MELODY PERCEPTION AND EXTRACTION FROM AUDIO

Andrea Cogliati
University of Rochester
Dept. of Electrical and Computer Engineering
andrea.cogliati@rochester.edu

ABSTRACT

Human beings have a very sophisticated sense of hearing. While the physiological aspects of the auditory systems are well established, the perceptual and cognitive aspects are still not well understood. One active field of research in computer audition is automatic melody extraction from audio. The applications range from query-by-humming to genre classification and cover detection. On the cognitive side, the research in melody perception has not been very active in the past 15 years, the main reason being an intrinsic difficulty in designing and conducting experiments. To understand how humans perceive melodies we have to understand the mechanism for pitch perception, stream segregation and melody perception. Only the first two tasks have been researched extensively while most researches on melodies have been relegated to simple unaccompanied melodies.

In this paper I describe the state-of-the-art of melody perception and extraction from both a cognitive and a computational points of view. Some of the proposed algorithms for melody extraction model the human hearing to a certain degree and exploit perceptual cues to improve accuracy, while other use signal processing and machine learning techniques with little or no regard to physiological or cognitive models. Proposed algorithms have achieved good results in some circumstances but they are far from being perfect. From our analysis it appears that perceptually motivated algorithms can improve the accuracy but future research in signal processing might find alternative ways to solve the problem.

1. INTRODUCTION

The field of music perception and its subfield of melody perception are relatively young areas of research, especially in comparison to related research fields like vision, memory or even speech recognition. While the human auditory system has been been extensively studied, we can only understand a small part of it, namely from the outer ear to the auditory nerve [23].

The brain mechanisms for melodic processing are still not very well understood [31]. Established and novel non-invasive diagnostic techniques, like functional magnetic resonance imaging (fMRI), even-related potential (ERP), electroencephalography (EEG), and positron emission tomography (PET), can be used to observe and analyze how the human brain reacts to musical stimuli. Unfortunately there are two main challenges to understanding how music is perceived and processed in the human brain: the first problem is the overwhelming complexity of the human brain, the second is the complexity of music itself. Typical experiments in psychoacoustics are conducted with simple signals, like sequences of pure sinusoidal tones. The rationale for this is for the experimenters to be able to precisely control each variable involved in the experiment, i.e., pitch, loudness, rhythm and tempo, independently from other variables [14]. These very crude experiments can only shed some light on how real music is perceived, though. The perception of melody in a complex texture, like in a song or in an orchestral passage, is an almost entirely uncharted territory. This is also reflected in automatic methods for pitch recognition. The detection of pitches from a single source is a solved problem but multi-pitch detection and streaming are still open problems. [8] One major challenge for future research is thus how to construct ecologically valid experiments, that is using real music as stimuli for experiment rather than dull laboratory music, while retaining a good control on the involved variables [14].

This paper is divided in two major sections: section 2 will describe the current understanding on how humans perceive melodies and what perceptual cues can be involved in the process; section 3 will describe different melody extraction algorithms showing different approaches to the problem.

2. MELODY PERCEPTION

According to the Oxford Dictionary of Music, a melody is “a succession of notes, varying in pitch, which have an organized and recognizable shape.” [15] Other definitions include a positive connotation, like being pleasant or musically satisfying for the listener. Cook provides a more precise definition on how the shape of a melody is recognizable: a melody maintains its identity over certain musical transformations, like starting pitch, tempo, timbre, loudness and rhythm [5]. As in vision, transformations of a shape can be tricky: a chair is still a chair even if the

