment length. This calculated the scaling factor which would bring the measured pitch track into as close alignment as possible with the ideal pitch track. Because the tracks have already been converted to cents, the pitch intervals are relative and so two identical melodies in different keys differ only by a scaling factor. In this case, the ratio is found which best aligns the measured pitch to the ideal pitch.

5 RESULTS

In this section a set of figures is presented showing examples of the pitch track phenomena observed in the course of this study. These figures are presented in pairs, with the first figure containing a complete melody track with a highlighted pitch segment, and the second figure containing an enlarged plot of the segment in question. Both plots are shown in relative pitch, with different scales. The complete melody lines are constructed so the root note of the target melody line is at 0 cents. The individual segment plots are built with the target pitch at 0 cents, which clearly shows the deviation from the target in cents. Figures 1 and Figures 2 show an example of these plots. The singer of this clip is subject 211, and the "g" refers to the 7th prompt in the generalized list (corresponding to the singing of "O Canada")

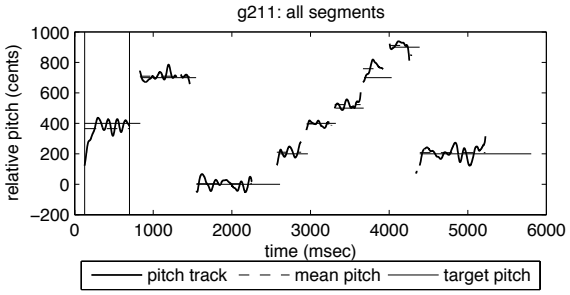

Figure 1: Pitch track with mean and target pitches.

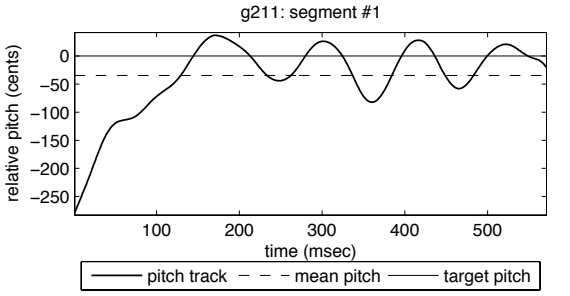

Figure 2: Indicated segment from Figure 1.

This first pair shows the success of the alignment algorithm as well as the deviation in pitch scale of the singer. In general, the singer is "in tune," and the melody track is recognizable, but there are some things of note here with regard to the deviation from the target. First, the singer glides up to the highlighted note. This is of particular interest because, as indicated previously, no cueing pitches were given. Subjects sang in whatever key they chose. If the subjects were singing in arbitrary keys, why should subjects glide up to the first note they sing? Perhaps it is because they have a target pitch in their head, and as they start singing they notice and correct the mis-tuning, until the target pitch is reached. Control systems work this way as well—a target is chosen, and the system can only respond in a finite amount of time, thereby approaching the target over a period of time. Professional singers avoid this initial glissando by holding a mental model of the note to be sung before vocalizing the note. Choir directors instruct their singers to "Think the note before you sing the note."

The next sections and figures provide a discussion of some of the typical deviations from the target pitch that were observed in the course of this research.

5.1 Vibrato

Vibrato is a well-known phenomenon in musical analysis, wherein the frequency of a voice or instrument is modulated by a pseudo-sinusoidal waveform. Prame (1994) Showed that in singing, this modulation is usually around 6.0 Hz. Typically, the formants which characterize the phoneme being sung do not change with the vibrato, which means that as the frequency partials oscillate in and out of the frequency peak of the formant, their amplitude increases and decreases as well. Vibrato blurs the pitch realization, making it more difficult to determine the intended pitch target.

Many novice and expert singers studied in this research used vibrato in their singing. It is theorized (although as yet unsubstantiated) that individuals using a query-by-humming system may, consciously or subconsciously, attempt to reduce the amount of vibrato in their singing, and to flatten their pitch tracks, to clarify and regularize their query. Vibrato has the perceptual effect of tightly coupling the partials of the note being sung, allowing it to be heard above other sounds. This is one of the reasons that opera singers utilize higher-frequency and higher-amplitude vibrato than popular singers, who use microphones and amplifiers to achieve the same purpose.

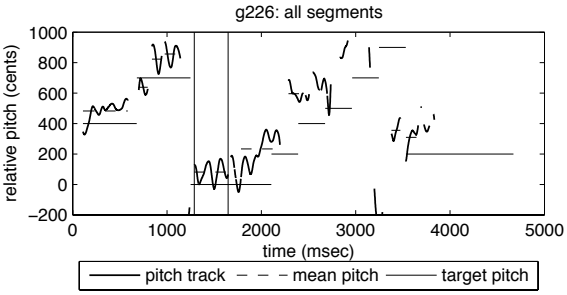

Figure 3: Mis-aligned pitch track with high-amplitude vibrato.

