Some Considerations About Processing Singing Voice for Music Retrieval

Emanuele Pollastri
Dipartimento di Scienze dell'Informazione
Università degli Studi di Milano
via Comelico, 39/41
20135 Milan – Italy
+39-02-50316297
pollastri@dsi.unimi.it

ABSTRACT

The audio processing and post-processing of singing hold a fundamental role in the context of query-by-humming applications. Through the analysis of a sung query, we should perform some kind of meta-information extraction and this topic deserves the interest of the present paper. Considering the raw output of a pitch tracking algorithm, the issues of note estimation and the study of singing accuracy have been addressed. Further, we report an experiment on the deviations from pure tone intonation in performances of untrained singers.

1. INTRODUCTION

The need for content-based music retrieval tools has been already stressed by a number of publications; in this context, one of the most appealing applications seems to be querying-by-humming [1, 2, 5, 7]. The singing voice as another possible channel of interaction with machineries is considered easy, natural and enjoyable for both music professionals and casual users. Nonetheless many problems remain open in building an effective system for querying-by-humming. In particular, the singing voice should be better investigated. The aim of this paper is to highlight useful aspects of singing voice to improve the translation from acoustic events into a note-like representation.

2. NOTE ESTIMATION AND SINGING VOICE ARTICULATION

A raw pitch tracker output will produce an accurate pitch contour in the "microintonation" sense. In music retrieval we are interested in the "macrointonation" aspect, which is the sequence of notes sung by a singer. Determining the latter contour starting from the former one is a difficult task especially when we are dealing with untrained voices, for which the degree of pitch variability is affected by a number of factors like the vocal range and singer's emotions. The actual pitch of a note is only an abstraction of the microinto-nation. In order to be able to reconstruct the intended sequence of notes, some "hard" smoothing, like low-pass or median filtering, is surely necessary but it is not sufficient.

In our experimentation, we are using a set of heuristic rules obtained by studying the singing sequences contained in a dataset made of forty tunes sung by non-professional singers. The quasi-stationary conditions are met for frames 5 milliseconds long; both pauses and notes could not be less than 120 milliseconds in duration. The pitch-tracker employed has been presented in [6].

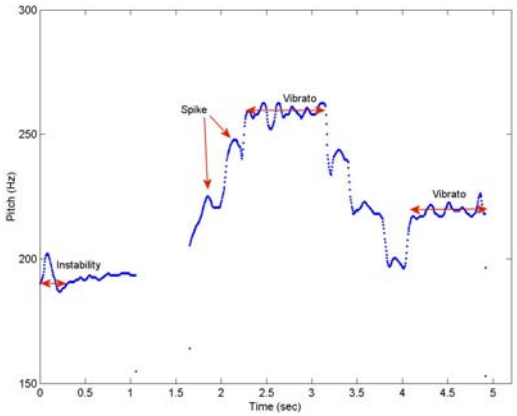

Figure 1. Example of pitch events related to the translation from microintonation to macrointonation.

Aiming to avoid local discontinuities, thus improving note recognition, the frame-level pitch contour is post-processed with a median filter and a running mean block with windows 100 msec. long, overlapped by 95 msec (i.e. shifting each window of 5 msec). Each note hypothesis (duration and pitch), is evaluated to match one of the following situations:

1- *Spike*: it is a tone, often short (around 100 msec), characterrized by a monotonically increasing sequence of pitches followed by a monotonically decreasing one. Spikes are often encountered in short note repetitions; sometimes they are used as ornamental notes. The value of pitch for this kind of event is the local maximum.

2- *False Spike*: event 60 msec. long with a mean pitch value in a range of ±0.6 semitones within the next tone. A false peak should be merged with the following note.

3- *Ascending/Descending Intervals*: in a sequence of two or more ascending/descending notes (not separated by silence) is common to have an unstable region during the first 70/120 msec. in the shape of a glissando. Usually, it does not happen with short notes, so the solution consists in discarding the unstable fragment and taking the mean or the median of the remaining pitch values.

4- *Vibrato*: an event corresponding to a quasi-sinusoidal modulation of pitch [8]. The center of the oscillation is the pitch.

In a practical implementation, the situation indicated as number three is the rule while the others are the exceptions. In Figure 1, a curve indicating a microintonation contour is commented to highlight spikes, vibrato regions and local instability.

3. ACCURACY OF INTONATION

In the previous section we reviewed the problem related to transform a sequence of frame-level pitch contour to a note-level contour. We did not mention that we are operating in the framework of the equal tempered musical scale, so the reference unit is the semitone and the fractions of semitone (or cents). While this choice is debatable, it is worth to be adopted for two main reasons. The precision of a query-by-singing system depends on the understanding of the query. Therefore, a nearly perfect translation of the input at the semitone level remains an important goal, even for system based on simpler representation like 3 or 5 level interval contour. Moreover, previous works on the question are still based on equal tempered musical scales, so for the sake of comparison we will continue with this convention.

In a previous paper [3], the author stressed the importance of adopting a musical scale relative to the singer. This solution is needed because singers very rarely have a tone-absolute pitch. It has been confirmed that singers made constant-sized errors, regardless of note distance in time or in frequency [4]. Thus the relative scale coincides with a shifted equal tempered musical scale. The amount of the shifting can be calculated out of the performance of the singer. An open question regards the accuracy of intonation in untrained singers. We can speculate that the overall deviations from pure intervals on average tend to go to zero. In other words, singers are able to adjust their intonation during singing. It should confirm Sundberg's intuition for which singers must hear the next target pitch before starting to change the pitch [9]. From our experience, we should add that trained singers are able to do this operation in their mind before giving any sound, while untrained singers need to physically hear their voice. However, what about the accuracy of intonation of every single interval?

