GRAZ UNIVERSITY OF TECHNOLOGY, AUSTRIA (TUG)

INSTITUTE OF ELECTRONIC MUSIC AND ACOUSTICS (IEM)

UNIVERSITY OF MUSIC AND PERFORMING ARTS, GRAZ (KUG)

Detection of singing voice signals in popular music recordings

Diploma Thesis

Author: Amir Rahimzadeh

Graz, Nov. 3rd, 2009

Supervisor: Alois Sontacchi

Assessor: Prof. Robert Höldrich

Keywords: Voice / Music Separation, Multi Pitch Estimation, Multi Pitch Tracking

Table of Contents

Abstract

Automatic music content analysis is an important, diverse and challenging research field. In this field, automatic (singing voice) melody transcription is of special interest being the key to many applications like music structure analysis, score-following, query-by-humming and karaoke like applications (voice removal). The basis for this wide field of applications is the robust and reliable estimation of the singing voice fundamental frequency (F0) trajectory.

This thesis deals with the detection of singing voice signals in polyphonic popular music recordings. The main challenges are to somehow recognize individual sound sources from the complex music mixture signal and to classify them as vocal or non-vocal. To do so we propose analysis of the musical content on the basis of pitch tracks. These are extracted from the audio signal applying frame wise multi pitch estimation followed by a tracking algorithm which uses cubic interpolation to improve correct grouping of the estimates across frames. Auditory motivated preprocessing is applied to the audio signal reinforcing weak or missing fundamental frequency components before multi pitch estimation and tracking is performed. From the pitch tracks a set of features is derived that has been found to bear discriminability between vocal and instrumental sounds and which is finally used for identification of the singing voice F0 trajectory.

The proposed method has been evaluated using the ”MIREX 2005 – Training data set – vocal“. The pure pitch accuracy of the algorithm for vocal F0’s in polyphonic music mixtures yields 90.4% while classification between singing voice and instrumental sounds reaches 79.1%.

Kurzfassung

Die automatisierte Analyse des musikalischen Inhalts polyphoner Audiosignale ist ein immer wichtiger werdendes Forschungsgebiet. Von großem Interesse sind vor allem (Gesangs-) Melodie Extraktionsalgorithmen welche die Basis für eine Reihe interessanter Anwendungen bilden. Zu diesen zählen Strukturanalyse von Musikstücken, score-following (Synchronisation zwischen Notentext und akustischem Signal), query-by-humming (Durchsuchen digitaler Musikdatenbanken durch Singen/Summen einer markanten Passage) sowie Anwendungen im Karaoke Bereich, wie das Entfernen der Gesangstimme aus einem Musiksignal. Ausgangspunkt für eine zuverlässige Detektion der Gesangsmelodie ist die korrekte Schätzung des Zeit-Frequenzverlaufs des Grundtons der Gesangsstimme.

Ziel dieser Arbeit ist das Auffinden von Gesangssignalen in polyphonen Popmusik Aufnahmen. Die Herausforderung besteht darin, die einzelnen gleichzeitig auftretenden Klangquellen im komplexen Musiksignal zu erkennen und sie aufgrund Ihrer Eigenschaften als Gesang oder Instrumentalklang zu klassifizieren.

Der vorgeschlagene Ansatz beruht auf der Analyse von Grundtontrajektorien, welche in einem zweistufigen Verfahren aus dem Musiksignal geschätzt werden. Dazu wird das Audio Signal einer mehrfachen segmentweisen Tonhöhenschätzung (Multi Pitch Estimation, MPE) unterzogen, gefolgt von einem „Tracking“ Algorithmus, der die Tonhöhenkandidaten über Segmentgrenzen hinweg zu kontinuierlichen Frequenz-Trajektorien verbindet. Der Tracking“-Algorithmus verwendet kubische Interpolation um eine genauere Vorhersage des tatsächlichen Gundtonverlaufs einer Klangquelle zu ermöglichen. Außerdem wird das Signal vor der Tonhöhenschätzung einer dem menschlichen Gehör nachempfundenen Vorverarbeitung unterzogen, welche in der Lage ist schwache oder fehlende Grundtonkomponenten aus der Obertonstuktur eines Klanges zu regenerieren. Die so extrahierten Grundtontrajektorien werden schließlich aufgrund der Eigenschaften des Zeit-Frequenz-Verlaufs als Gesangs- bzw. Instrumentalklänge klassifiziert.

Die entwickelte Methode wurde mittels des „MIREX 2005 – Training data set – vocal“ evaluiert. Die Genauigkeit der Tonhöhenschätzung von Vokalklängen in polyphonener Musik liegt bei 90,4% während die Klassifizierung zwischen Instumentalklang bzw. Gesang bei ca. 79,1% liegt.

Danksagung

An dieser Stelle möchte ich ganz besonderen Dank meinen beiden Eltern aussprechen, die an mich geglaubt und mich auf meinem Weg immer unterstützt haben. Sie waren es, die es mir ermöglicht haben, mich in diesem Ausmaß dem Studium und der Entwicklung meiner Interessen und Fähigkeiten zu widmen.

Weiters möchte ich meiner Freundin, für die Kraft und den Rückhalt den sie mir gibt, danken.

Besonderen Dank möchte ich an dieser Stelle auch meinem Betreuer Hrn. Alois Sontacchi aussprechen, der durch seine Begeisterungsfähigkeit, sein weitreichendes Fachwissen, seine Persönlichkeit und vor allem seinen Sinn für Humor immer eine Quelle von Inspiration für mich dargestellt hat.

Vielen Dank auch an Prof. Robert Höldrich, Brigitte Bergner und an alle meine Kollegen.

Chapter 1 - Introduction to the topic

1.1. Motivation

The fast growth of digital music markets and the associated consumer electronic industry induced a need for automated music analysis and indexing methods. Broad scientific attention has been drawn to the research field of Music Information Retrieval (MIR) or computational music content analysis. Research dates back to the 1970’s but up to now algorithms generating full score transcriptions (harmonic-, melodic- and rhythmic content) have proven elusive [Poliner2007]. Higher transcription accuracy has been obtained by algorithms seeking to perform only partial transcriptions consisting of the chord sequence, the drum track or the melody.

Melody is a highly descriptive attribute of music enabling us to distinguish one musical excerpt from another [Selfridge-Field98]. Therefore algorithms able of automatically extracting the main melody (being the most salient one at a time) from audio recordings would open the field to a wide range of applications comprising music indexing, lyrics alignment, voice-removal (karaoke), score following, query-by-humming and other Music Information Retrieval based applications.

The focus of this thesis will be on the detection of the singing voice melody since in popular music recordings the main melody is usually carried by a human voice. The key to successful transcription of the singing voice melody is the robust and reliable estimation of the singing voice fundamental frequency (F_0) trajectory from the complex audio signal exemplified in Fig.0. The main difficulty is that the vocal and instrumental sounds usually significantly overlap in time and frequency. Moreover the F_0 of singing voice varies a lot with time especially at note beginnings which is challenging in terms of tracking (the correct association of frame wise multi pitch estimates to F_0-trajectories corresponding to the F_0s of underlying sound sources). Therefore singing voice F_0 estimation is closely related to the research fields of computational auditory scene analysis, source separation and multi pitch estimation (MPE).

