Computational model of pitch detection, perceptive foundations, and application to Norwegian fiddle music

Olivier Lartillot,[1] Hans-Hinrich Thedens,[2] and Alexander Refsum Jensenius[3]

[1,3] *RITMO Centre for Interdisciplinary Studies in Rhythm, Time and Motion, University of Oslo, Norway*

[2] *National Library, Norway*

[1]olivier.lartillot@imv.uio.no

Abstract

Automated detection of pitch in polyphonic music remains a difficult challenge. Implementation of perceptive/cognitive models have been so far less successful than engineering methods.

We present a model that is neither based on a machine-learning training on a given set of samples, nor explicitly relying on stylistic rules. Instead, the methodology consists in conceiving a set of rules as simple and general as possible while offering satisfying results for the chosen corpus of music. We present a new method for harmonic summation that penalizes harmonic series that are sparse, in particular when odd partials are absent, as it would indicate that the actual harmonic series is a multiple of the given pitch candidate. Besides, a multiple of a fundamental can be selected as pitch in addition to the fundamental itself if its attack phase is sufficiently distinctive. For that purpose, we introduce a concept of pitch percept that persists over the whole extent of the note, and that serves as a reference for the detection of higher pitches at harmonic intervals.

The proposed method enables to obtain transcriptions of relatively good quality, with a low ratio of false positives and false negatives. The construction of the model is under refinement. We are applying this method to the analysis of recordings of Norwegian folk music, containing a large part of Hardanger fiddle pieces and a cappella singing.

By attempting to design computer models based on general rules as simple as possible rather than on machine learning, while resulting in a behavior in terms of pitch detection that comes closer to human capabilities, we hypothesize that the underlying mechanisms thus modelled might suggest general computational capabilities that could be found in cognitive models as well. In the same time, an improvement of the model based on expertise in music perception and cognition is desired.

Cognitive models of pitch perception

Cognitive modelling of pitch perception remains an open and controversial issue (McDermott & Oxenham, 2008). The perception of the fundamental frequency (or F0) of a given pitch is explained using various competing–and complementing–mechanisms. The cochlea operates a spatial (or tonotopic) decomposition of sound along frequencies at each successive instant. But at the same time a precise pitch estimation requires the study of the time waveform for each individual critical frequency in that tonotopic decomposition. Because each critical frequency is actually the center of a frequency band–with increasing bandwidth as frequency increases–the first harmonics of a given pitch will be properly resolved while higher harmonics will interact with each other in frequency bands and will not be resolved. Those mechanisms have not been fully understood yet (Oxenham, 2012). Another mechanism possibly helping pitch perception relates to the phase locking of individual auditory nerves. There is evidence for cortical neurons beyond primary auditory cortex that are tuned to pitch (Bendor & Wang, 2006; McDermott & Oxenham, 2008).

Extensive studies have attempted to understand how multiple pitches could be perceptually combined to form chords and voices. The knowledge brought by these works is not sufficiently detailed to be directly translated into computational models.

In fact, the mere core mechanism of single note perception does not seem to be sufficiently understood in a cognitive point of view to allow computational modelling. For the simple case of individual sound event, this could be conceived as a simple detection based on detection of attack, onset and decay in the temporal representation of the energy of the signal. But when considering complex polyphonies, the notes detection seems to depend mainly on pitch perception, probably through a tracking of F0 over time and a detection of attack and decay along each F0 separately.

Computational models of pitch extraction

Computational approaches for pitch extraction can be decomposed into several types.

Machine learning approaches

The most dominant type of approach currently giving the best results is based on machine learning: the learning system is trained on particular audio recordings for which the corresponding transcriptions are given as well. “Learning”, here, means that the program automatically optimizes the inference algorithm so that it predicts the correct transcription when given as input its corresponding audio recording. This approach is dependent on the transcriptions provided during the learning phase. The resulting algorithm does not generalize well on audio examples that have very different musical characteristics. If we consider for instance the transcription of fiddle music, a proper training would require to provide as learning examples detailed transcriptions. And if we aim to get detailed transcriptions showing the pitch fluctuation within each note, we would have to provide such example transcriptions beforehand.