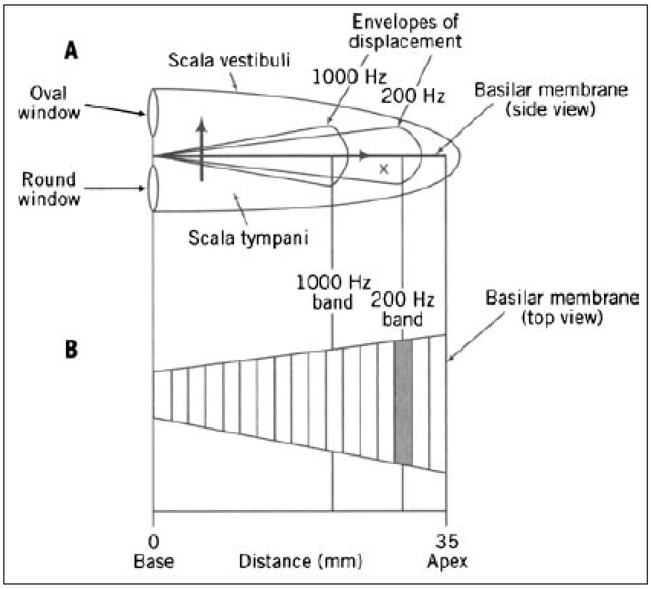

Figure 1. Pictorial representation of activity along the basilar membrane [10]

legs are shortened or lengthened slightly; but if the legs are lengthened too much a chair might become a stool. The same happens with melodies, especially with tempo and rhythm transformations.

The problem of melody perception in humans is twofold: the first aspect is the perception of melody in itself, the second aspect is melody perception in an ensemble, e.g. the melody of a song. The second problem is essentially a segregation problem with the additional constraint of identifying the most important part [2,24].

A popular hypothesis among researchers is that melody perception, similarly to vision, is governed by a few Gestalt principles, like the law of proximity and the law of continuity [2, 14]. One important open question though is "how does the brain recognize musical Gestalts?" [18] Also, Gestalt principles do not explain other important perceptual features like tonality and expectation [7, 14, 29]. One interesting problem is whether the tonal hierarchy and thus musical expectation is innate or learned. At this time the research is inconclusive. Some experiments suggest that melody perception is a learned ability [27] while other researches indicate that there is also a strong innate component to it [5,14,21].

2.1 Pitch Perception

Pitch is a perceptual measure strongly correlated to the fundamental frequency F0 of a stimulus [5, 6, 10, 14]. There are two approaches to estimate the F0: the first one involves the spectrum, the second one the waveform. Physiological experiments show that the cochlea performs a frequency analysis of auditory stimuli, see Figure 1, thus suggesting the spectral model [6, 10, 23]. On the other side, some early psychoacoustical experiments involving the missing fundamental can not be explained by the spectral model̃citemoore,cheveigne, thus suggesting temporal mechanisms involving autocorrelation. Meddis et al. proposed a unified model of pitch perception taking into account both spectral and temporal information [22], see Figure 2. It is reasonable to assume that humans use both mechanisms, even though the latter is not well understood yet [5,6,10,14].

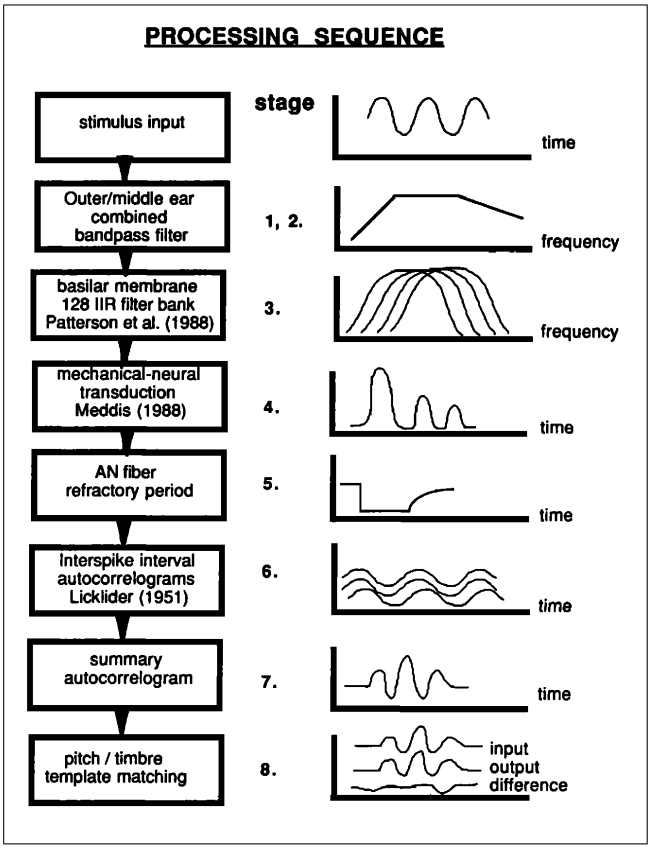
PROCESSING SEQUENCE
stage
stimulus input
time
Outer/middle ear combined bandpass filter
1, 2.
frequency
basilar membrane 128 IIR filter bank Patterson et al. (1988)
3.
frequency
mechanical-neural transduction Meddis (1988)
4.
time
AN fiber refractory period
5.
time
Interspike interval autocorrelograms Licklider (1951)
6.
time
summary autocorrelogram
7.
time
pitch / timbre template matching
8.
input
output
difference

