PITCH CONTENT VISUALIZATION TOOLS FOR MUSIC PERFORMANCE ANALYSIS

Luis Jure[1] Ernesto López[2] Martín Rocamora[12] Pablo Cancela[2] Haldo Sponton[2] Ignacio Irigaray[2]
[1]School of Music and [2]Faculty of Engineering, Universidad de la República, Uruguay
lj@eumus.edu.uy {elopez,rocamora,pcancela,haldos,irigiaray}@fing.edu.uy

ABSTRACT

This work deals with pitch content visualization tools for the analysis of music performance from audio recordings. An existing computational method for the representation of pitch contours is briefly reviewed. Its application to music analysis is exemplified with two pieces of non-notated music: a field recording of a folkloric form of polyphonic singing and a commercial recording by a noted blues musician. Both examples have vocal parts exhibiting complex pitch evolution, difficult to analyze and notate with precision using Western common music notation. By using novel time-frequency analysis techniques that improve the location of the components of a harmonic sound, the melodic content representation implemented here allows a detailed study of aspects related to pitch intonation and tuning. This in turn permits an objective measurement of essential musical characteristics that are difficult or impossible to properly evaluate by subjective perception alone, and which are often not accounted for in traditional musicological analysis. Two software tools are released that allow the practical use of the described methods.

1. INTRODUCTION

Most of the established techniques for musical analysis do not work directly on the acoustic signal, but on some kind of symbolic representation of it [1]. This representation reduces the continuous and complex sound flow into a set of discrete events, usually determined by their most salient parameters, such as temporal location, duration and pitch. Applications of spectrographic analysis of sound to the development of new techniques of musical analysis began to be explored systematically with the work by Robert Cogan [3]. Using time-frequency representations of the audio signal, Cogan proposes an analytical method applicable to both structural and local aspects of a musical piece, that exemplifies analyzing music from very varied corpus. Recently, techniques based on sonographic representation have been applied extensively to the analysis of electroacoustic music [8]. These tools are also being applied to notated music or music from traditions not based on scores, to discuss aspects of music not represented in symbolic notation by the analysis of recordings. This may include both components that depend on the performance [7] (such as temporal and tuning micro-deviations), or the precise determination of the tuning system of a certain music [6].

Different software tools for computer-aided analysis, visualization and annotation of recorded music have been developed, for instance Sonic Visualiser.[1] They typically include traditional time-frequency representations and digital signal processing tools intended for music information retrieval, such as onsets or pitch detection. Some mid-level representations are also available, i.e., signal transformations that tend to emphasize higher semantics than the energy in the time-frequency plane [4]. Those mid-level representations are usually devised to facilitate the subsequent feature extraction and processing of an automatic algorithm. However, as suggested in [5], they can also be used by humans to study performance nuances such as pitch modulations or expressive timing.

In this article, examples are given of the type of analysis that can be done with an implementation of the pitch salience representation proposed in [2] by an end-user with a musicological background (first author). In addition, two graphical software tools are released that allow the practical use of the described methods by the research community. The representation proposed, called F0gram, is based on the Fan Chirp Transform (FChT) [13] and seeks two main goals: firstly, the precise time-frequency location of the components of a complex sound, using recent analysis techniques that overcome the limitations of the classical tools; secondly, the automatic grouping of all the components that are part of the spectrum of a single harmonic source, highlighting the fundamental frequency, f_0. This makes it possible to obtain an accurate graphical representation of the temporal evolution of the melodic content of a music recording, that allows the detailed study of performance aspects related to pitch intonation and timing (e.g. tuning system, vibrato, glissando, pitch slides).

The remaining of the document is organized as follows. Sections 2 and 3 briefly describe the time-frequency analysis and the pitch salience computation respectively. Examples of performance music analysis using the released tools are provided in section 4. The paper ends with some discussion on this work and ideas for future research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.
© 2012 International Society for Music Information Retrieval.

[1] http://www.sonicvisualiser.org/

2. TIME-FREQUENCY ANALYSIS

Music audio signals often exhibit ample frequency modulation, such as the typical rapid pitch fluctuations of the singing voice. Precisely representing such modulations is a challenging problem in signal processing. It is reasonable to look for a signal analysis technique that concentrates the energy of each component in the time-frequency plane as much as possible. In this way, the representation of the temporal evolution of the spectrum is improved and the interference between sound sources is minimized, simplifying the task of higher level algorithms for estimation, detection and classification.