This thesis is organized as follows. In the first chapter we will introduce the concept of pitch which is essential to human hearing and the basis for the perception of musical sounds. Pitch perception will further be investigated on the basis of the physiology of the human auditory system which motivated the auditory signal pre-processing able to reinforce weak F0 components. Then the signal characteristics of musical sounds and singing voice will be investigated and compared. The chapter ends with an introduction of the most basic concepts concerning the organization of musical pitches in western music. In the second chapter we will start with a general overview of (multi-) pitch estimation methods ending with the presentation of two recently proposed approaches towards singing voice F0 estimation that have been influential for this thesis. In the third chapter the proposed approach is introduced and explained in detail. Chapter four explains the evaluation framework and corresponding results are presented. Then in chapter five the proposed method will be reflected and finally the thesis ends with chapter six drawing conclusions for future improvements.

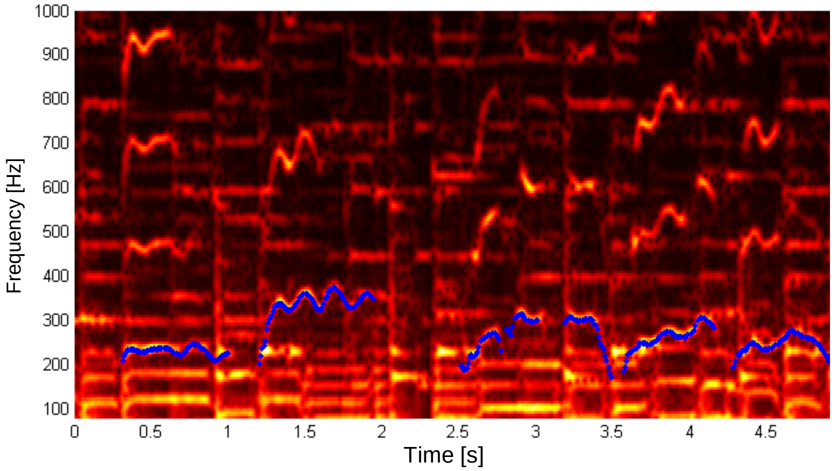


Fig. 0.: Singing voice F_0 trajectory (blue) plotted over the spectrogram representation of the corresponding audio excerpt

1.2. Perception of musical sounds

The capability of humans to orient to musical sounds is well known [Bregman90]. The basic acoustical characteristics of sounds namely pitch, loudness, timbre and duration are easily perceived, even by listeners without former musical education. Among these pitch is the most important for discriminating between concurrent sounds and thus essential for the transcription process.

1.2.1. Pitch

Pitch is a perceptual quality of sound which is fundamental to the hearing process. According to the ANSI standard 1994 [ANSI94], *“Pitch is that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high. Pitch depends mainly on the frequency content of the sound stimulus, but it also depends on the sound pressure and the waveform of the stimulus."* This is often referred to as the verbal definition of pitch which is rather impractical and inexact for analytical examination and comparison of sounds. A more practical definition of pitch is given by [Stevens75] who associates pitch with the measurable physical quantity of frequency given as follows: *“A sound has a certain pitch if it can be reliably matched by adjusting the frequency of a sine wave of arbitrary amplitude."* This is often referred to as the operational definition of pitch.

Sinusoids are the simplest kinds of sounds that evoke a pitch percept consisting of one single frequency component. There is a relation between temporal periodicity of the waveform and spectral energy distribution which can easily be observed under Fourier analysis which is illustrated schematically in Fig 1.1 and 1.2.

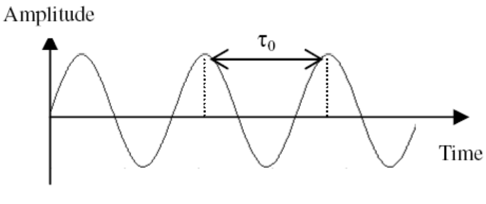


Fig. 1.1.: Fundamental period of a single sine wave

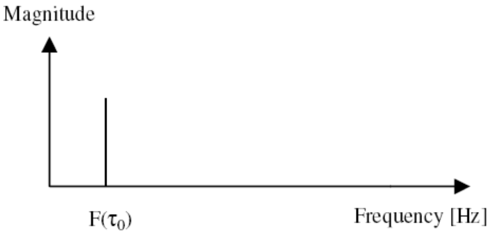


Fig. 1.2.: Fourier Spectrum of a sinusoid (schematically)

Temporal periodicity is described by the fundamental period τ_0 in seconds (see Fig. 1.1), corresponding to the fundamental rate of repetition in the waveform which is related to the frequency F in Hz by the simple equation

$$F = \frac{1}{\tau_0} \ [Hz] \qquad \text{(Eq. 1.1)}$$

However, natural instruments and also the voice tend to produce sounds with several frequency components which are called harmonics or also partial tones of a sound. They are found at distinct frequency locations forming a spectral pattern which exhibits energy concentrations at integer multiples of the lowest frequency component (see Fig. 1.4), which therefore is called fundamental frequency F_0. The partial tone frequencies are related to the F_0 by the following equation:

$$F_k = k * F_0 \quad \text{[Hz]} \qquad k = 1,2,3... \quad \text{(Eq. 1.2)}$$

These frequency components result from partial resonances largely determined by the sound generation mechanism. The waveform and the partial tone series of a complex tone is illustrated schematically in Fig.1.3 & Fig.1.4. The spectrum of a sound showing this property is said to be harmonic.

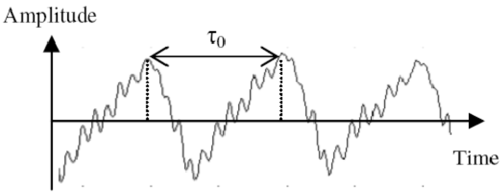


Fig. 1.3.: Fundamental period τ_0 indicated in the waveform of a complex tone

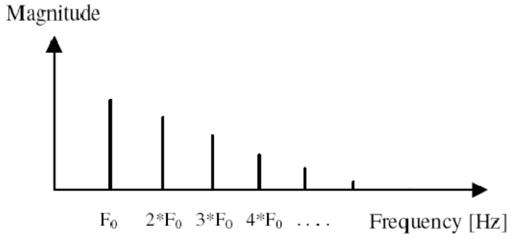


Fig. 1.4.: Partial tone series of a complex tone under Fourier Analysis (schematically)

Real instruments tend to produce partial tone frequencies slightly deviating from the ideal harmonic positions due to imperfect vibrating conditions. Inharmonicity is especially observed for instruments that use plucked (piano) and struck (guitar) strings due to the stiffness of real strings [Fletcher98]. Since popular music makes heavy use of these instruments this property has to be considered when designing multi pitch estimation (MPE) methods for music analysis.

The perceived pitch of complex harmonic sounds corresponds over a wide range of frequencies to the F_0 measured in Hertz. Since pitch is a pure subjective quality of sound which needs a human listener to make a perceptual judgment, it has established to computationally analyze the pitch of sounds on the basis of their F0, the lowest frequency component of a certain partial tone series. The two termini pitch and F0 are often mixed though conceptually different. Psychoacoustic findings for example show that a pitch is also perceived for harmonic sounds where the F0 component has been removed or masked which is referred to as the "missing fundamental" phenomenon. The pitch that is heard at F0 though actually not present in the sound stimulus is referred to as "virtual pitch" or also "residue pitch" indicated in the following illustration as "$F0_{virt.}$". This phenomenon can also be observed for sounds where only part of the partial tone series (e.g. Partial Tones: 3,4,5) is present. However, since certain terminology has established we will also use F_0 and pitch as synonyms for each other throughout this thesis.