Figure 2. Schematic outline of pitch model by Meddis et al. [22]

2.2 Perceptual cues for segregation and identification

Auditory stream segregation is an important component of melody perception. Human listeners can almost effortlessly identify sequences of notes played by different instruments. That requires to segregate concurrent stimuli coming from different sources and aggregate successive stimuli coming from the same source. The most important perceptual cue for segregation appears to be timbre [5, 13, 14]. See the next section for more details.

Also important to melody perception and identification appear to be asynchrony, i.e. playing the melody note slightly earlier than other notes, and loudness [9].

Many melody detection algorithms use an equal loudness curve filtering at the initial stage. Humans are more sensitive to sounds at particular frequencies and less sensitive at other frequencies. Perceptual experiments have been conducted to establish the equal loudness curve, i.e. a curve in the frequency and sound pressure level space that is perceived with constant loudness at different frequencies of the audible spectrum. Since loudness is an important perceptual cue for melody detection, making sure to analyze the perceived loudness and not just the energy at a given frequency is critical for estimating the salience of a particular pitch [5].

Finally, masking can affect perception of notes and thus melody identification. An established model of the basilar membrane and auditory nerves assumes the existence of critical bands or auditory filters. Simultaneous stimuli with frequencies inside the same critical band interact and, under certain circumstances, one sound can mask, i.e., cancel, the other [5,23]. While it is hardly arguable that music composers exploit the masking effect intentionally when composing music, masking plays a definite role in listening and perception. One of the methods discussed in section 3 exploits masking to improve the accuracy of the melody identification.

2.3 Stream segregation by timbre

Since the most significant perceptual cue for segregation of auditory streams appears to be timbre, it is interesting to review the status of research. Unfortunately, very little success has been attained in the recent years, probably because of the intrinsic difficulty in realizing ecologically valid experiments, as previously mentioned. One of the most significative results are still those presented by Iverson in 1995 [13].

Previous works established that onset transients play a fundamental role in instrument identification. Identification accuracy decreases when onsets are removed, while accuracy is very high when only transients are presented. Nevertheless, Iverson et al. also discovered that dynamics attributes present throughout tones are also responsible for similarity judgements [12]. In his 1995 paper, Iverson claims that both identification and similarity play a role in auditory segregation, thus concluding that onsets and sustained regions of tones must be analyzed.

Iverson used sequences of equally loud tones produced by orchestral instruments in an experiment similar to Bregman's seminal experiment on segregation [2]. Two set of sequences were presented to the participants: a physically isochronous sequence and a perceptually isochronous sequence. The former was comprised of the first 130 ms of a tone followed by a 10-ms decay to silence. The latter was created estimating the Perceptual Attack Time (PAT) for different instruments. The reason for this is that the onset of a note is perceived at a later time respect to when a player starts playing a note. Intuitively, percussive instruments have a faster attack, while woodwinds and brass have slower ones. The difference can be as low as 1.6 ms for a vibraphone and as high as 52.6 ms in the case of a clarinet. Iverson used Multidimensional Scaling Analysis to analyze the results.

The main result of the experiments was that judgements of streaming are highly correlated with similarity judgments, i.e. it is easier to segregate dissimilar instruments. Shorter PATs segregate more respect to longer PATs, and instruments with dissimilar PATs also segregate more.

Finally, the author discusses the challenges in interpreting the results because "natural musical instrument tones had many uncontrolled acoustic factors." I would add that even if we know the importance of timbre in stream segregation, we still do not know how to exploit timbre in machine extraction algorithms when multiple auditory streams are mixed together. The difficulty of controlling acoustic features of natural musical instruments might be partially overcome using sampled libraries. Modern virtual instruments closely reproduce their natural counterparts, so more realistically sounding experiments can be designed while maintaining control over experiment variables.