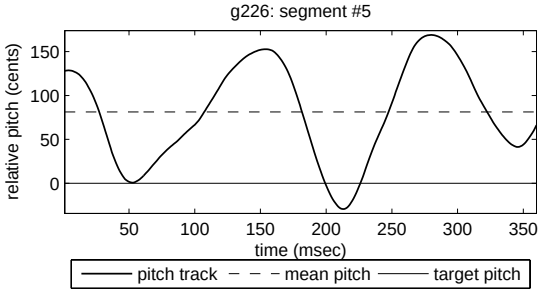

Figure 4: Indicated segment from Figure 3.

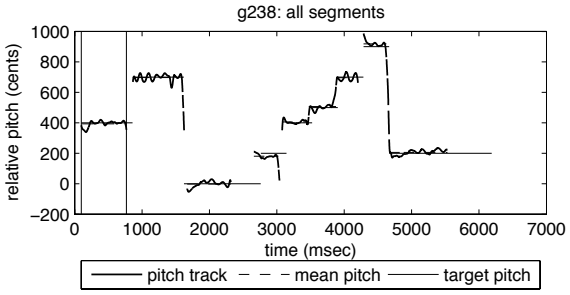

Figure 5: Pitch track with low-amplitude vibrato.

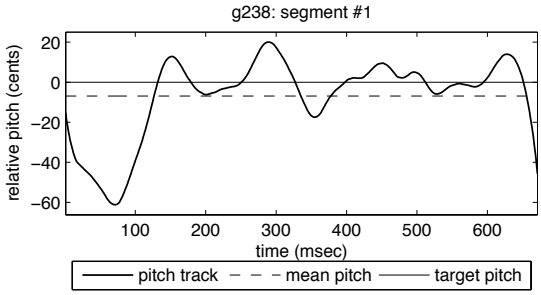

Figure 6: Indicated segment from Figure 5.

Figures 3 and 4 show an example of a situation where high-amplitude vibrato can interfere with the retrieval of pitch target information. Subject 226 is a trained tenor soloist, and produces vibrato which ranges almost two semitones from lowest to highest pitch in the segment shown in Figure 4. Figure 3 shows how the vibrato has make the investigation of the pitch track difficult. It is clear that the alignment algorithm has failed in this case, and it is difficult even to follow the extracted pitch track with the eye. The general contour is there but without a template to match to, identifying the melody of that utterance using this or any other method would be quite difficult. This is in contrast to Figure 1 where the discrete notes and melody of the extracted pitch track are easy to follow visually. It is interesting to note here that the recordings from Subject 226 are arguably the most perceptually cohesive, and while many would consider 226 the "best" singing of the 50 singers in the set, it is one of the most difficult to track algorithmically.

In contrast, Figures 5 and 6 show an example of a low-amplitude tight vibrato. The pitch track is the most visually consistent of the set, and the most easy to follow with the eye. It would not be difficult to design a system to extract the melody from this signal without any *a priori* knowledge of the ideal pitches or ideal rhythm—both are strongly adhered to by the singer.

5.2 Expertise of the Singer

Subjects were asked to indicate their level of experience with singing, with music, and with public speaking. Some subjects claimed little or no experience, while others were expert or professional. The differences in the pitch tracks between the novice and the expert is quite interesting.

Figures 7 and 8 show the pitch track for subject 219,

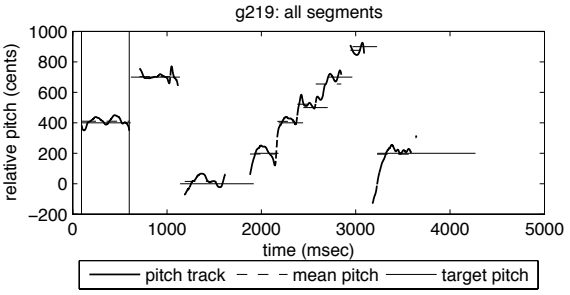

Figure 7: Pitch track of a novice singer.

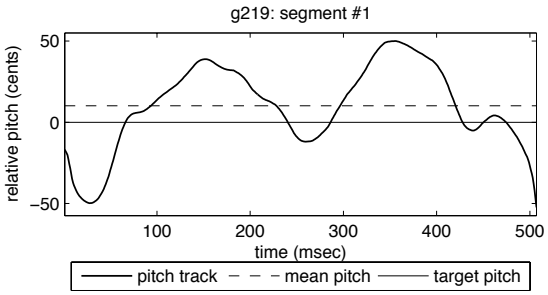

Figure 8: Indicated segment from Figure 7.

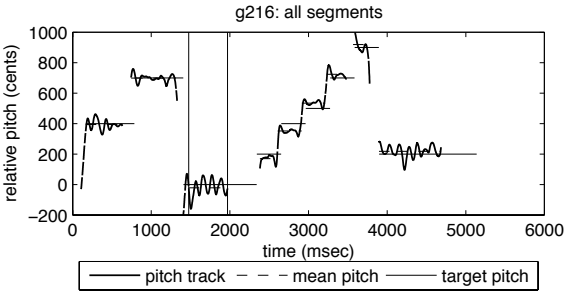

Figure 9: Pitch track of a trained singer.

