Pitch-Based Solo Location

Gilles Peterschmitt
Emilia Gomez
Perfecto Herrera
Music Technology Group, Pompeu Fabra University
http://www.iua.upf.es/mtg
{gilles.peterschmitt, emilia.gomez, perfecto.herrera}@iua.upf.es

Abstract

The aim of this work is to study how a pitch detection algorithm can help in the task of locating solos in a musical excerpt. Output parameters of the pitch detection algorithm are studied, and enhancements for the task of solo location are proposed. A solo is defined as a section of a piece where an instrument is in foreground compared to the other instrument and to other section of the piece.

1 Introduction

Music browsing is a manifold activity that comprehends behavior as diverse as retrieving songs by different musical criteria, creating play lists that follow subjective criteria, selecting special excerpts, visualizing rhythmic or harmonic structures, etc. Recent advances and research projects on music content processing [1, 2, 3] give way to think that some of those activities will be performed soon in an automatic user-configurable way.

Instrumental solos are interesting and characteristic parts of a musical piece with a special status not only for a musicologist but even for a home-listener. A music browser should then provide with some functionalities in order to allow the user:

- to spot and fast browse solos inside music works;
- to visualize relevant music information of the solo (i.e. the score);
- to compile a list of solos by a given performer, or by a given instrument from the available music database; provided an adequate constraint satisfaction system, this mega-solo play list could follow some subjective and musical directions [4].

Besides the mentioned practical motivations, solo location has a research-related interest as a type of pre-processing intended to be useful for deriving instrumentation descriptions of complex music mixtures.

As far as we have been able to trace, there is no specific literature on automatic solo location. Therefore we have started our study with a conceptual analysis of the possible acoustic differences between what we may consider a "solo" and what we may consider an "ensemble" performance.

In this paper, a solo is defined as a section of a piece where an instrument is in foreground compared to the other instruments and to other sections of the piece. In physical terms, this means that spectra of solo sections should be dominated by one instrument. It is clear that the previous definition is highly debatable from the musicological point of view, but it should be accepted as a reasonable starting point from where some refinements can be done after careful study and testing. For a more formal definition of what can be considered a musical solo, the reader might consult the Grove Dictionary of Music [5].

One of the first approaches to getting some discriminative data for solo sections is looking at some spectral complexity measure. We assume that in audio segments where ensemble performance is predominant, the spectrum is more complex (i.e. with larger variability of spectral peaks location and amplitudes). Given that in the context of music browsing some pitch information is required, we thought that the above-mentioned measurements could be obtained as a "side-effect" of a pitch extraction process (hence without increasing the computational load of a system). We argued that in solo sections, pitch could be reasonably tracked by a common monophonic pitch detection algorithm, the *Two-Way Mismatch* [6], and therefore the pitch error indexes to be found in solo sections would be smaller than those to be found in "ensemble" sections. As we will see, the error indexes did not show enough discriminative power, and further enhancements were attempted.

2 TWM algorithm description

The used pitch estimation algorithm is described at [6]. This algorithm tries to extract a fundamental frequency from a set of spectral maximum of the magnitude spectrum of the signal. These peaks can be compared to the predicted harmonics for each of the possible candidate note frequencies. A particular fitness measure is described in [6] as a "Two-Way Mismatch" procedure. For each candidate, mismatches between the harmonics generated and the measured partials frequencies are averaged over a fixed subset of the available partials. The discrepancy between the measured and predicted sequences of harmonic partials is referred as the *mismatch error*. The solution presented on [6] is to employ two mismatch error calculations.

The first one is based on the frequency difference between each partial in the measured sequence and its nearest neighbor in the predicted sequence. The second is based on the mismatch between each harmonic in the predicted sequence and its nearest partial neighbor in the measured sequence.

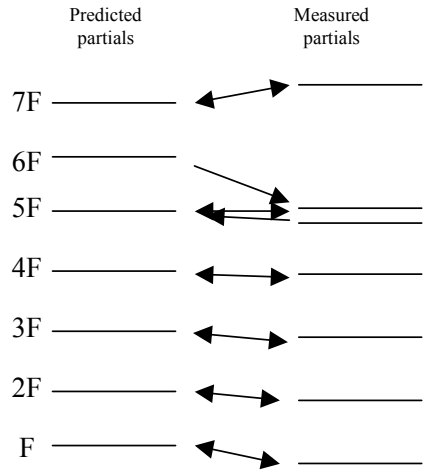

Illustration 0: TWM procedure

The two error measurements are computed as following:

- Predicted-to-measured mismatch error:

$$Err_{p\to m} = \sum_{n=1}^{N} E_\omega(\Delta f_n, f_n, a_n, A_{\max})$$

$$= \sum_{n=1}^{N} \Delta f_n \cdot (f_n)^{-p} + \left(\frac{a_n}{A_{\max}}\right) \times \left[q\Delta f_n \cdot (f_n)^{-p} - r\right] \quad (1)$$

where a_n, f_n correspond to the amplitude and frequency of the predicted partial number *n*, A_{max} is the maximum amplitude, and Δf_n is the difference between the frequency of the predicted partial and its closest measured partial.