Nowadays machine learning is implemented using neural networks and notably using deep learning techniques. Despite the name, “neural networks” are not supposed here to provide an actual cognitive modelling of pitch perception, since humans generally do not learn to perceive pitch based on supervised training.

Template-based models

Another method consists in recording individual notes played along the whole range of pitches under consideration and for various types of instruments. Those notes are then retrieved automatically on the audio recording to be analyzed, through a mathematical decomposition. Evidently, this engineering approach is not supposed to mimic human cognition. One major practical limitation is that the approach will not work properly for instruments with very different timbre that those prerecorded or with larger pitch range.

Cognitive computational models

The aforementioned cognitive models of pitch perception have been translated into detailed computational models (Medis & Hewitt, 1991). It has also been shown that some simplification can be carried out without degrading the general quality of the model (at least in the particular domain of application under consideration), while allowing to solve complex problems that the more detailed models were not able to tackle yet. The particular step of periodicity analysis of the critical bands, was not much explained in music cognition. In many reference works, this analysis has been modeled using autocorrelation function. But alternative models such as comb filter have been proposed as well. One of the most advanced computational models for pitch extraction based on cognitive theories has been proposed by Klapuri (2006a). Because pitch perception models so far focus mainly on single F0 extraction, additional engineering-based methods are developed to enable multipitch extraction, mainly based on time-domain cancellation.

Other approaches

Klapuri (2006b) has proposed another model, which offers significant improvements and has become a reference in Music Information Retrieval (MIR). Interestingly enough, this model does not follow cognitive theories as closely as before (such as tonotopic decomposition followed by autocorrelation function, as discussed above) but instead develop engineering-based strategies. This seems to indicate that current cognitive understanding of pitch perception is not mature enough to be directly implemented into a state of the art computational model. In many recent approaches, periodicity analysis is based on frequency *spectrum* representation, computed for instance using the *Fourier Transform*. This is the case in the subsequent work by Klapuri (2006b). Moreover, this approach is funded on the concept of *harmonic summation*, that we will discuss further in the next section.

Most current MIR approaches for pitch detection, including Klapuri's, search for F0s for each successive instant (or time frame more precisely) independently. The F0s detected frame by frame are then tracked over time in a second step, in order to form pitch contour based on time and frequency continuity, using heuristics based on auditory streaming cues or additional musical knowledge.

Still an open problem

Despite the significant advance in multipitch estimation, this task remains one of the main challenges in the MIR field, which needs to address many difficulties, such as masking, overlapping tones, mixture of harmonic and non-harmonic sources and the fact that the number of sources might be unknown (Schedl, Gómez & Urbano, 2014). Even on simple polyphonies, the performance obtained by multipitch estimation methods reaches moderate note accuracy for relatively simple music material, such as quartet, woodwind quintet recordings, and rendered MIDI, with a maximum polyphony of 5 notes (MIREX, 2016).

One way to compensate the limitations of current approaches in multi-pitch extraction is through the addition of music language models, which represent sequences of notes and other music cues based on knowledge from music theory or from constraints automatically derived from symbolic music data. Such approach would however not generalize well to various kinds of music genres and cultures, unless the models are updated accordingly. Besides, in a cognitive point of view, if listeners are able to detect the pitches without necessarily knowing that particular genre of music, we would surmise that a computational model should work independently on the stylistic rules.

Proposed model

Spectrum representation

Similar to many MIR methods, we first represent the audio signal in the frequency domain (or spectrum) using the Fourier Transform. By decomposing the audio signal into short parts (*frames*) and computing the spectrum representation for each frame successive, we obtain a bi-dimensional diagram, where the horizontal axis is the temporal evolution of the audio signal, and the vertical axis the different frequencies found in each frame. An example is shown in Figure 1. This spatial representation is similar (in terms of axes dimensions) to the representation of the pitch curves shown in Figure 2.