3. MELODY EXTRACTION

MIREX 2013 defines the audio melody extraction task as "to identify the melody pitch contour from polyphonic musical audio. Pitch is expressed as the fundamental frequency of the main melodic voice, and is reported in a frame-based manner on an evenly-spaced time-grid. [1] " The task is divided in two sub-tasks: voicing detection, i.e. deciding whether a particular frame contains a melody pitch or not, and pitch detection, i.e. selecting the most likely melody pitch in a voiced frame. The evaluation of the methods is based on the two sub-tasks. The Voicing Detection is the accuracy of detecting voiced frames. The Raw Pitch Accuracy is the accuracy of detecting the right pitch (within ± 1/4 tone) in voiced frames. The Overall Accuracy combines both voicing detection and pitch estimation. Table 1 shows the results of the methods discussed in this section. Starting from 2006, MIREX introduced multiple datasets for evaluating the audio melody extraction task. Most algorithms are very sensitive to the dataset so the table shows the minimum and maximum accuracy.

Several methods for automatic melody extractions have been proposed in the past 15 years, and many of them have been submitted to MIREX for evaluation. The biggest group is comprised of salience-based methods. They start with a spectral representation of the audio signal, then they compute a time-frequency representation of pitch salience. The peaks of this salience function are potential F0 candidates for the melody. Finally, the best F0 in each frame is selected based on some tracking model. The two most important components of these methods are the salience function and the pitch selection function. Some methods use a model of the cochlea to detect perceptual dominant fundamental frequencies [24, 26], other methods use harmonic summation with weights learned from instrument training data [16], while others let different F0s compete for harmonics using Expectation–Maximization algorithm to estimate latent harmonic components [11].

Some algorithms are limited to detect and extract sung melodies, i.e. extract the vocal line in a song while ignoring melodies or solo parts played by instruments. These methods exploit some features specific to the human voice, like the range, the timbre and the presence of vibrato and tremolo [4].

Finally, several other methods have been proposed. Some approaches, like Poliner's [25], while providing good results overall have not been further explored since their ap-

[1] http://www.music-ir.org/mirex/wiki/2013:Audio_Melody_Extraction

pearance. Others, like Arora's [1] are very new and promising and it is arguable that they will be studied more in the future.

The next three sections summarize some of the methods for melody extraction proposed in the past decade. Section 3.1 describes salience based methods, section 3.2 covers voice separation methods, finally section 3.3 describes other approaches.

3.1 Salience based methods

The method proposed by Paiva et al. [24] starts with computing a cochleagram followed by a correlogram. According to the authors this is sufficient for melody-detection task as it allows to capture only the pitches that most likely contain the main melody. Then the pitches are grouped into contiguous segments to detect notes. The algorithm is based on a minimum note duration of 125 ms. This threshold is based on empirical tests after an observation by Bregman, "Western music tends to have notes that are rarely shorter than 150 ms in duration" [2]. The melodic notes are selected using Gestalt principles, in particular sequences of pitches are established by intensity and frequency proximity. One of the most common error in pitch estimation and melody detection is the octave error (i.e., picking the right chroma but on a different octave, typically an octave or two higher than the right note). The algorithm uses another perceptual rules suggested by Bregman to eliminate the ghost octave notes, the rule of harmonicity and common fate [2]. The method looks for notes with common onset or ending and common modulation, that is whose frequency and salience sequences move in parallel octaves. Finally the most salient notes are selected by intensity and melody smoothness. The whole algorithm is summarized in figure 3.

The method proposed by Salamon et al. [26] starts by applying an equal loudness filter to the original audio followed by FFT. Due to the poor frequency resolution of FFT in the lower range, given the short frame window, the method also uses the phase vocoder method to provide a more accurate estimate of the peak's true amplitude and frequency. The salience function is computed as the sum of the weighted energies of the harmonics of the frequency peaks. The peaks of the salience function are the melody F0 candidates. The F0 candidates are then grouped into pitch contours after pruning the peaks that fall under a salience threshold. The grouping is performed using Gestalt proximity principles. Finally the melody is selected by detecting the non-voiced sections and filtering out octave errors and pitch outliers. The use of the phase vocoder to improve the resolution at lower frequencies might be the biggest advantage of this method compared to the previous one. The authors also claim to have improved the voicing detection over previous approaches, even though this task seems to be very dependent on the dataset, as shown by the false alarm rate shown at the end of the paper, which ranges from 5% in the best case to 24% in the worst.