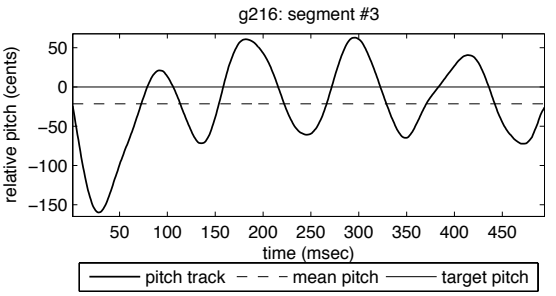

Figure 10: Indicated segment from Figure 9.

an example of a novice singer, and Figures 9 and 10 show the pitch track for subject 216, an example of a trained singer and musician. When asked to describe their musical, choral or spoken voice training or experience, subject 219 indicated 2 years of theatrical speaking, and subject 216 indicated 10 years of piano, one semester of a voice workshop, and 5 years in choirs. Note particularly that although the pitch rises and falls above the intended target, the novice singer does not have a consistent vibrato and

appears to be attempting to hold the pitch constant, while the experienced singer has a well-formed and intentional vibrato, closely resembling a sinusoid. It should also be noted here, however, that the novice is singing "in tune" just as well as the experienced singer.

The experience level of the singers whose tracks appear in this paper are summarized in the following list:

211 (age 59) 10 years trumpet, 3 years piano, 25 years choirs

216 (age 32) 10 years piano, 1 semester voice workshop, 5 years choirs

219 (age 24) 2 years theatrical voice

226 (age 26) 4 years classical voice, 20 years choirs, solo training, 2 years theory

238 (age 39) 6 years guitar, 5 years guitar, 1 year ukelele, 2/3 year public speaking, 9 years informal speaking

Another interesting observation is that the "quality" of the singing of subjects with experience but without specific vocal solo or opera training was uncorrelated with the amount of musical or vocal experience they had. Most singers in this range produced consistent pitch and tone with relatively well-formed vibrato. Only the subjects with very little or very much training exhibited exceptional characteristics in their pitch track, and even then, the single subject who claimed no experience whatsoever produced a pitch track comparable to those claiming years of experience in choirs.

5.3 Onset and offset

Looking at Figure 9, it is clear that as the singer ascends the major scale consisting of 2, 4, 5, 7, and 9 semitones above the root, that the pitch descends slightly before the pitch transition, and rises above the target pitch before settling into a vibrato oscillation. This is a common observation across the data set. Singers rarely make a clean break between notes, and whether they glide up to or down to the target pitch depends on the previous note, if there is one. Observe in Figure 5, toward the end of the clip the pitch glides down in a very short period of time, but remains continuous during that glide. This is unusual in the data set—most singers produce pitch tracks more like that seen in Figure 1, where a complete break is made and the pitch glides up to the target of the last note.

6 CONCLUSIONS

A singing human aims for a consistent note, and uses an oscillation around that note to solidify the perception in the ears of a listening human. Listening machines must therefore take this intentional deviation from the target into account in order to accurately transcribe human singing. In many cases, pitch is one of the only available features for transcribing human singing. Deviations from the ideal are not universal, and models of vibrato in a song recognition system must take these into account while being able to distinguish between vibrato pitch changes and note transition pitch changes.

The level of experience of the singer has an impact on the type of deviation shown in the singing. Novice singers and singers beginning to study "professional" singing may have erratic vibrato and unpredictable glide transitions, while experienced singers tend to have more regularized vibrato and note transitions. Knowledge of the singer in question would be very useful, and for that reason it may be worthwhile to investigate ways of classifying human singing with the intent of developing personalized deviation models or a set of standardized deviation models based on feature clustering.

ACKNOWLEDGEMENTS

This work is supported by the Natural Sciences and Engineering Research Council of Canada and the Canadian Foundation for Innovation.

REFERENCES

A. de Cheveigné and H. Kawahara. YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, 111(4), 2002.

D. Gerhard. A human vocal utterance corpus for perceptual and acoustic analysis of speech, singing and intermediate vocalizations. *Journal of the Acoustical Society of America*, 112(5):2264, November 2002.

B. Kedem. Spectral analysis and discrimination by zero-crossings. *Proceedings of the IEEE*, 74(11):1477–1493, November 1986.

P. Loizou. COLEA: A matlab software tool for speech analysis. [Online] Retrieved March 18, 2003, from http://www.utdallas.edu/~loizou/speech/colea.htm, 2003.

E. Prame. Measurements of the vibrato rate of ten singers. *The Journal of the Acoustical Society of America*, 96 (4):1979–1984, October 1994.

S. Rossignol, P. Depalle, J. Soumagne, X. Rodet, and J.-L. Colette. Vibrato: Detection, estimation, extraction, modification. In *Proceedings of the COST-G6 Workshop on Digital Audio Effects (DAFx-99)*, December 9–11 1999.

T. Weyde. The influence of pitch on melodic segmentation. In *Proceedings of the 5th International Conference on Music Information Retrieval*, pages 128–132, Barcelona, Spain, October 2004.