This spectral representation can be somewhat related to the basic principles of tonotopic decomposition, that is, decomposing the energy into frequencies. Another perceptive aspect that needs to be included in the spectrum representation is the frequency filtering operated by the outer ear, with an emphasis on frequencies around 3000 Hz (Terhardt, 1979). Figure 1 shows the result of this filtering, with most of the energy near 3000 Hz. It turns out that this frequency emphasis significantly improved the pitch detection performed in the subsequent steps, for instance when analyzing fiddle music (cf. next section). This illustrates the fact that the acoustic of fiddle instrument is tuned to human ears, for instance with respect to pitch clarity.

Harmonic summation

When representing a given pitch in the frequency domain, there is a peak of the frequency F0 of the fundamental and peaks as well at harmonic partials, whose frequencies are multiples of F0. A convenient method to detect F0s from the frequency domain is called *harmonic summation*: it consists in associating a score to each F0 by summing together the spectrum amplitudes at multiples of F0. For instance, for the frequency F0 = 440 Hz, we sum the spectrum amplitudes related to the frequencies 440, 880, 1320 Hz and so on. The highest scores would correspond to the F0s of the actual pitches in the signal.

In previous works, the score associated to each F0 candidate is computed through a simple summation of each partial magnitude. Klapuri (2006b) also follows this strategy, although using weighted summation.

In our view, a simple summation, even weighted, does not sufficiently describe the way harmonic sequences of partials are perceived. We have developed a new strategy, where instead of considering each partial p_i individually, we consider each partial in relation with its previous partial p_{i-1} or its two previous partials p_{i-1} and p_{i-2}. If p_i is an even harmonic, its contribution to the total score is computed by multiplying the spectrum magnitude at p_i and p_{i-1}. If p_i is an odd harmonic, its contribution is computed by take the maximum between (1) multiplying the spectrum magnitude at p_i and p_{i-1} and (2) multiplying the spectrum magnitude at p_i and p_{i-2}. For instance, a harmonic sequence where all odd harmonics of F0 are absent will give a score of 0, indicating that the actual pitch should rather be at 2*F0. This method penalizes harmonic series that are sparse, in particular when odd partials are absent, as it would indicate that the actual harmonic series is a multiple of the given pitch candidate.

Pitch detection based on dynamic evolution of partials

One major difficulty when detecting multiple pitches is that, mathematically speaking, any harmonic Fi of a given F0 present in the signal could itself appear as a possible F0: its own harmonic series is a subset of the harmonic series of the lower F0. For instance, the pitch F1 at an octave above a given pitch F0 has an harmonic series that corresponds to the even partials of the harmonic series of F0. Evidently, all harmonics Fi of a given F0 are not themselves perceived as individual pitches.

We may hypothesize that for a given harmonic Fi to be perceived as the fundamental of an additional pitch, superposed to the pitch at F0, its harmonic series being added to the harmonic series of F0, in the resulting harmonic series, the peaks corresponding to the series related to Fi should be higher. Klapuri (2006b) uses this hypothesis, and develops a method where each time a given F0 is found, its harmonic series is removed from the signal so that eventual harmonic series corresponding to other pitch at Fi can be detected as well. We also follow this hypothesis in our model, but instead of removing harmonic series for each pitch found in the signal, we search for predominant subseries in each given harmonic series.

In our experiments, we came to the conclusion that this hypothesis is too strong and does not explain all conditions for the appearance of pitches at harmonic intervals above other pitches. We have found another more subtle characterization: a pitch Fi can appear at such harmonic interval above F0 if the harmonic series starting from Fi increases over time. So even if the harmonic series of Fi is not particularly dominant compared to the series starting from F0, if there is a significant increase over time of some of its components, this suggests the detection of pitch Fi.

Adopting this strategy requires to rethink the way pitches are detected. As aforementioned, in most approaches in MIR, pitches are detected for each successive frame separately; they are then combined into notes by tracking the F0s values over time in order to form pitch contour based on time and frequency continuity. The decision concerning the selection of F0s is made in the first step, for each successive frame separately.