The standard method for time-frequency analysis is the Short Time Fourier Transform (STFT), which provides constant resolution in the time-frequency plane. A typical alternative for multi-resolution analysis is the Constant Q Transform (CQT). Both representations produce a Cartesian tiling of the time-frequency plane, as depicted in Figure 1. This may be inappropriate for non-stationary signals, for instance a frequency modulated sinusoid, namely a chirp. The virtue of the FChT is that it offers optimal resolution simultaneously for all the partials of a harmonic linear chirp, i.e. harmonically related chirps of linear frequency modulation. This is well suited for music analysis since many sounds have a harmonic structure and their frequency modulation can be approximated as linear within short time intervals.

The FChT can be formulated as [2],

$$X(f,\alpha) \triangleq \int_{-\infty}^{\infty} x(t)\ \phi'_{\alpha}(t)\ e^{-j2\pi f \phi_{\alpha}(t)} dt, \qquad (1)$$

where $\phi_{\alpha}(t) = (1 + \frac{1}{2}\,\alpha\, t)\, t$, is a time warping function. The parameter α, called the chirp rate, is the variation rate of the instantaneous frequency of the analysis chirp. Notice that by the variable change $\tau = \phi_{\alpha}(t)$, the formulation can be regarded as the Fourier Transform of a time warped version of the signal $x(t)$, which enables an efficient implementation based on the FFT. If a harmonic chirp is analyzed and the correct α value is selected for the transform, the warping yields sinusoids of constant frequency so the spectral representation is a set of very narrow peaks.

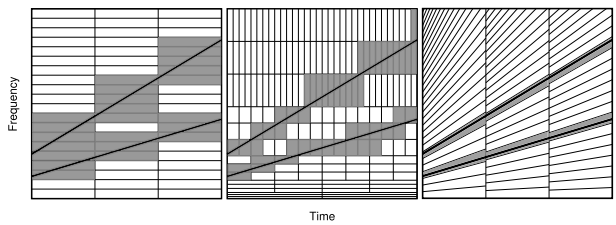

Figure 1. Time-frequency tiling sketch for the STFT, the Short Time CQT and the Short Time FChT and the resulting resolution for a two-component harmonic linear chirp.

A time-frequency representation can be built by computing the FChT for consecutive short time signal frames, namely a Short Time FChT (STFChT). This requires the determination of the optimal α value for each signal frame. For polyphonic music analysis there is no single optimal α value, so the approach followed in [2] is to compute several FChT instances with different α values. This yields a multidimensional representation made up of various time-frequency planes. The selection of the α values that produce the better representation of each sound present is performed by means of pitch salience. A comparison of the STFT and the STFChT applied to a polyphonic music audio clip is provided in Figure 2.

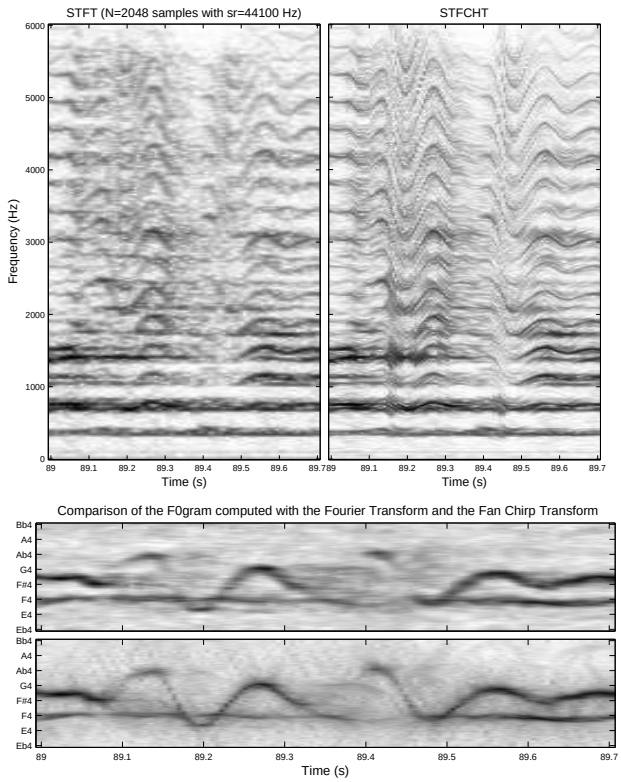

Figure 2. Above: Comparison of the STFT and SFChT for an excerpt from the example of section 4.1. The chirp rate of the most prominent sound source is selected for each frame. Note the improved representation obtained for this source while the rest is blurred. Below: F0grams obtained from the DFT and FChT. Rapid pitch fluctuations are better represented in the latter.