The method proposed by Liao et al. [19] tries to exploit difference perceptual cues including loudness and timbre similarity and also accounts for masking effect. After the FFT four peaks are selected as follow: two peaks are based on loudness and masking, to select perceptually dominant peaks, and two more peaks are selected based on cepstral envelope and noise envelope respectively to select energy dominant peaks. The melody is then selected out of the candidate peaks using a Hidden Markov Model with a transition probability trained to privilege trajectory smoothness and spectral envelope similarity. The use of HMM to select the candidate peaks seem to be a sensible choice, instead of applying a priori melody principles the HMM might learn voice leading rules and practices from real examples. In practice the algorithm performs generally worse than the previous one, particularly on certain dataset. The poor performances might be due to a poor voicing detection mechanism, though.

3.2 Voice separation methods

The method proposed by Hsu et al. [4] is limited to detect sung melodies and is based on voice separation and trend estimation. The first step is critical for the algorithm. Since single pitch detection algorithms are much more robust and accurate than multi-pitch detection algorithms, if the voice source can be separated from the mixture, a single-pitch detection algorithm can be applied to the separated audio and get more a accurate result. Source separation is hardly a solved problem in the general case. Previous attempts to separate a voice source from the accompaniment were based on the limited range of the human voice. Still the range of fundamental frequencies for a singer can go as low as 80 Hz and as high 1100 Hz, even though lower or higher frequencies can be sung at times. Such a large range impacts the accuracy of the voice separation methods so this proposed algorithm uses a trend estimation step to reduce the range to be searched at any given time frame. The trend estimation process is shown in figure 4. The vocal component enhancement step is based an Harmonic/Percussive Sound Separation algorithm proposed by Tachibana et al. [28] The two step process is aimed at attenuating the energy of harmonic instruments and percussive instruments. The sinusoidal partial extraction step is similar to the same step in the salience based algorithms. The instrumental partial pruning step uses a Gaussian Mixture Model trained on voice partials and instrumental partials and uses the fact that human voice naturally contains both strong vibrato and tremolo while most of the musical instruments contain only one of them. Finally the pitch range estimation estimate the most probable vocal F0 in each frame then extends its boundary to tolerate for pitch shifting in the vocal range. Limiting the extraction to sung melodies has the definite advantage of lowering the complexity of the task, since some specific features of the human voice can be exploited to detect voiced regions and extract the melody. Unfortunately, it appears that there is still a lot of variability between different singers in range and timbre, so the overall accuracy is not higher than general, salience-based methods.