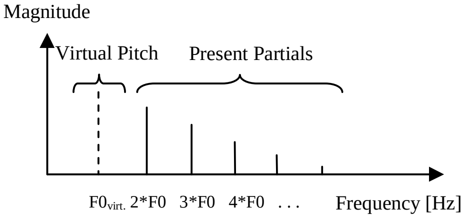


Fig. 1.5.: The "missing fundamental" phenomenon: Human listeners hear a virtual pitch at $F0_{virt.}$ though no spectral energy is present in the sounds stimulus

The spectral pattern formed by the partial tone series seems to be of importance for the pitch perception of complex tones. Indeed pitch perception models of Goldstein (1973) and Terhardt (1974) try to explain the missing fundamental phenomenon by some kind of pattern matching mechanism that is assumed to take place in the auditory system to derive pitch sensations for harmonic sounds according to [Plack04]. Apart from that, inharmonic sounds that don't have a distinct F_0 (like the sound of a church bell) might also evoke a pitch sensation which not necessarily corresponds to the lowest frequency component present.

Besides, psychoacoustic evidence emphasizes the fact that temporal regularity of the waveform of a sound stimulus and its envelope is of relevance too. For amplitude modulated white noise for example that has a random fine structure a pitch is heard according to the modulation frequency. In another experiment it has been demonstrated that for white noise that is periodically switched off listeners could perceive a pitch sensation according to the frequency corresponding to the inverse of the interruption rate. In both cases the long term average magnitude spectra of the sound stimuli are flat and don't show distinct spectral peaks [Plack04].

Obviously there is no simple relation between the frequency content or the waveform of a sound and the perceived pitch. Pitch perception is rather based on a complex interaction of physiological, neurological and high level cognitive processes working together to form a sensation of tone height.

However, the peripheral parts of the auditory system and its behavior to acoustic sensations are quite accurately known. According to [Plack04] the main characteristics of the peripheral part can be effectively simulated by a sequence of different signal processings which comprise auditory filtering, compression, half-wave rectification, and low-pass filtering. [Klapuri08] demonstrated that pitch estimation algorithms for music transcription might benefit from such a processing. Therefore our method makes use of the auditory motivated processing proposed by [Klapuri08] which is able to reinforce weak or missing F_0 components from the partial tone series which will be explained in detail in Chapter 3. First we will study the characteristics of the peripheral part of the auditory system which is the basis for the auditory motivated signal processing.

1.3. The human auditory system

Human listeners show a great capability to listen out individual sound sources in complex acoustical scenes and musical mixture signals [Bregman90]. It has been found [Plack04] that pitch is of major importance to this high level cognitive process which motivates the study of the human auditory system. Knowledge about how an acoustical signal is processed along the auditory path can be beneficial for pitch estimation algorithms [Klapuri08].

1.3.1. Physiology of hearing

Since there is plenty of work covering the physiology of hearing (e.g. [Pickles08]) we will mainly concentrate on the parts essential for the task of pitch perception. The human auditory system can be divided into the peripheral hearing system and the auditory cortex in the brain. While the characteristics of the peripheral part are quite accurately known the cognitive processes remain a matter of debate.

The peripheral part of the human auditory system can be divided into outer ear, middle ear and inner ear as illustrated in Fig. 1.6. The outer ear mainly contributes to directional hearing and acts as a resonator for frequencies around 4 kHz, while the middle ear performs an impedance match between the outer and inner ear. The inner ear contains the cochlea which is responsible for the transduction of the mechanical vibration into a neural representation, transmitted to the brain via the Auditory nerve. The cochlea is a sophisticated organ which performs a frequency-to-place transform of the input signal and therefore is of large interest for understanding human pitch perception.

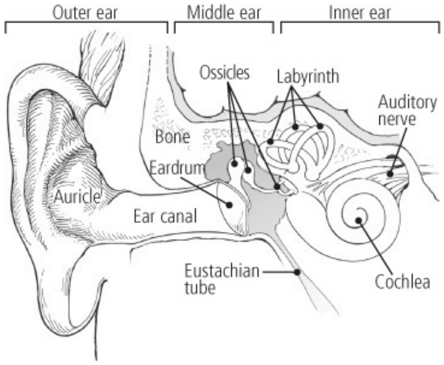


Fig. 1.6.: The peripheral part of the human auditory system (from [1])

Physiologically, the cochlea is a long coiled, tubular structure which is filled with liquid and which tapers towards its end. It is divided into two main sections by the basilar membrane over the whole length (see Fig. 1.7). Mechanical vibration transmitted by the middle ear enters the cochlea via the oval window resulting in hydraulic pressure fluctuations in the contained liquid causing the basilar membrane to vibrate in a specific way (see Fig. 1.10). Interestingly the points of maximal resonance on the basilar membrane vary with frequency. High frequency stimuli result in maximal excitation near the apex while low frequency stimuli result in maximal displacement towards the base which is referred to as the tonotopical organization of the BM illustrated in Fig. 1.8.

[1] Connect to research, 11.11.2009, URL: http://www.connecttoresearch.org/publications/72

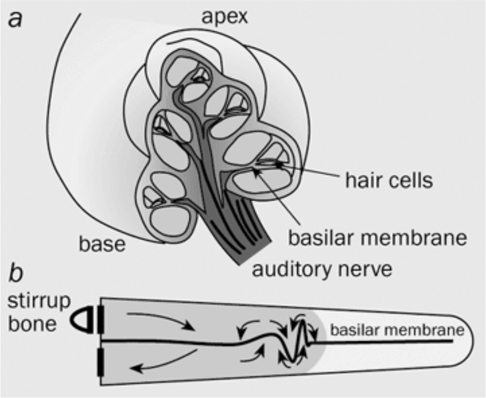


Fig. 1.7.: a.) Cut through the cochlea b.) Illustration of the excitation of the Basilar Membrane in the unrolled cochlea (from [Gelfand01])

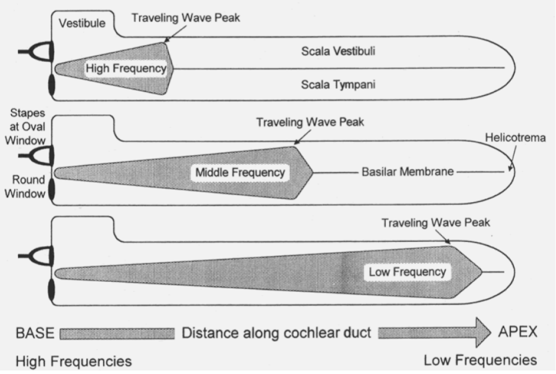


Fig. 1.8.:Tonotopical organization of the BM; the point of maximal displacement on the BM depends on the frequency of a sound stimulus(from [Gelfand01])

While the width of the basilar membrane increases almost linearly from base to apex, the stiffness decreases logarithmically. Therefore the relation between the frequency of a pure tone and distance of points of maximal resonance on the basilar membrane (BM) is nonlinear (see Fig. 1.9).

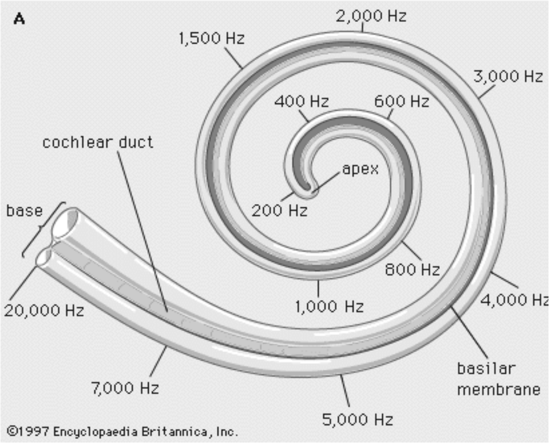


Fig. 1.9.: Nonlinear relation between points of maximal resonance along the basilar membrane frequencies of sinusoidal stimuli (from Encyclopædia Britannica, Inc.)