- Measured-to-predicted mismatch error:

$$Err_{m\to p} = \sum_{k=1}^{K} E_\omega(\Delta f_k, f_k, a_k, A_{\max})$$

$$= \sum_{k=1}^{K} \Delta f_k \cdot (f_k)^{-p} + \left(\frac{a_k}{A_{\max}}\right) \times \left[q\Delta f_k \cdot (f_k)^{-p} - r\right] \quad (2)$$

where a_k, f_k correspond to the amplitude and frequency of the measured partial number k, A_{max} is the maximum amplitude, and Δf_k is the difference between the frequency of the measured partial and its closest predicted partial.

The total error for the predicted fundamental frequency is then given by:

$$Err_{total} = Err_{p\to m} / N + \rho \cdot Err_{m\to p} / K \quad (3)$$

The parameters *p, m, r* and ρ are set empirically and vary for each instrument.

3 TWM Output Errors Behavior

2.1 Errors

The three errors are crucial for pitch detection. However, from its definition, the PM error will be of first interest for our purpose. Indeed, the PM matches a set of predicted peaks with the set of measured spectral peaks. Its values are therefore usually lower and less erratic than that of the MP error, which tries to match a great number of measured peaks with the predicted peaks. The total error, which is a weighted combination of both errors, does not need precise description. We will therefore focus our attention to the study of the PM error.

2.2 PM Error Behavior

The optimal parameters input to the pitch detection algorithm are set carrying out tests on monophonic recordings of the instrument considered. If the parameters are optimally set for this instrument, the algorithm estimates the pitch correctly and the PM error is usually minimal. Using the algorithm on polyphonic sounds does not enable good pitch

estimation but leads to interesting output errors behavior.

For example, if the parameters are set optimally for a saxophone, good pitch estimation occurs if the saxophone plays on its own. The PM error is at its lowest as there is an evident match between the predicted peaks and the peaks present in the spectrum. In the presence of other instruments, the error is high (due to the addition of spectral peaks that belong to different harmonic series and instruments) and pitch estimation is usually corrupted. But if the saxophone "dominates" enough, some of the pitch can still be estimated. In the spectrum, some of the harmonic peaks of the saxophone are detectable and sensible matching is possible. The PM error in this case will be higher than in the monophonic case, but lower than when no instruments are in foreground. Moreover, the parameters being optimized for a given instrument -for example the saxophone-, it should give a lower error if the saxophone is in foreground than if an instrument with very different spectral characteristics is in foreground.

Figure 1 below shows the PM error output to the analysis of an extract of a piece by a Miles Davis ensemble. In this extract, the background (piano, drum and bass) is very quiet, and the saxophone plays clear and relatively loud notes, making the example visually explicit (note that it is not always the case).

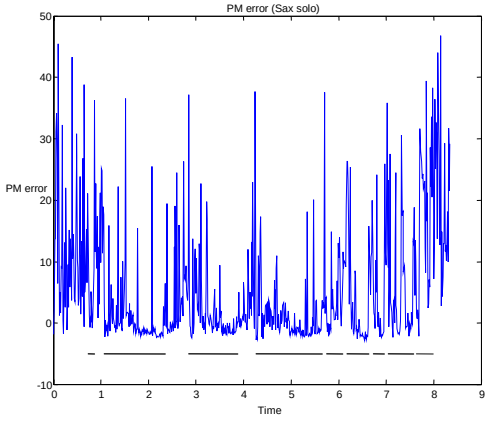

Figure 1: PM error against time for an extract of a piece by a jazz ensemble. The thick lines under the curve show the saxophone notes. Clear decreases can be observed when the saxophone plays.

The PM error behavior described above can be observed. The PM error, high for the ensemble mode, decreases a lot when the saxophone presence is dominant.

4 Solo Location

This behavior of the PM error, which in turn gets reflected in the Total error, can be used as a sign of spectral complexity, and therefore help in the task of solo location. Theoretically, the discrimination between "ensemble" and "solo" mode should be possible by detecting long-term changes in the mean of, say the PM error –high values representing "ensemble" mode and low values, solos.