3. PITCH SALIENCE REPRESENTATION

A representation intended for visualizing the pitch content of polyphonic music signals should provide an indication of prominence or salience for all possible pitch values within the range of interest. A common approach for pitch salience calculation is to define a fundamental frequency grid, and compute for each frequency value a weighted sum of the partial amplitudes in a whitened spectrum [5, 13]. A method of this kind was used in [2] and is briefly described in the following.

Given the FChT of a frame $X(f,\alpha)$, salience of fundamental frequency f_0 is obtained by summing the log-

spectrum at the positions of the corresponding harmonics,

$$\rho(f_0, \alpha) = \frac{1}{n_H} \sum_{i=1}^{n_H} \log |X(if_0, \alpha)|, \qquad (2)$$

where n_H is the number of harmonics located up to a certain maximum analysis frequency. This is computed for each signal frame in a certain range of f_0 values.

Some postprocessing steps are carried out in order to attenuate spurious peaks at multiples and submultiples of the true pitches, and to balance different fundamental frequency regions [2]. Finally, for each f0 in the grid, the highest salience value is selected among the different available α values. In this way, a representation that shows the evolution of the pitch of the harmonic sounds in the audio signal is obtained, namely an F0gram. Examples of the resulting representation are depicted in Figure 3 for two short audio clips. The F0gram produces a fairly precise pitch evolution representation, contrast balanced and without spurious noticeable peaks when no harmonic sound is present. Note that simultaneous sources can be correctly represented, even in the case that they coincide in time and frequency if their pitch change rate is different. Figure 2 shows a comparison of the F0gram obtained from the DFT and the FChT. The improvement in time-frequency localization provides a more accurate representation of pitch.

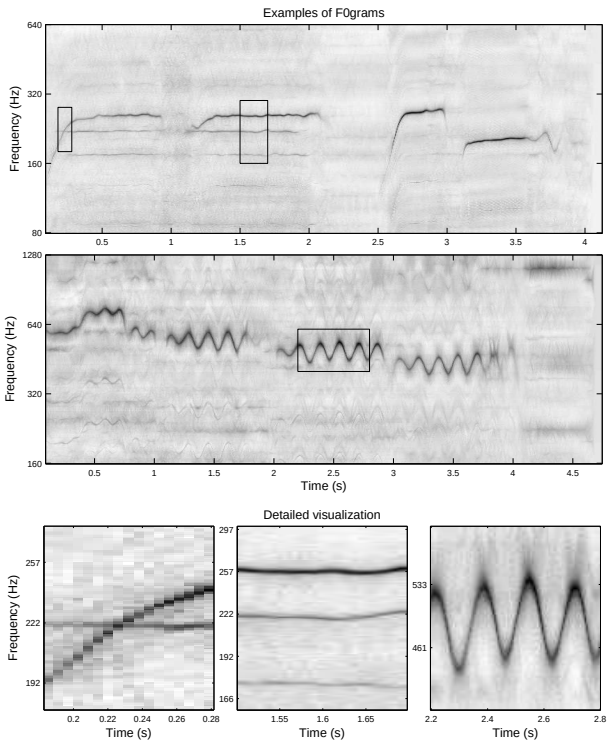

Figure 3. Above: F0gram examples for audio excerpts of *pop1.wav* and *opera_fem4.wav* from the MIREX melody extraction test set. Below: Detailed visualization. Crossing pitch contours are well resolved, and simultaneous sources and rapid pitch fluctuations are precisely represented.

4. CASE STUDIES

In order to exemplify the application of these techniques to musicological analysis, we have selected two pieces of non-notated music, both of them with vocal parts exhibiting complex pitch evolution, very difficult or downright impossible to notate with precision using Western common music notation: a field recording of a folkloric female vocal trio from west-central Bulgaria, and a commercial recording by noted blues singer and guitarist Muddy Waters.