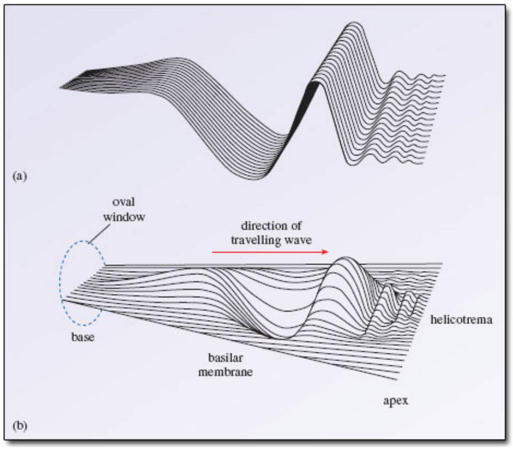


Fig. 1.10.: Excitation pattern of the basilar membrane, (a) if BM would not be fixed laterally (b) simulation of the actual BM motion (from [Gelfand01])

So in fact frequency is perceived logarithmically. Also musical instruments reflect this property. Sounds in octave relationship corresponding to a doubling of F0 are perceived particularly similar, resulting (at least for sinusoidal sounds) in equidistant points of maximal excitation on the BM. According to [Plack04] the excitation along the BM can be effectively modeled by a bank of band-pass filters with logarithmically spaced center frequencies and having bandwidths increasing with frequency.

It is known [Plack04] that in the high frequency (basal) region of the cochlea a compression of the input signal as much as 5:1 takes place, resulting in inter-modulation distortion products if more than one sinusoidal component is present, as it is the case for harmonic sounds. These distortion products occur at frequencies f_2-f_1 and at f_1-$k(f_2$-$f_1)$. For harmonic sounds the first term would always result in the fundamental frequency of that specific sound for every two neighboring partial tones which might partly explain the "missing fundamental" phenomenon.

Along the BM there are about 15.500 [Moore04] hair cells which react to mechanical displacement of the same, thus they are responsible for the transduction of mechanical vibration into neural impulses. The hair cells are connected to the auditory nerve which connects to the brain. Motion at different places along the BM causes activity in different neurons in the auditory nerve. Therefore it is believed that frequency of a tone is represented by a pattern of neural activity evaluated at higher cognitive levels.

There are two competing theories to the perception of pitch, the "place theory" and the "temporal theory" [Moore04]. The "place theory" states that pitch is perceived according to the specific point of maximal resonance on the BM which results in activity in distinct neurons. This is supported by measurements of the excitation of the BM in response to pure tones. Unfortunately this concept fails to explain the perception of complex tones for which the point of maximal resonance on the BM not necessarily corresponds to the perceived pitch. The "temporal theory" on the other hand states that a pitch percept is derived from the temporal pattern of neural impulses. This is supported by the observation that nerve spikes occur at a particular phase of the stimulating waveform for F0's up to about 5 kHz which is referred to as "phase locking" of nerve fibers [Moore04]. Neither of the two theories can explain the diversity of psychoacoustic phenomena that have been observed in humans and it might be more realistic to assume that both mechanisms work together. However, the cognitive processes involved in the perception of sound can only be studied indirectly, and therefore are not accurately known and remain a matter of debate.

Consequently modern pitch perception models aim at simulating the main characteristics of the peripheral part of the auditory system which are quite accurately known and comprise auditory filtering (simulation of the BM motion), neural transduction (compression, half-wave rectification, low pass filtering) followed by some kind of periodicity analysis. A processing strategy for MPE following the mentioned structure has been proposed recently by [Klapuri08] which is able to reinforce weak or missing F0 components by auditory motivated signal processing which has been adopted to a large degree and will be explained in detail in Chapter 3.

1.4. Organization of musical pitches

Western music is highly structured in time and frequency. The basic musical objects are notes which are specified by their pitch value, the duration and the onset time and rests which are moments of silence. Notes may be arranged sequentially forming melodies or vertically building chords which are explained by the concept of harmony.

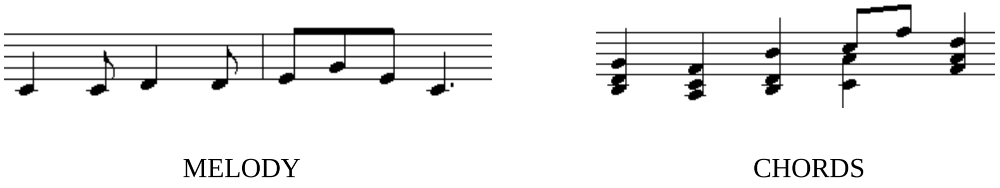


Fig. 1.11.: Possible arrangements of musical notes (from [1])

Musical notes are organized in intervals, which can be expressed as frequency ratios between F_0's. The smallest interval in western music is the semitone which refers to a F_0 ratio of $f_1/f_2=2^{(1/12)}$ between neighboring notes.

One particular interval is the octave, corresponding to a doubling or halving of the fundamental frequency. An octave is divided into 12 pitches according 12-tone equal tempered tuning which has established for modern western popular music. These pitches are referred to using letters C, C#, D, D#, E, F, F#, G, G#, A, A#, B and the corresponding F_0's are equally spaced on a logarithmical frequency scale according to $f_n=2^{(n/12)}f_{base}$ where n refers to the integer offset in semitones from the reference note. So the reference note f_{base} determines the pitch of the remaining semitones (usually $f_{base} = 440\ Hz$ or close to that). A distinct set of these 12 notes is referred to as scale. An example of a particular scale the C-major that corresponds to the white keys of the piano keyboard is illustrated in the following.

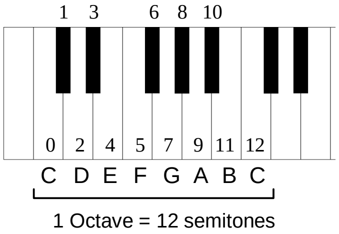


Fig. 1.12.: An octave and the 12 semitones illustrated on a piano keyboard

[1] WikiVisual, 18.10.2009, URL: http://en.wikivisual.com/images/5/55/Mussorgsky_Pictures_at_an_Exhibition,_chords.PNG

The notes of a scale repeat themselves for each octave as illustrated for the C-major scale in Fig. 1.13 over a range of 4 octaves. Two notes in octave relationship are perceived as particularly similar and share the same note- or pitch class symbol.

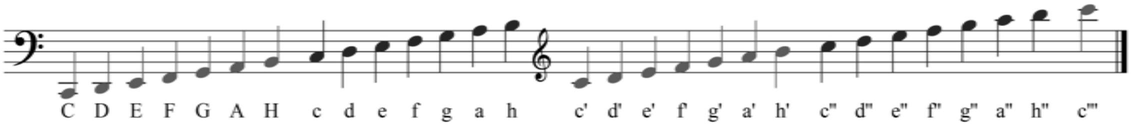

C D E F G A H c d e f g a h c' d' e' f' g' a' h' c'' d'' e'' f'' g'' a'' h'' c'''

Fig. 1.13.: The C-major scale for 4 different octaves. Notes of the same pitch class share a common class symbol referred to using letters (from [1])

It has been observed that the range of F_0's that can be used to produce melodies is limited to high frequencies and that recognition of musical intervals breaks down for F_0's above 5 kHz [Plack04]. It doesn't seem coincidental that the highest note in an orchestra played on the piccolo flute has a frequency of 4.096 Hz corresponding to c5.

The similarity in perception of tones in octave relationship is repetitive. This cyclic attribute of similar pitch sensation is referred to as "chroma". The relation between notes in octave relationship (thus notes sharing the same pitch class) and absolute frequency is illustrated in the following for the note A at different octaves.