The algorithm by Yeh et al. [30] is an improvement of

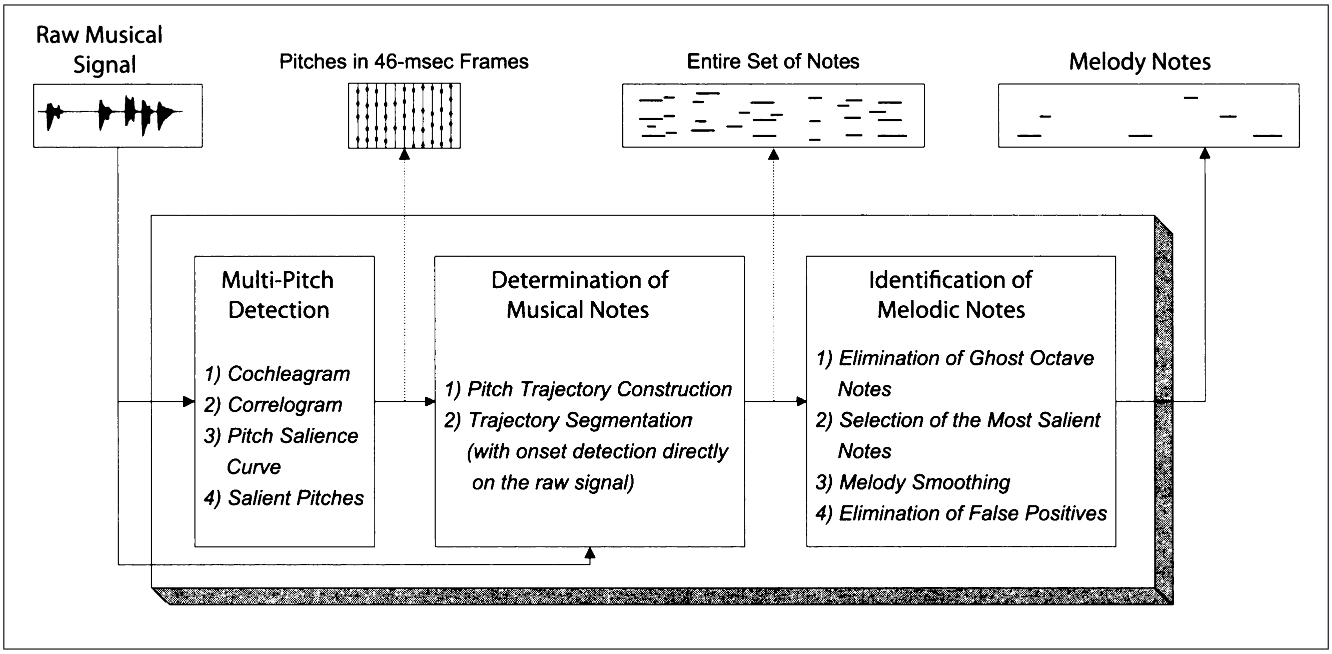
Raw Musical Signal
Pitches in 46-msec Frames
Entire Set of Notes
Melody Notes
Multi-Pitch Detection
1) Cochleagram
2) Correlogram
3) Pitch Salience Curve
4) Salient Pitches
Determination of Musical Notes
1) Pitch Trajectory Construction
2) Trajectory Segmentation (with onset detection directly on the raw signal)
Identification of Melodic Notes
1) Elimination of Ghost Octave Notes
2) Selection of the Most Salient Notes
3) Melody Smoothing
4) Elimination of False Positives

Figure 3. Overview of the melody detection system by Paiva et al. [24]

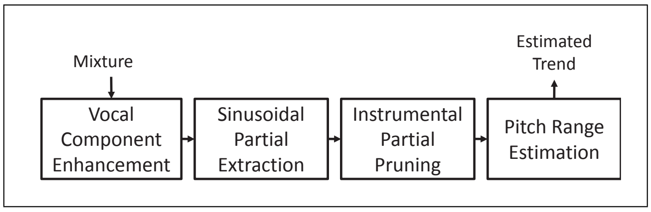
Mixture
Estimated Trend
Vocal Component Enhancement
Sinusoidal Partial Extraction
Instrumental Partial Pruning
Pitch Range Estimation

Figure 4. Schematic diagram of trend estimation by Hsu et al. [4]

the previous algorithm. It is still limited to sung melodies but it integrates three methods: forward and backward trend estimate and HMMs. The final step of the algorithm, the pitch combination scheme, gets multiple pitch contours as input. The papers propose three different mechanisms for picking the best contour: median, mean and dynamic programming. Experiments on the MIR-1K dataset shows that the median method always gives the best overall accuracy so it is proposed as the preferred mechanism. Integrating multiple approaches improves the trend estimation algorithm, as shown by the results presented by the authors, but since the accuracy of the previous method is heavily dependent on the dataset, it is arguable that this method is too. Interestingly, this method was not submitted to MIREX for independent evaluation.

3.3 Other methods

The method proposed by Arora et al. [1] uses a novel approach to sung melody extraction. The method, inspired from the Kalman Filter framework, aims at tracking a cluster of harmonic partials, called a comb. While other methods try to estimate the F0 calculating the salience of each partial, this method tracks all the harmonics from the same source at the same time, improving the accuracy and avoiding the octave error altogether. This method can also be implemented online. This method has several advantages over the previous ones, including the fact that it attempts to track all the harmonics from the same source at the same time, which is arguably something that human beings naturally do. Streaming pitches by timbre seems to be the natural approach to source separation as well.