4.1 Diaphonic chant of the Shope country

Throughout the world, folkloric forms of polyphonic singing are relatively scarce, one of the most notable exceptions being the diaphonic singing of the Shope region in west-central Bulgaria. A closely related form can be found in the Pirin region in the south-west of the country, and extending into the Republic of Macedonia.

As a general rule, these polyphonic songs are performed by female singers, and the sound itself of the voices is usually enough to impress listeners not familiar with this idiom. But the treatment of pitch in these two-part songs, or *dvuglas*, also has some unique characteristics.

In a typical setting, the melody part is sung by one singer, and the second part by two or sometimes more. The upper part has several classified melodic gestures, one of the most characteristic being a sort of—usually fast—glottal trill called *tresene*. Another characteristic gesture is the *izvikvane*, a form of ending the phrases with a fast upward leap on the vowel sound *"eee"*. The second part is more static, and has been described as a "drone" or pedal. It usually stays on the tonic of the mode, with occasional deviations to the sub-tonic when the melody descends to the tonic. Both parts join, however, to perform the *izvikvane* together. Apart from some fast swoops, the melody part moves within a very limited range, especially in the Shope region. This results in a preponderance of narrow intervals between the voices [9, 10].

For our case study, a commercially available field recording of a folkloric group from the Shope region was used [12]. The recording is identified, without further information, as a "Harvest Song" performed by a female vocal trio from the village of Zheleznitsa. The recording date can be placed around 1980. The song consists of 9 short phrases of similar duration (ca. 10∼12 s), structured as three variations of a group of three distinct phrases.

Figure 4 shows an F0gram of the third of these phrases, exhibiting all the characteristics described above: the second part begins in the sub-tonic and soon moves to the tonic for the rest of the phrase, while the first part moves both above and below the tonic, singing a more embellished melody that includes faster and slower *tresene*. The cadential *izvikvane* covers a narrow octave before descending back to the tonic area, and sounds like a unison of the three voices, in accordance with the prevailing description of *izvikvane*. The F0gram allows us to appreciate, however, that there is actually a slight separation of the voices, very difficult to perceive by listening alone.

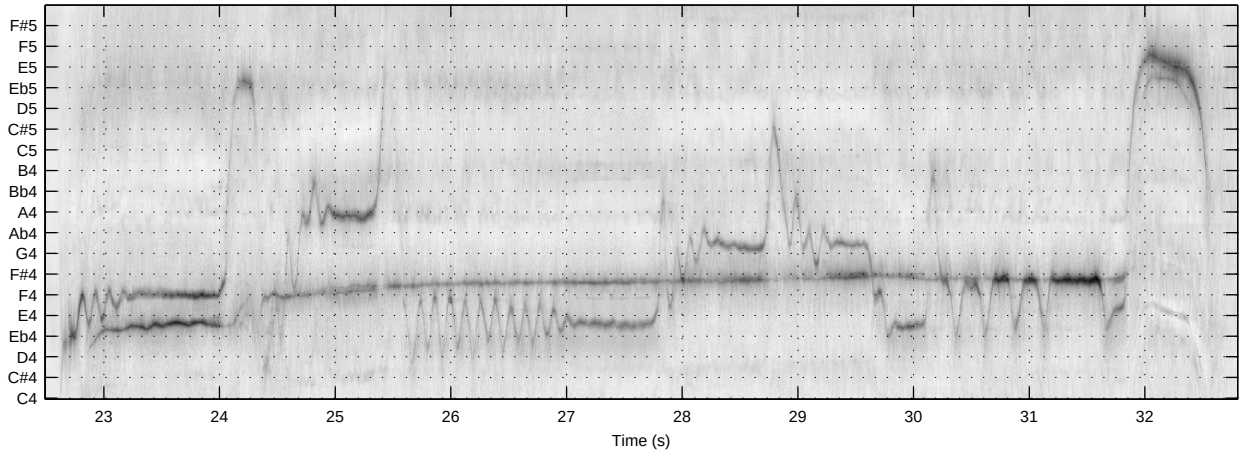

Figure 4. One phrase of the Harvest Song, showing characteristic traits of Shope diaphonic singing: melodic ornamentation in the first part (including *tresene* and a final *izvikvane*), and a pedal on the "tonic" in the second part, with a slow glissando.