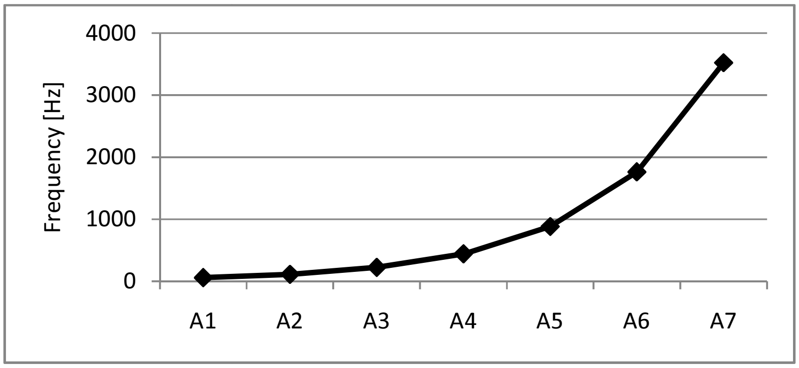


Fig. 1.14.: The relation between pitch of musical notes and F0 in Hz, displayed for the note A in different octaves

Obviously there is a nonlinear relation between frequency and perceived musical pitch which corresponds with physiological measurements of the inner ear.

[1] Wikimedia, 19.10.2009, URL: http://commons.wikimedia.org/wiki/File:Octaves.gif

1.5. Singing voice characteristics

In the following the main characteristics of singing voice will be analyzed and compared to the ones of speech. Further the differences between singing voice signals and instrumental sounds will be investigated.

Sounds produced by the human voice can largely be divided into voiced and unvoiced sounds. The main characteristic of voiced sounds is that the vocal folds vibrate to generate a sound. In fact, an airstream coming from the lungs is continuously modulated by the vocal folds resulting in pressure-pulse train. This spectrally rich signal exhibits frequency components at integer multiples of the F0, called harmonics. Unvoiced sounds in contrast don't show these properties.

Singing voice and speech signals are of course similar in many aspects due to the sound generation mechanism they share. Nevertheless certain differences in the signal characteristics exist and which will be explained in the following.

1.5.1. Ratio – Voiced / Unvoiced

It has been observed that in singing the ratio between voiced and unvoiced parts is significantly larger than for speech. The amount of voiced parts is about 60 % in speech while for singing it can increase up to 95% [Cook90]. This is due to the fact that singers intentionally stretch voiced parts to match the sounds of accompanying instruments. Since the voiced parts carry most of the musical information they are of major interest for the transcription process.

1.5.2. Singing Formant

The spectrum of voice sounds shows energy concentrations in certain frequency regions indifferent of the F_0 of a sound. These are called formants and arise due to the wave propagation properties of the vocal tract, the mouth and the nasal cavity which act as acoustical resonators. A well known difference between speech signals and operatic singing is the presence of an additional formant, called the singing formant, at frequencies around 2000 – 3000 Hz which can be seen in Fig. 1.17. After [Sundberg70] who first documented the existence of the singing formant, it helps the voice of a singer stand out of the accompaniment.

Unfortunately the singing formant seems to be an exclusive attribute of operatic singing which can be seen from the following comparison (Fig 1.17) of long term average spectra (LTAS) of a pop singer and an operatic tenor singer which performed the same excerpt in the same key using tempo and phrasing of their own choice [Borch02]. Averaging time was about 17 seconds.

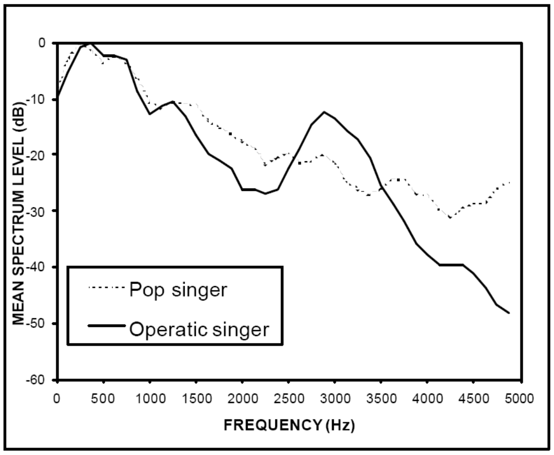


Fig. 1.17.: Comparison of the long term average spectra of a pop singer and an operatic tenor singer performing the same excerpt in the same key (from [Borch02])

1.5.3. Pitch range

Another difference is the pitch range which is about 80 Hz – 400 Hz for normal speech while it is about 80 - 1000 Hz for (operatic) singing [Li05]. For the songs of the vocal training data set (9 song excerpts of ~30 s each, detailed description in Chapter 4) that singing voice F_0’s concentrate on a narrower frequency range. This is illustrated in Fig.1.18, displayed parameters are the absolute frequency range (indicated by “*” and “o”), the standard deviation and the mean values of the reference vocal F0 trajectories (temporal resolution 10ms) for individual songs excerpts. Although the number of songs is small and therefore might not be representative for all songs of the genre popular music it can be expected that singing voice F0s in popular music will mainly concentrate on a narrow frequency range usually smaller than the range described by Li and Wang in [Li05].

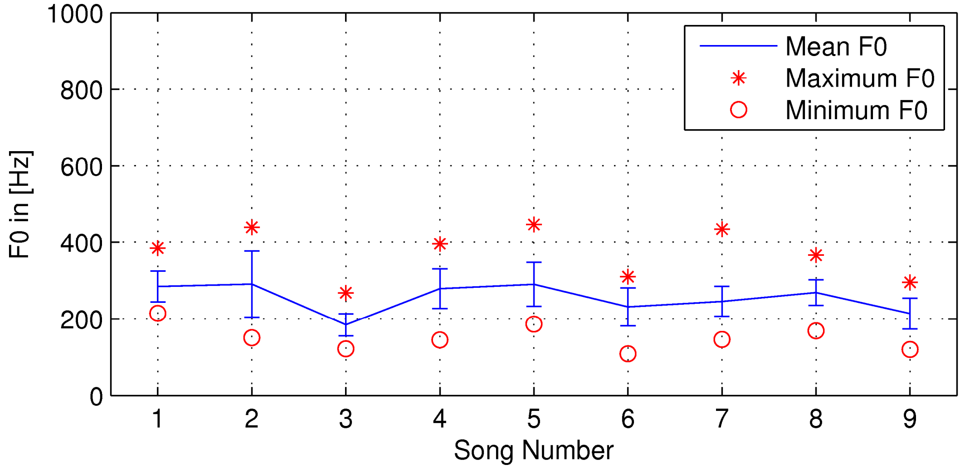


Fig. 1.18.: F0 Statistics of the training data set vocal (detailed description in Chapter 4). Displayed parameters: mean value (line), Standard deviation (bar), minimum-(circle) & maximum-(asterisk) F0 for individual song excerpts (1-9) of approximately 30s each

1.5.4. Pitch variability

While singing voice exhibits large pitch variability notes played on instruments result in pitches that are more or less stable in frequency during one note event [Sutton05]. This characteristic has been exploited recently for singing voice detection [Shenoy05]. A visual comparison of the spectral characteristics of a female singing performance and notes played on the piano is given in Fig.1.19 and Fig. 1.20.