From the statistical analysis of frame by frame data, it is clear that frames corresponding to solos have different average error values than frames corresponding to ensemble sections, and that both data distributions might have different origins (t-student -for means- and Kolmogorov-Smirnov -for distribution- tests results are omitted for convenience). Notwithstanding, an attempt to use a linear classifier (PDA) with the three features as predictors of category (solo/ensemble) yielded a mere 56% of success, after cross validation with a Jackknifed procedure. This was a disappointing result that could not be improved even using a quadratic classifier.

It seems clear, then, that solo location using the error parameters for the discrimination cannot be made on a frame by frame basis, as the variability of the error is large compared to the change in the mean we want to detect. Also, the change in the mean is neither very large nor neat at the solo boundaries. Using the TWM output PM and Total error alone, does not enable proper solo location.

A proposed way to tackle this problem was to use segmentation techniques such as Foote's similarity matrix [7] prior to the discrimination. Using such a technique with appropriate descriptors should enable to locate the boundaries of the different "parts" of the piece. The ensemble/solo discrimination using the long-term PM error level changes can then be done on these pre-located segments.

Studies were carried out in order to find relevant descriptors for locating these specific boundaries. Three ones were found to be particularly useful for this purpose: the spectral centroid, the skewness and the kurtosis. Adding the PM and Total error in the feature vector enabled better boundary location in a few cases where the spectral parameters were not sufficient for a good segmentation. However it sometimes misleads the segmentation result more than it helps. On 15 tests, adding these parameters helped 5 times (3 of which enabling segmentation when it was not previously possible), and corrupted the results importantly on only one occasion.

A summary of the procedure is given below:

1. The TWM input parameters, that are empirically set, are adapted to the solo instrument in order to improve the fundamental frequency estimation. In this step, monophonic sounds have been used.

2. The algorithm is applied to polyphonic sounds and the spectral features are calculated (Spectral Centroid, Skewness, Kurtosis). The analysis is done by frames of 0.0161s (512 samples at 44.1 kHz). This short frame length is necessary for a good performance of the pitch detection algorithm. The PM error, Total error and the three spectral parameters are extracted.

3. The features, averaged over segments of 50 frames, are input to the Foote's segmentation algorithm, and the candidates for Ensembles-Solos boundaries are automatically located according to the value of a "novelty score" (see figure 2) (the parameters for this step have to empirically set, although in the future the algorithm could adapt to the data).

4. The PM error is averaged over these pre-located segments and a decision taken for Solo or Ensemble mode.

5 A Case Study

As our initial database for testing is still rather small (15 songs), no significant and robust numerical data can be provided. Anyway, a case study will illustrate some specificities, pros, and cons of the procedure.

The following example illustrates the analysis for a piece by a John Coltrane ensemble. The ensemble is formed of alto saxophone, trumpet, piano, drum and bass. The piece starts with the ensemble until a saxophone solo starts around 37 seconds into the piece. The following parameters were input to the algorithm:

TWM input parameters (Optimal saxophone parameters):

- Window length of 0.0161 s (512 samples at sampling rate of 44100 Hz);
- Pitch range: the pitch detection analysis was carried out between 1000 and 3000 Hz;
- TWM parameters: p=0.5, q=1.4, r=0.5 in Equations (1) and (2), ρ=0.33 in Equation (3).

Segmentation parameters:

- Features were averaged over segments of 50 samples;
- Mahalanobis distances were calculated between feature vectors;
- Size of kernel: 30 averaged segments;
- Threshold for peak picking in Novelty Score curve: 3000 (empirically set).

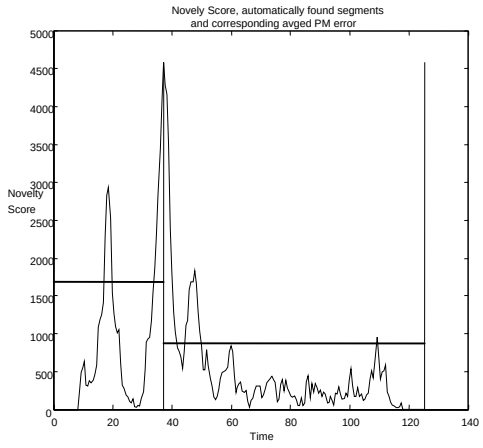

Figure 2: Novelty Score against time, with automatically found segments and corresponding PM error levels for an extract of "Blue Train" by John Coltrane. The transition form the theme to the solo was accurately found, and the error levels enable good discrimination.