An analysis of the simultaneities confirms that narrow intervals prevail, and a variety of intervals can be found between the unison and the major third. The F0gram obtained from the FChT permits a precise measurement of these type of intervals, as can be appreciated in Figure 2. Of special interest was the location of the sub-tonic, and the interval most frequently found lies half-way between one and two semitones below the tonic (sec. 23–24). This same kind of "second" can often be found above the tonic, in the upper part (sec. 28–29). The speed and range of the *tresene* can also be assessed with good precision. Typical rates are around 8~9 Hz (sec. 26–27), but slower rates can also be found (sec. 30–31). The width is variable, extending through intervals of up to three semitones (sec. 26).

So far, the analysis of the F0gram confirms—and permits a better measurement of—the characteristics described. Observing the second part with more detail, however, its behavior should be striking: instead of remaining on a fixed note, as its supposed character of "drone" would suggest, it performs a slow upward glissando, covering roughly the equivalent of a semitone (from F to F♯) during approximately 7 seconds. This displacement of the tonic not only occurs in various degrees in all the phrases throughout the song, but it also covers different pitch areas in each one (e.g., between E and F, or F♯ and G), resulting in a sort of "roving" tonic. Field recordings from different villages in the Shope region were analysed, and a similar behaviour was found in most of them, with glissandi of the "pedal" notes typically spanning between 50 cents and a semitone within a phrase. In the course of the song, the intonation of the local "tonics" can vary as much as three semitones.

The implications are two-fold and of paramount importance: unlike a typical drone or pedal point, essentially static, this second part has a dynamic character, and this kind of slowly ascending movement imposes on the polyphony a very particular tension and expressiveness. Additionally, the fact that the "tonic" varies between phrases, turns somewhat fuzzy the idea itself of modal tonic.

This phenomenon is not mentioned in the consulted bibliography and is not represented in the available transcriptions, although it was found in various degrees in several recordings analysed, suggesting that these traits conform a characteristic feature of the Shope musical idiom and should be considered an essential component of the powerful expressiveness of this particular form of folkloric polyphonic singing. These analysis techniques should be applied to a wider corpus to properly assess the importance of this performance practice.

4.2 Muddy Waters - Long Distance Call

The Blues is a genre of popular music deeply rooted in the African-American folksong tradition of the rural South of the United States, and as such it shows several traits that differ considerably from those found in the traditional European musical system. The most characteristic of these traits are the so-called *"blue notes"*, the precise definition of which has been elusive and even somewhat controversial. A simplistic but widely circulating definition reduces them to the use of the minor third and minor seventh degrees (sometimes also the diminished fifth) in a major-key context, for example, E♭ and B♭ in C major. Actually, this performance practice is much more complex, and entails two related but distinct aspects: the use of pitches that lie outside the standard Western tuning system, and continuous variations of pitch within certain tonal regions. Rather than fixed tones in a discrete scale system, blue notes would be flexible areas in the pitch space. For the analysis of the behaviour of these pitch complexes in actual performance, we chose a recording by Muddy Waters, one of the most important blues musicians of all time, regarded as an unsurpassed performer both as a guitarist *and* as a singer.

On January 23, 1951 he recorded his own composition "Long distance call" for Chess Records. He sings and plays electric guitar, and is accompanied by Marion "Little Walter" Jacobs on harmonica and Willie Dixon on double bass. The song is a standard 12-bar, three-line stanza blues, where the second line in each verse repeats the first, and the third is a rhyming conclusion. After a 4-bar introduction, the