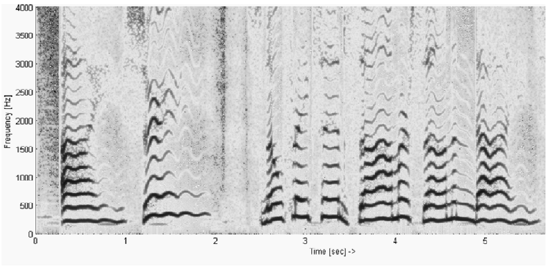


Fig. 1.19.: Spectrogram of a female vocal performance (~6s)

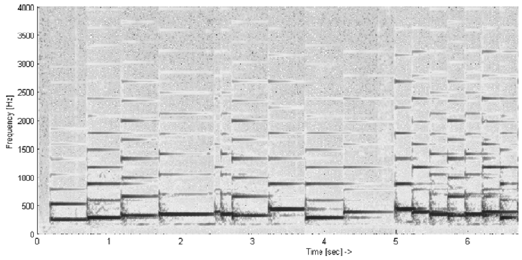


Fig. 1.20.: Spectrogram of a sequence of notes played on the piano (~7s)

It has to be noted that there are certain instruments and playing styles that don't allow such a clear distinction between vocal and instrumental sounds. Trombones for example are able of altering the frequency of a tone while playing resulting in continuous rather than discrete note transitions referred to as glissandi. Another example would be pitch bendings, a playing style on the guitar where individual strings are bended to continuously alter the frequency of a note.

Finally both singing and speech signals exhibits F0 fluctuations and they have been found to be larger for the former. For singing three distinct types of F0 fluctuations have been identified by [Saitou02] which have been found essential for the naturalness of vocal performances. These types are referred to as overshoot, vibrato and preparation and are indicated in Fig 1.21. With respect to our application the identification of these explicit types is not necessary. More important is the fact that there are certain signal characteristics that seem to be exclusive attributes of the singing voice.

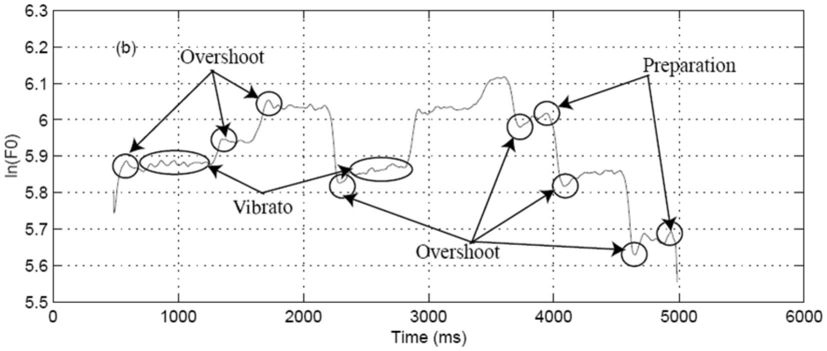


Fig. 1.21.: Three types of F0 fluctuations characteristic to singing voice signals (Note the logarithmic scaling of the frequency axis), from [Saitou02]

Knowledge of the mentioned singing voice characteristics will be exploited later on facilitating the discrimination of vocal and instrumental sounds. As will be explained in chapter 3 in detail we will derive a feature set that aims at capturing the described properties which will be finally used for voice recognition.

Chapter 2 - Literature Review

The transcription of a song refers to the derivation of a symbolic representation of the musical content in terms of notes which have been played by the individual instruments. Automatic polyphonic music transcription algorithms have gradually improved over the last decades but accuracy of full instrument transcription methods is still far from a satisfactory level and musically trained people still outperform computational methods in delivering reliable transcriptions [Klapuri06]. In order to reduce the complexity of the problem researchers have concentrated on the development of algorithms able to predict the most prominent pitch sequence in sound mixtures which is referred to as the main melody. It is the specific sequence of notes that human listeners usually agree on when reproducing an excerpt of a song and thus seems to be of high informative character when comparing two music performances [Selfridge-Field98]. We focus on vocal melody transcription algorithms since in popular music the main melody is usually carried by a human singer. The key to successful melody transcription is reliable and robust estimation of the singing voice fundamental frequency (F_0) trajectory. Numerous algorithms have been proposed sharing one general structure that we adopted which illustrated in Fig. 2.1.

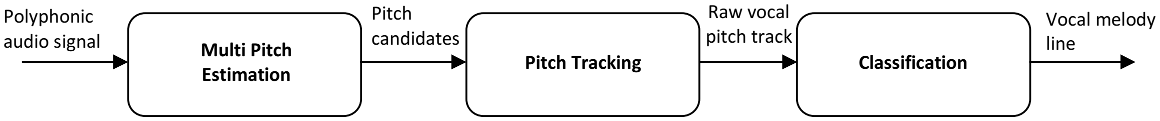


Fig. 2.1.: Basic structure underlying singing voice F0 estimation algorithms

The common primary step is to derive multiple pitch candidates from the polyphonic music signal. Since the acoustical waveform doesn’t allow the direct extraction of the individual pitches of the underlying sound sources the data has to be transformed to yield a representation which reveals the desired information. Multiple pitch estimation algorithms differ in the way they make use of the complex information contained in the acoustical waveform. They may largely be divided into algorithms that make use of spectral information using the well known Short Time Fourier Transform (STFT) and algorithms that analyze temporal periodicity of the waveform using correlation based methods, also hybrid algorithms exist. A third class of algorithms applies auditory motivated processing to the input signal before periodicity analysis is performed. This is also the approach followed in this thesis which is motivated by the well known ability of human listeners to orient to musical sounds. A good overview of existing multi pitch estimation methods can be found in [Klapuri06].

We adopted a pre-processing strategy proposed in [Klapuri08] which is able to reinforce weak or missing F0 components. This is very useful since MPE algorithms sometimes fail in predicting the exact pitch due to weak F0 components or spectrally interfering partial tones leading to F0 doubling or halving errors. Moreover the method proposed by Klapuri has been favored over others to serve as front end for our approach towards singing voice F0 estimation since the reported error rates [Klapuri08] are substantially low. They used solo instrument recordings (on the whole 2842 samples of individual note events comprising 32 instruments) to generate 4000 semi-random mixtures for different numbers of simultaneously played notes (N=1,2,4,6 – 1000 test cases each) which have been used for evaluation. Sounds have been mixed with equal mean amplitude and correct F0 estimates have been defined to deviate less than 3% from the reference (corresponding roughly to +/- ½ semitone). It has to be noted that the polyphonic test cases did not necessarily contain only musically meaningful note combinations. Reported error rates [Klapuri08] for multi pitch estimation are ~10% for combinations of 4 notes.

Given multiple pitch candidates over time the next challenging step is deciding which of the pitch candidates has most likely originated from a human voice and which ones have not, often referred to as vocal/non-vocal discrimination. Early methods have applied voicing decisions on a frame level assuming that the voice is constantly the strongest component in the mixture over time. This is not necessarily true and it would be more adequate to assume the voice to be predominant in a certain frequency region, namely the mid- and high frequencies, while the low frequencies (F0 <150 Hz) are usually dominated by the bass line. Moreover these early approaches didn’t particularly address the explicit differences in the signal characteristics of voice and instrumental sounds, the most apparent ones being pitch instability of vocal sounds and the mentioned dominance region. In the following, two recent approaches to vocal melody transcription specifically addressing the peculiarities of vocal sounds will be discussed.