The method proposed by Poliner et al. [25] differs from the other methods shown so far as it does not assume any prior knowledge of the domain, i.e. no assumptions are made on the harmonicity of the sounds. The authors claim that in other fields, like speech recognition, it is possible to build classifiers for particular events without any prior knowledge of how they are represented in the examined features. They tried to apply the same approach to melody extraction using a Support Vector Machine (SVM). SVM is a supervised classification technique. The authors trained the SVM with labeled audio synthesized from MIDI tracks, where the lead melody track is labeled, and extracted from multi-track recordings, where the vocal track is recorded separately. This method has the advantage of applying pure machine learning techniques without any assumptions from the domain. While domain knowledge is obviously very important when solving a problem, music is so diverse that most rules and principles have a limited scope of application; e.g., popular music is strcuturally different from classical music in many aspects; harmonic and melodic practices common in tonal music cannot be applied to modern and contemporary music.

The method proposed by Sam Myer for MIREX2012 is also radically different from the other methods presented so far as it is not based on any perceptual or signal processing techniques. Instead it leverages MIDI data mining over the Internet. The method has not been published nor peer-reviewed so further details are not available, nonetheless it

Method	Overall Accuracy
Paiva et al. [24]	61.1%
Poliner et al. [25]	61.1%
Salamon et al. [26]	61%–85%
Liao et al. [19]	35%–73%
Hsu et al. [4]	61%–83%
Yeh et al. [30]	82.6%
Arora et al. [1]	50%–80%
Myer et al.	47%–77%

Table 1. Accuracy of the presented methods.

performs surprisingly well compared to the other methods on the MIREX datasets.

4. CONCLUSIONS

This paper has given an overview of the melody extraction problem, the human mechanisms for melody perception, and the methods proposed for automatic melody extraction from audio. Despite very little understanding of human cognitive mechanisms of melody perception, most algorithms exploits one or more perceptual cues to improve accuracy. On the other sides, a few algorithms that do not explicitly exploit perceptual cues have been proposed and they achieve comparable accuracy in most tests. It might be argued that machine learning techniques might learn the perceptual cues from the training sets, though.

It would be interesting to apply machine learning techniques, especially neural networks and hierarchical clustering algorithms, to an annotated corpus of music and analyze the results. This approach has been successfully applied to other fields, like linguistic and vision, and has provided useful insights on human cognition processes. One important question from a cognitive point of view is what makes a melodic line stands out? Furthermore, what makes a musical line memorable? The last question is especially important for applications like query-by-humming where users are more likely to query a database with memorable features of a song.

On the perceptual side, more research is needed on pitch determination, timbre perception and stream segregation. Understanding how a complex auditory stimulus, e.g., a single note played by an instrument, can generate a stable perceptual representation, i.e., a stable auditory image, might lead to a model that can be simulated with higher accuracy than current methods. Timbre perception and stream segregation are strictly related as humans naturally and effortlessly combine all the harmonics produced by a given instrument into a single auditory percept. Once again, a deep understanding of these complex mechanisms might provide insights on how to automate the process via signal processing.

In conclusion, more research is needed in both the cognitive and computer audition fields. On the cognitive side the biggest open problem is how to design ecologically valid experiments while maintaining a full control on the experiment variables. On the signal processing side, the biggest challenge is how to segregate streams and track pitches played by the same source, i.e. tracking pitches by timbre.

Exploiting perceptual cues seems to be a sensible approach so far, but more sophisticated signal processing and machine learning approaches should be pursued further.

5. REFERENCES

[1] V. Arora and L. Behera. On-line melody extraction from polyphonic audio using harmonic cluster tracking. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(3):520–530, 2013.

[2] Albert S. Bregman. *Auditory Scene Analysis: the perceptual organization of sound*. MIT Press, Cambridge, Mass, 1999.

[3] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney. Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, 96(4):668–696, 2008.

[4] Hsu Chao-Ling, Wang DeLiang, and J. S. R. Jang. A trend estimation algorithm for singing pitch detection in musical recordings. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 393–396.

[5] Perry R. Cook. *Music, Cognition, and Computerized Sound : An Introduction to Psychoacoustics*. MIT Press.

[6] Alain de Cheveigné. *Pitch Perception Models*, volume 24, pages 169–233. Springer New York, New York, NY, 2005.

[7] W. Jay Dowling. Expectancy and attention in melody perception, 1990.