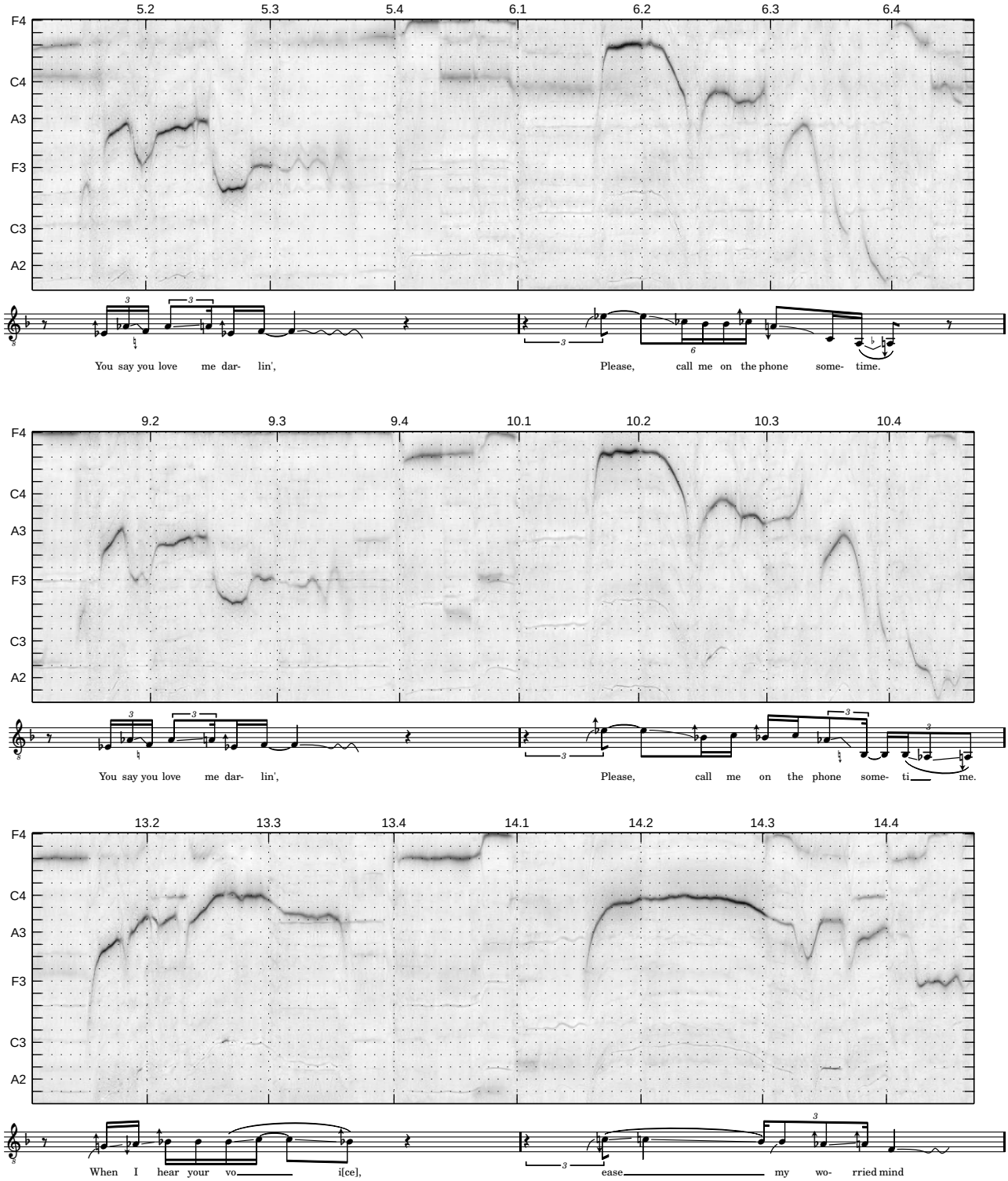

Figure 5. Muddy Waters, "Long distance call" (1951): F0gram showing continuous pitch contours of the voice, and approximate musical transcription informed by the analysis of the F0gram.

first stanza extends from measure 5 to 16. Figure 5 shows the F0gram and the transcription of the six measures where Muddy Waters sings the lyrics: mm. 5-6, 9-10 and 13-14. Each of these 2-bar vocal phrases is followed by a 2-bar instrumental response, omitted in the figure.

The musical transcription offered here is informed by the analysis of the F0gram, [2] and differs in many substantial details from published transcriptions [11], as well as from what was perceived by highly trained musicians that

[2] Just as with pitches, Muddy Waters' treatment of durations is equally flexible. The note values chosen for the transcription are approximate, and the vertical alignment with the F0gram is not always perfect.