2.1. Comparison of two recent approaches

[Goto06] proposed a three stage processing strategy comprising frame based pitch likelihood calculation, vocal probability calculation and F0 tracking based on Viterbi search. The front end of the famous “PreFEst” algorithm is used to perform frame based multi pitch estimation. A given spectrum is assumed to have originated from the superposition of multiple harmonic sounds sources which are modeled as probability density functions to enable the application of statistical methods. For each frame a maximum number of 10 pitches having the highest likelihood are selected. These predominant pitches are tracked over time and finally re-synthesized individually with a sinusoidal model using the parameters extracted from the specific locations in the power spectrum and phase spectrum corresponding to the partial tone frequencies. In this way separation of the individual sources is achieved. This time domain representation of the individual pitches is used to derive features which serve for vocal probability calculation. These features comprise Linear-Prediction-Mel-Frequency-Cepstral-Coefficients (LPMFCC’s, MFCC of the LPC derived spectrum) and ∆F0’s which both aim at capturing the mentioned voice characteristics. Two different types of GMM’s (Gaussian Mixture Models) are used to calculate voice probability of individual F0’s. They used a vocal-GMM and a non-vocal GMM which have been trained on features of vocal solo parts and polyphonic interlude sections, respectively. The training data set comprises 21 songs of 14 singers of the “RWC music database: Popular”. Finally given the vocal probabilities and considering continuity of F_0’s the most probable F0 series over time is found using Viterbi search. The algorithm has been evaluated using 10 songs from the “RWC music database: Popular” and pitch accuracy of 84.3% and chroma accuracy of 85.5% is reported in [Goto06].

Our method resembles the described one in that both apply frame wise multi pitch estimation and that both methods aim at generating continuous frequency trajectories which are used for the discrimination between singing voice and instrumental sounds. The two methods differ in the way how they make use of the information contained in the frequency trajectory. In contrast to the approach followed in [Goto06] we do not re-synthesize individual tracks since we believe the key characteristics of singing voice sounds being the F0 variability which we try to capture using a set of features derived from the frequency trajectory.

[Sutton06] has proposed vocal melody transcription based on two distinct pitch estimators which exploit characteristics of the human singing voice. A Hidden-Markov Model (HMM) is used to fuse the individual pitch estimates and to make voicing decisions. The first vocal pitch estimator consists of a pre-processing stage, where semi-tone cancellation is applied to emphasize vocal parts followed by a standard two-way mismatch (TWM) monophonic pitch transcription algorithm. Semi-tone cancellation is based on the fact that vibrato in singing voice (+/- 60-200 cent) has been observed to be larger than for instruments (+/- 20-35 cent). According to that spectral energy is removed at constant note F0 positions by zeroing the corresponding FFT bins in the vicinity of +/- 20 cent. This is done for all notes in the frequency range of interest spaced by one semitone. Vocals generally survive this procedure due to larger pitch variability while the accompaniment is attenuated. It seems feasible to apply a monophonic pitch transcription algorithm to the resulting signal. The second vocal pitch estimator consists of a correlogram based monophonic pitch transcription algorithm. It has been found that power of the upper partials of voice is generally larger than that of instruments and that accuracy of correlation based vocal pitch estimation is higher for high frequency channels than for low frequency channels [Li05]. According to that 19 channels between 3-15 kHz are used to derive individual pitch estimates and the most frequently occurring estimate is selected. Moreover this allows the formulation of a reliability measure reflecting whether these multiple frame-wise pitch estimates cluster or scatter. The method has participated in the MIREX 2006 melody extraction contest and was ranked third (overall accuracy ~67.3 %) on the “MIREX 2005 dataset – vocal” (9 songs of approximately 30 sec. each, described in Chapter 4) having the lowest voicing false alarm rate (12.3%) of all entered algorithms. The winning algorithm with an overall accuracy of ~73.7% did not specifically address the properties of singing voice and showed a voicing false alarm rate of 28.7%.

Similar to the described method we make use of two different multi pitch estimators to increase pitch estimation accuracy. Moreover both approaches aim at separating the voice from the accompaniment while using different strategies to achieve this. Our approach towards singing voice detection is based on the estimation of multiple parallel F0 trajectories present in polyphonic music signals. The representation of individual sound sources as time-frequency tracks is considered the main informative domain for discrimination between singing voice and instrumental sounds. In the following chapter our approach towards estimation of the singing voice F0 trajectory in polyphonic music will be presented.

Chapter 3 - The proposed method in detail

The analysis framework follows a hierarchical structure (see Fig. 3.1). The individual processing stages will be explained in detail one after another. First the auditory motivated preprocessing will be presented and the effects on an audio signal will be demonstrated. Then a twofold multi pitch estimation strategy is introduced and the resulting frequency discriminability will be investigated. Next a tracking algorithm is proposed that uses cubic interpolation to facilitate grouping of pitch candidates across frame boundaries to pitch tracks. The post-processing stage is responsible to reject unreliable pitch tracks. Next the feature extraction stage and the derivation of training data from audio recordings is described and discriminability of the training data based on the feature set will be investigated in different ways (Fisher’s Ratio, Linear Discriminant Analysis LDA). Then the K-Nearest Neighbor classifier will be presented which is used to discriminate between singing voice and instrumental sounds. Finally we will explain possibilities how to convert the F_0-trajectory into discrete note events. The whole programming has been done in MATLAB.

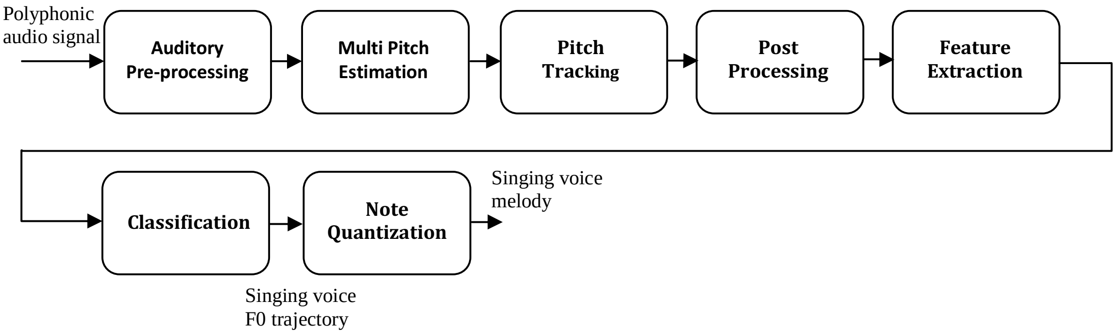


Fig. 3.1.: Block Diagram of individual processing stages of the proposed method towards singing voice F0 estimation

3.1. Data Preparation

Stereo audio files will be converted to mono files by simply adding the left channel and the right channel together. All input audio signals are normalized to their absolute maximal value.

3.1.1. Segmentation

Every input signal is divided into partly overlapping segments according to the selected frame size and hop size. We use a frame size of 92,9ms and a hop size of approximately 5ms. The frame size and the sampling frequency determine the discriminability between spectral components in the FFT magnitude spectrum. Further details will be explained in *3.1.4. Frequency discriminability of the periodicity analysis*.

3.2. Auditory Preprocessing

The human auditory system shows great capability in resolving and organizing musical sounds based on spacial, temporal, and timbral information. Therefore it seems natural to apply a similar kind of processing to the signal that happens in the auditory system before deriving further information. We adopted a processing strategy and periodicity analysis proposed by [Klapuri08] being the basis for further analysis. The individual processing steps explained in the following aim at simulating the transform characteristics of the inner ear and follow the common structure of pitch perception models [Cheveigne05] as illustrated in Fig. 3.2.