In our proposed model, F0s are still searched for in each successive frame, but once a F0 has been detected, a new pitch contour is created and further tracked in the subsequence frames. The pitch contour stores the dynamic evolution of the magnitude of each partial. This information can therefore been used to detect any pitch Fi at harmonic interval.

The proposed model will be presented in more details in an upcoming paper and will be integrated into MIRtoolbox (Lartillot & Toiviainen, 2007) and the MiningSuite.

Analysis of Hardanger fiddle music

This approach has been specifically developed for the analysis of traditional Norwegian folk music played on the Hardanger fiddle. The Hardanger fiddle is slightly smaller than a regular violin, with a shorter neck and a flatter bridge, which allows to play more than two strings at the same time. In addition to the bowed strings, there are four or five sympathetic strings that run under the board. The sympathetic strings resonate when the bowed strings are played, contributing to the rich sound of this fiddle and also giving the music its characteristic drone. The fiddle playing is also extensively ornamented (Haugen, 2016).

In the current state of this research, the algorithm has been tested on traditional Norwegian folk dances. Figure 2 shows the beginning of the analysis of a performance of the tune called *Gibøens bruremarsj*.

The first results of this new model are promising. We can generally detect all the notes, even if they are repeated very quickly and played on several strings at the same time. We can also get a detailed dynamic envelope of each note, that could be used for further research about rhythm and attacks. The algorithm also output false positives and false negatives sometimes.

Discussion

As we have seen in the study of the state of the art in computational models for pitch extraction, the implementation of perceptual models describing as closely as possible the way pitch is perceived has not lead so far to particularly successful solutions for polyphonic music transcription. It is telling in this respect that the author of the most detailed computational implementation of psychological modelling of multi-pitch extraction has turned later to a more engineering approach. His engineering approach proved more successful, as it became a reference model in MIR. It seems that the cochlear model and in particular the filterbank decomposition proves somewhat problematic, as it leads to a significant distortion of the audio signal without clear advantages, compared to more simple engineering approaches, such as the use of Fourier transforms. We hope new advances in cognitive understanding (or works already published unknown to us) will offer new guidance for computational improvements.

Nonetheless the strategies developed in computational approaches might suggest hypotheses concerning the cognitive modelling of pitch perception. In particular if a simple computational mechanism is shown to offer particularly good results in a large range of music styles, we may suppose that similar operations are performed in the auditory system.

In that respect, we might wonder whether (or not) the concept of harmonic summation might have some cognitive validity. It seems computationally more simple and effective than selecting individual frequencies in the spectrum. One main objection would be that harmonic summation would not work in the same way for inharmonic sounds. However inharmonic sounds are usually associated with attacked sound, whose pitch might be detected using complementary mechanisms taking benefit of the attack phase.

One particularity of our proposed model is that it introduces a concept of pitch percept that persists over the whole extent of the note, and that serves as a reference for the detection of higher pitches at harmonic intervals. The perceptual and cognitive implications of this notion of pitch percept need to be investigated.

Acknowledgements. This work was carried out in the context of the innovation project *SoundTracer* supported by the University of Oslo and developed in collaboration with the National Library of Norway.

References

Bendor, D., & Wang, X. Cortical representations of pitch in monkeys and humans. Current Opinion in Neurobiology 2006;16:391–399.

Haugen, M. R. Investigating Periodic Body Motions as a Tacit Reference Structure in Norwegian Telespringar Performance. Empirical Musicology Review 2016;11:272–294.

Klapuri, A. (2006a). Auditory Model-Based Methods for Multiple Fundamental Frequency Estimation. In A. Klapuri & M. Davy (Eds.), *Signal processing methods for music transcription* (pp. 229-265). Springer.

Klapuri, A. (2006b). Multiple Fundamental Frequency Estimation by Summing Harmonic Amplitudes. In *Proceedings of the Seventh International Conference on Music Information Retrieval*. US: University of Victoria.