were asked to listen to the recording. Observing the F0gram it is easy to see why: the melody consists mostly of continuously varying pitches, with few moments of stability other than the resolution on the tonic, and these exhibit a wide terminal vibrato. In this context, the perception of definite notes requires a decision on the part of the listener, that is partly subjective. For example, the first three notes ("You say you") are normally perceived as F-A♭-F, but a closer inspection reveal that the first note is actually a fast "scoop" around a slightly high E♭, and the second a continuous glide from below A♭ to around A♮. This behavior is consistent when the phrase is repeated (m. 9). A similar treatment of the third as a blue note (i.e., as a flexible pitch area) can be observed on the word "phone" on mm. 6 and 10. The previous words ("call me on the"), also move within a continuous pitch region, this time around the 4th and 5th degrees (B♭ and C). The long notes that begin the second half of each line ("*plea*-se" around E♭ on the first and second line, and "*ea*-se" around C on the third) exhibit all the same arch-like melodic contour, with wider and faster ascending and descending movements at the beginning and the end of the note, and a slow curve during the sustain part. A particularly expressive effect results from the fact that the C is hardly reached for an instant at the peak of the arch, the rest of the time the melody is kept moving slowly around a somewhat flat fifth. The F0gram also shows, through different shades in the grayscale, that the dynamics of the phrase follow a similar arch-like contour. The most ambiguous moments in terms of pitch are the endings of the first and second lines ("sometime"), with fast portamentos into the lower register that give these passages a speech-like quality.

In many Western vocal practices, continuous inflections of pitch are common when connecting the different —stable—notes of a melody (*legato* singing), as well as in the form of vibrato, fluctuations of pitch around a central perceived note. The application of the analysis tools proposed here permits a clear visualization of two salient traits of this passage: a melody consisting mostly of time varying pitches with relatively few moments of stability, and the establishment of continuous tonal regions non reducible to single pitches in a discrete scale.

5. DISCUSSION AND FUTURE WORK

By means of the analysis of two music recordings, the usefulness of the introduced techniques for computer aided musicology was illustrated, in particular for discussing expressive performance nuances related to pitch intonation. The result of the analysis by itself reveals important aspects of the music at hand, difficult to asses otherwise. The computational techniques implemented are oriented towards the precise representation of pitch fluctuations.

Two graphical software tools are released with this work to allow the application of the described methods by the research community.[3] One of them is a Vamp plugin[4] for Sonic Visualiser that computes the pitch contours representation. Within this application several other features are available that can assist the analysis. A Matlab® GUI is also released that includes additional functionalities and information, better suited for signal-processing researchers.

The improvement of the pitch contours representation and its application to different music scenarios are the following directions for future research.

[3] Available, as well as the audio clips, from `http://iie.fing.edu.uy/investigacion/grupos/gpa/ismir2012`

[4] `http://www.vamp-plugins.org/`

6. ACKNOWLEDGMENTS

This work was supported by the R+D Program of the Comisión Sectorial de Investigación Científica (CSIC), Universidad de la República, Uruguay.

7. REFERENCES

[1] I. Bent and A. Pople. "Analysis". *Grove Music Online*. Accessed June 16, 2012.

[2] P. Cancela, E. López, and M. Rocamora. Fan chirp transform for music representation. In *13th Int. Conf. on Digital Audio Effects, Austria*, Sep. 2010.

[3] R. Cogan. *New images of musical sound*. Harvard University Press. Cambridge, Massachusetts, 1985.

[4] J. Durrieu, B. David, and G. Richard. A musically motivated mid-level representation for pitch estimation and musical audio source separation. *Selected Topics in Signal Processing, IEEE Journal of*, 5(6):1180 –1191, Oct. 2011.

[5] A. Klapuri. A method for visualizing the pitch content of polyphonic music signals. In *10th Int. Society for Music Information Retrieval Conf., Japan*, 2009.

[6] A. Krishnaswamy. Pitch measurements versus perception of south indian classical music. In *Proc. of the Stockholm Music Acoustics Conf., Sweden, Aug.*, 2003.

[7] D. Leech-Wilkinson. *The Changing Sound of Music: Approaches to Studying Recorded Musical Performance*. Published online, London: CHARM, 2009.

[8] T. Licata, editor. *Electroacoustic Music - Analytical Perspectives*. Greenwood Press, 2002.

[9] L. Litova-Nikolova. *Bulgarian folk music*. Bulgarian academic monographs. Marin Drinov Academic Pub. House, 2004.

[10] S. Petrov, M. Manolova, and D. Buchanan. "Bulgaria". *Grove Music Online*. Accessed April 5, 2012.

[11] F. Sokolow and D. Rubin. *Muddy Waters - Deep Blues*. Hal Leonard Corporation, 1995.

[12] Musics & musicians of the world: Bulgaria. AUVIDIS/UNESCO, 1983.

[13] L. Weruaga and M. Képesi. The fan-chirp transform for nonstationary harmonic signals. *Signal Processing*, 87(6):1504–1522, 2007.