Using our dataset (40 tunes), we collected some data trying to answer this question. A word of warning is needed since in this experiment singers were not asked to repeat a set of tone with intervals studied to cover all the possible configurations. Instead, they sung well-known tunes prepared for a music retrieval test. Within sung intervals, we measure the average "flatness" or "sharpness" without taking care if that interval was the right one. For instance, if an interval was estimated to be 3.75 semitones, we considered it as 4 semitones with 0.25 semitones flat, no matter if the right interval for the tune was 3 semitones. The amount of correct intervals can be inferred by the difference between the expected and the estimated distribution of intervals.

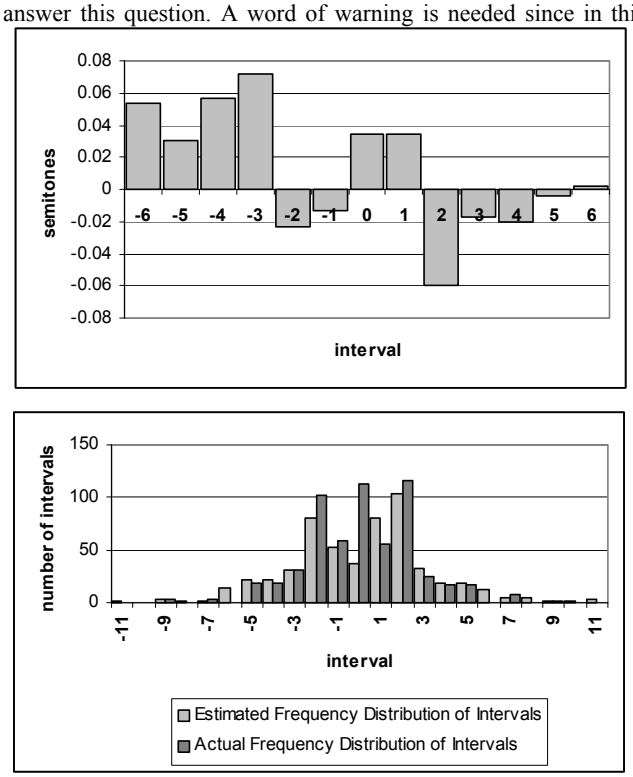

Figure 2. Deviations of intervals from pure tones in our dataset (top). Number of estimated and actual intervals (bottom)

Even with all the given precautions, untrained singers did not show considerable deviation for any interval (see Figure 2, top). These results contrast with the accepted notion that there is a trend to tune narrowly minor intervals and widely major intervals [26]. Furthermore, the curves of estimated and expected distributions are quite similar with the exception of the interval of zero size (note repetition) (see Figure 2, bottom); the segmentation algorithm employed in the audio analysis did not probably track that interval correctly [6]. A deeper investigation, in fact, has shown that most missing notes were repetitions. We over-estimated intervals of minor second (interval=1 semitone) but we missed intervals of major second and unison (interval=0 and 2 semitones). Thus we can argue that some difficulties actually exist in tuning repetitions and second major intervals, which are often tuned respectively sharp and flat. In the light of these findings, one could debate that the average measure of deviation tends to be wider than a semitone. Therefore, the presented statistics are at least questionable. Lacking formal investigations on this topic, further experiments are needed to dispel all doubt.

4. REFERENCES

[1] Ghias, A., Logan, D., Chamberlin, D., Smith, S.C. Query by humming – musical information retrieval in an audio database. In Proc. of ACM Multimedia'95, San Francisco, CA., Usa, Nov. 1995.

[2] Haus, G. and Pollastri, E. A multimodal framework for music inputs. In Proc. of ACM Multimedia 2000, Los Angeles, CA, Usa, Nov. 2000.

[3] Haus, G. and Pollastri, E. An Audio Front End for Query-by-Humming Systems. In Proc. of ISMIR 2001, Bloomington, IN, Usa, Oct. 2001.

[4] Lindsay, A. Using contour as a mid-level representation of melody. M.I.T. Media Lab, M.S. Thesis, 1997.

[5] McNab, R.J., Smith, L.A., Witten, C.L., Henderson, C.L., Cunningham, S.J. Towards the digital music library: tune retrieval from acoustic input. In Proc. of the 1st ACM Int. Conf. on Digital Libraries, Bethesda, USA, March 1996.

[6] Pollastri, E. A pitch tracking system dedicated to process singing voice for music retrieval. To appear in Proc. of IEEE International Conf. on Multimedia and Expo 2002, Lausanne, Switzerland, Aug. 2002.

[7] Prechelt, L. and Typke, R. An interface for melody input. ACM Trans. On Computer Human Interaction, Vol.8, 2001.

[8] Rossignol, S., Rodet, X., Soumagne, J., Colette J.L. and Depalle P., Automatic characterisation of musical signals: feature extraction and temporal segmentation. Journal of New Musical Research, Vol. 28, N. 4, Dec. 1999.

[9] Sundberg, J. The science of the singing voice. Northern Illinois University Press, Dekalb, IL, 1987.