[8] Zhiyao Duan, Jinyu Han, and Bryan Pardo. Multi-pitch streaming of harmonic sound mixtures. *Manuscript for IEEE Trans. Audio, Speech and Language Processing, 2013.*

[9] Werner Goebl and Richard Parncutt. Asynchrony versus intensity as cues for melody perception in chords and real music. In *5th ESCOM conference, Sept 8–13, 2003.*

[10] Ben Gold, Nelson Morgan, and Dan Ellis. *Models of Pitch Perception*, pages 218–231. John Wiley & Sons, Inc, Hoboken, NJ, USA.

[11] M. Goto A real-time music-scene-description system: predominant-f0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication*, 43:311-329, 2004.

[12] P. Iverson and C. L. Krumhansl. Isolating the dynamic attributes of musical timbre. *Journal of the Acoustical Society of America*, 94:2595-2603, 1993.

[13] Paul Iverson. Auditory stream segregation by musical timbre: Effects of static and dynamic acoustic attributes. 21:751–763, 1995.

[14] M. R. Jones, R. R. Fay, and A. N. Popper. *Music Perception*. Springer, 2010.

[15] Michael Kennedy and Joyce Bourne, eds. *The Oxford Dictionary of Music*. Oxford University Press, 2004.

[16] Anssi Klapuri. Multiple fundamental frequency estimation by summing harmonic amplitudes. In *Proc. 7th In.t Conf. on Music Inform. Retrieval, Victoria, Canada, October, 2006*, 216-221.

[17] Anssi Klapuri, Jouni Paulus, and Meinard Müller. Audio-based music structure analysis. In *ISMIR*, in Proc. of the Int. Society for Music Information Retrieval Conference.

[18] Stefan Koelsch and Walter A. Siebel. Towards a neural basis of music perception. *Trends in Cognitive Sciences*, 9(12):578–584, 2005.

[19] Wei-Hsiang Liao, Alvin W. Y. Su, Chunghsin Yeh, and Axel Roebel. On the use of perceptual properties for melody estimation. In *Proc. of the Int. Conf. on Digital Audio Effects (DAFx-11), Paris, France*, 2011.

[20] R. F. Lyon. Machine hearing: An emerging field [exploratory dsp]. *Signal Processing Magazine, IEEE*, 27(5):131–139, 2010.

[21] F. Marmel, B. Tillmann, and C. Delbe. Priming in melody perception: tracking down the strength of cognitive expectations. *J Exp Psychol Hum Percept Perform*, 36(4):1016–28, 2010.

[22] R. Meddis and M. J. Hewitt. Virtual pitch and phase-sensitivity studied using a computer model of the auditory periphery: I Pitch identification. *Journal of the Acoustical Society of America*, 89:2883-2894, 1991.

[23] Brian C. J. Moore. *Hearing*. Academic Press, San Diego, 1995.

[24] Rui Pedro Paiva, Teresa Mendes, and Amílcar Cardoso. Melody detection in polyphonic musical signals: Exploiting perceptual rules, note salience, and melodic smoothness. *Computer Music Journal*, 30(4):80–98, 2006.

[25] Graham Poliner and Daniel Ellis. A classification approach to melody transcription. *ISMIR 2005: 6th International Conference on Music Information Retrieval: Proceedings: Variation 2: Queen Mary, University of London & Goldsmiths College, University of London, 11-15 September, 2005*, 161–166, 2005

[26] J. Salamon and E. Gomez. Melody extraction from polyphonic music signals using pitch contour characteristics. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(6):1759–1770, 2012.

[27] G. Schwarzer. Analytic and holistic modes in the development of melody perception. *Psychology of Music*, 25(1):35–56, 1997.

[28] H. Tachibana, T. Ono, N. Ono, and S. Sagayama. Melody line estimation in homophonic music audio signals based on temporal-variability of melody source. *IEEE ICASSP*, 425–428, 2010.

[29] David Temperley. A probabilistic model of melody perception. *Cognitive Science*, 32(2):418–444, 2008.

[30] Yeh Tzu-Chun, Wu Ming-Ju, J. R. Jang, Chang Wei-Lun, and I. Bin Liao. A hybrid approach to singing pitch extraction based on trend estimation and hidden markov models. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 457–460.

[31] RJ Zatorre, AC Evans, and E Meyer. Neural mechanisms underlying melodic perception and memory for pitch. *The Journal of Neuroscience*, 14(4):1908–1919, 1994.