The analysis of this piece showed to be particularly successful. The theme/solo boundary, located at 37 seconds into the piece was automatically found at 37.2 s, and the change in the mean of the PM error goes from 8.8 during the theme to 11.2 during the solo, enabling good discrimination. It can be noted that the good performance of the algorithm is very dependent on the piece, and that this piece is particularly suited for this purpose. There is a clear spectral change from the theme to the solo enabling a good boundary location. The solo is clear and continuous against a quite background giving easy ensemble/solo discrimination. This is not the case for all pieces.

6 Discussion and further work

First of all, problems with the TWM algorithm performance were encountered when dealing with instruments whose spectra do not show clear harmonic behavior. For example, very good results were obtained with quite harmonic solo instruments as saxophone, trumpet and violin, but with guitar and piano, the TWM could not give reliable results. A solution proposed to this problem is to use pre-processing techniques, such as the "noise suppression" technique (to remove additive and convolutive noise) proposed by A. Klapuri [8]. This technique enables to boost the spectral peaks, possibly enhancing the TWM performance. Including

an inharmonicity factor to correct for the stretched harmonics of the piano could also be beneficial.

Another problem resides in the automatic segmentation, whose input parameters vary from pieces to pieces (i.e. threshold for peak detection in novelty score curve, size of kernel, etc.). Studies have to be carried out in order to find ways to adapt the algorithm to the data. Also, more features should be added to the segmentation algorithm in order to get more robust boundaries location. Measures of spectral peak variability and amplitude modulation patterns (in order to detect beatings) are under current scrutiny.

Finally, the main problem resides in the concept we want to extract. This algorithm is basic and uses low-level descriptors to extract a rather abstract concept. This causes a number of inconsistencies and limits the robustness of the algorithm. First of all, the location is done from low-level features, which enables to locate solos with respect to these features only. That is, in order for a solo to be detected as such, it has to be:

- clear: relatively loud solo instrument compared to the background;
- continuous: if the solo consists of solo instrument lines with short ensemble intervention in between each solo lines, the level of the PM error will be raised considerably and the discrimination might be corrupted;

The ensemble mode parts have to be very characteristic as well: for example, either if the theme is played by one single instrument or by two instruments at unison, it might be considered as a solo.

This raises the problem of extracting a musical concept with low-level descriptors. It shows to work well in a lot of cases, but in reality, what is extracted is a “physical” concept of a solo (an instrument dominating the spectra) rather than a solo in a musical sense of the term. The variability of the abstract concept is too high for low-level physical features to describe it in its entirety. Higher-level descriptors and more powerful classification techniques could be used to take into account musical knowledge on solo location. For example, we know that it is statically more robust to detect ensembles than solos. Post-processing techniques could be used to correct the uncertain chunks of data with respect to these observations. Finally, recent studies were made on spectral flatness and the associated coefficient of tonality [9]. First tests showed this feature to be potentially useful in the task of locating solos [10], especially in the discrimination step. It could be added to the PM error feature in order to increase the discrimination robustness.

7 Acknowledgments

The work reported in this article has been partially funded by the IST European project CUIDADO [3].

References

[1] Agrain, P. (Guest Editor) "Musical Content Feature Extraction". Journal of New Music Research, 28 (4), 1999.

[2] Music Content Analysis web pages: http://www.cs.tut.fi/sgn/arg/music/. Tampere University of Technology, Finland.

[3] CUIDADO project. European Comission IST Program. http://www.cuidado.mu/

[4] Pachet, F. Roy, P. "Automatic Generation of Music Programs", Proceedings of the CP (Constraint Programming) Conference, Washington (USA), October 1999.

[5] Grove Dictionary of Music, http://www.grovemusic.com.

[6] Maher, R. C and Beauchamp, J. W, "Fundamental frequency estimation of musical signals using a two-way mismatch procedure", Journal of the Acoustic Society of America, Vol. 95, page 2254-2263, 1993.

[7] Foote, J. “Automatic audio segmentation using a measure of audio novelty”, Proceedings of IEEE International Conference on Multimedia and Expo, vol. I, pp. 452-455, 2000.

[8] Klapuri, A. Virtanen, T. Eronen, A. Seppänen, J. “Automatic Transcription of Musical Recordings”, http://www.cs.tut.fi/sgn/arg/klap/crac2001/crac2001.pdf, 2001.

[9] Johnsotn, J.D. 1988. “Transform Coding of Audio Signals Using Perceptual NoiseCriteria”, IEEE Journal of Selected Areas in Communication, Vol. 6, pp. 314-323, 1988.

[10] Izmirili, O. “Using a Spectral Flatness Based Feature for Audio Segmentation and Retrieval”, http://ciir.cs.umass.edu/music2000/posters/izmirli.pdf, 2000.