Lartillot, O., & Toiviainen, P. A Matlab Toolbox for Musical Feature Extraction From Audio. In *Proceedings of the International Conference on Digital Audio Effects*. France: Bordeaux, 2007.

McDermott, J. H., & Oxenham, A. J. Music perception, pitch, and the auditory system. Current Opinion in Neurobiology 2008;18:452–463.

Medis, R., & Hewitt, M. J. Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. The Journal of the Acoustical Society of America 1991;89:2866–2882.

MIREX 2016: Multiple Fundamental Frequency Estimation & Tracking Results - MIREX Dataset http://www.music-ir.org/mirex/wiki/2016:Multiple_Fundamental_Frequency_Estimation_%26_Tracking_Results_-_MIREX_Dataset

Oxenham, A. J. Pitch perception. Journal of Neurosciences 2012;32:13335–13338.

Schedl, M, Gómez, E., & Urbano, J. Music Information Retrieval: Recent Developments and Applications. Foundations and Trends in Information Retrieval 2014;8:127–261.

Terhardt, E. Calculating virtual pitch. Hearing Research, 1979;1:155–182.

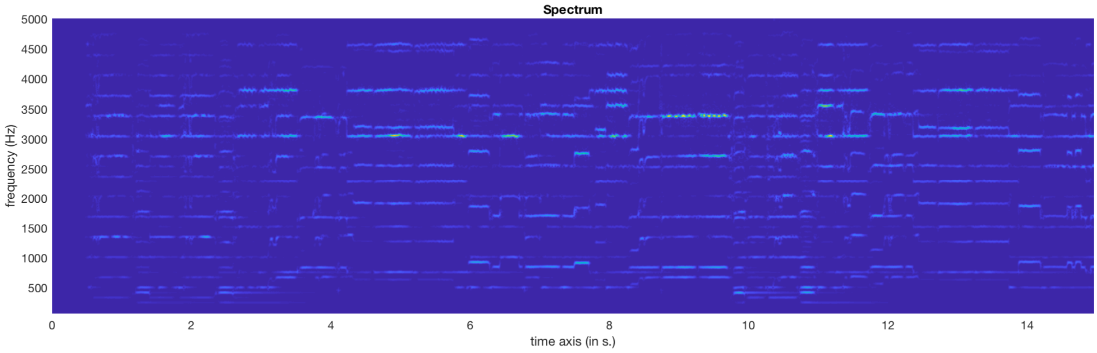

Figure 1. Frequency spectrum decomposition of the beginning of a performance of the tune called *Gibøens bruremarsj* played on Hardanger fiddle. Time in second is shown on the horizontal axis. Frequencies are decomposed along the vertical axis. The higher the magnitude, the brighter the colour.

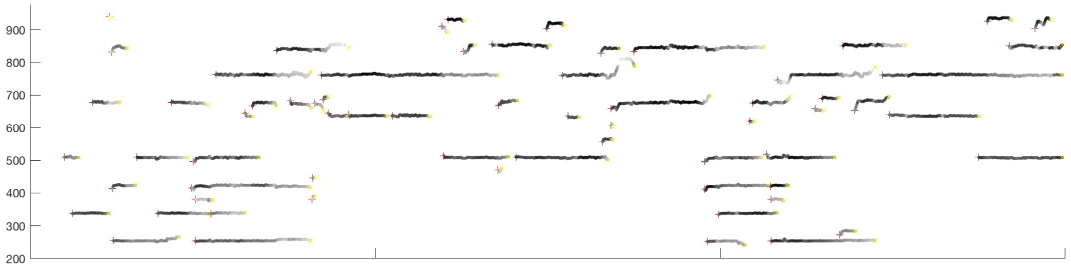

Figure 2. Pitch extracted from the audio recording shown in Figure 1. Each note is represented by a black line, starting with a red cross and ending with a yellow cross. Time is shown on the horizontal axis and frequencies on the vertical axis.