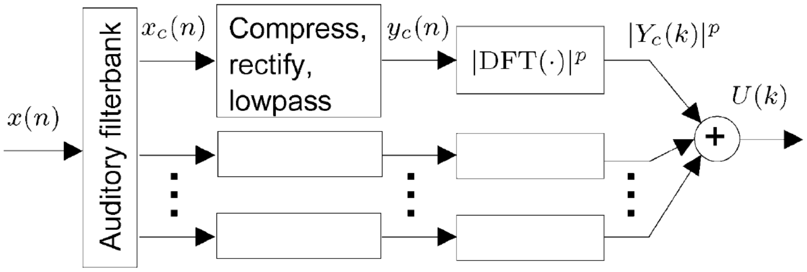


Fig. 3.2.: Common structure of pitch perception models
simulating the peripheral part of the auditory system from [Cheveigne05]

3.1.1. The auditory filter bank

The basilar membrane performs a frequency-to-place conversion. Sounds of different frequency result in maximal displacement at different points along the membrane. This frequency selectivity of the inner ear can effectively be modeled by an auditory filter bank [Plack04].

The power and impulse response of auditory filters have been studied in humans and animals and are quite accurately known [Patterson76], [Boer78]. The *gammatone* filter provides an excellent fit to the experimental data, and is therefore widely used [Patterson96].

The impulse response of gammatone filters is given by the following equation:

$$h(t) = b^{\eta} t^{\eta-1} e^{-2\pi bt} \cos(2 f_c t + \phi) \;\; \ldots t \geq 0, h(0) = 0 \qquad \text{(Eq. 3.1)}$$

The main parameters of the gammatone filter described by the impulse response h(t) in Eq.3.1 are b and η. According to [Patterson96] “b” largely determines the duration of the impulse response, and thus the bandwidth of the filter while “η” refers to the order of the filter determining its Q-factor. The parameter “f_c” corresponds to the filter center frequency in Hz and “t“ represents time in seconds.

As proposed by [Klapuri08] we use a total of 70 gammatone filters with center frequencies ranging from 65Hz to 5.2 kHz. The center frequencies are spaced uniformly on a critical-band scale resulting in a logarithmic frequency spacing (see Fig. 3.3) of neighboring auditory channels according to

$$f_c = 229 \times \left(10^{(\xi 1 * c + \xi 0)/21.4} - 1\right) \qquad \text{(Eq. 3.2)}$$

With ξ_0 being the critical-band-number of the lowest band, and $0<\xi_1<1$ determining the band density, in our case $\xi_0=2.3$ and $\xi_1=0.39$.

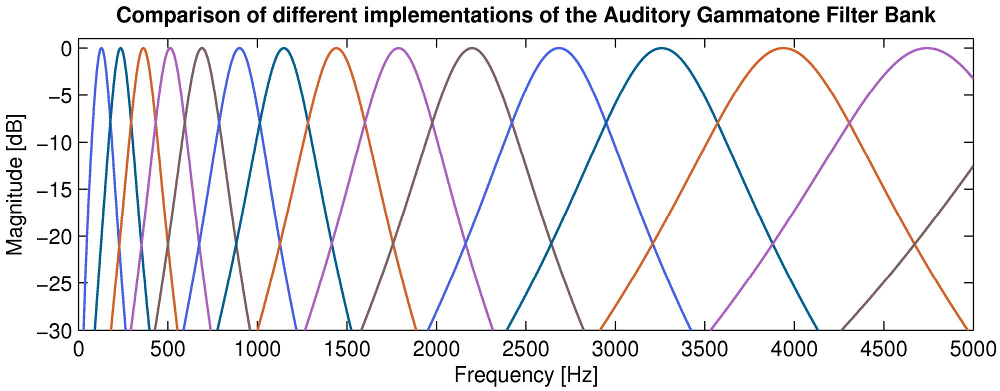


Fig. 3.3.: Magnitude Response for individual filters of the auditory filterbank, with logarithmically spaced center frequencies, every 4th filter displayed for better readability

One of the most important parameters of the filter bank is the bandwidth of the auditory filters. The equivalent rectangular bandwidth (ERB) of the filters used in this thesis have been reported in humans by [Moore95] given by

$$b_c = 0.108 f_c + 24.7 \text{ [Hz]} \qquad \text{(Eq. 3.3)}$$

The ERB of a filter is a measure for comparing the bandwidths of two filters. More specifically it is defined as the bandwidth of a perfectly rectangular filter which has an integral over its power response which is equal to the one of the specified filter. The auditory filters are implemented using a cascade of four second order infinite impulse response (IIR) filters. For a detailed description of the filter structure and an efficient implementation the reader is referred to [Klapuri08].

3.1.2. Neural Coding

At some point in the auditory system the physical waveform has to be transformed into a neural representation in order to be evaluated by the brain. This happens in the inner ear where hair cells excited by the BM motion generate nerve firings in the auditory nerve.

The main characteristics of the processing that a signal is subjected to in the inner ear can be effectively modeled as a cascade of signal processing operations: These comprise 1) dynamic gain control, 2) half-wave rectification, and 3) low-pass filtering [Klapuri06].

3.1.2.1. Compression

Measurements of basilar membrane motion show that the cochlea has a strong compressive nonlinearity over a wide range of sound intensities. The purpose of the strong nonlinearity may be recognized as an automatic gain control (AGC) that serves to map a huge dynamic range of physical stimuli into the limited dynamic range of nerve firings [Lyon95].

In order to become independent of the absolute level of the input signal a dynamic gain control is applied to the output signals of the individual auditory filters. To let all auditory channels contribute equally to the summary spectrum, the sub-band signals $x_c(n)$ of one analysis frame are scaled by the factor $\gamma_{c,t}$:

$$\gamma_{c,t} = \sigma_{c,t}^{\nu-1} \qquad \text{(Eq. 3.4)}$$

With c being the number of the corresponding auditory channel, t the analysis instant, and σ being the standard deviation of the signal $x_c(n)$ within the frame t. The parameter controls the amount of compression. For $0<\nu<1$ the auditory channel variances are normalized towards unity resulting in a spectral flattening of the summary signal (see Fig. 3.4). The value applied here is $\nu = 0.33$.

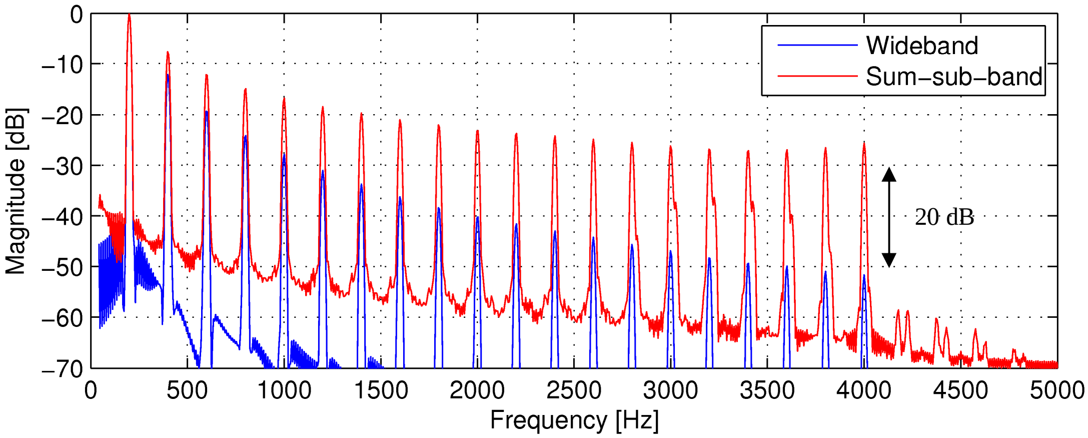


Fig. 3.4.: Effects of dynamic compression of the band-bass output signals of the auditory filter bank: Unprocessed wideband spectrum (blue) vs. normalized summary sub-band spectrum (red) calculated for an artificial harmonic tone complex of 220 Hz, maximal gain reduction ~25dB, FFT Settings: 92,9ms frame size, 4x zero padding

3.1.2.2. Half-wave rectification

Half-wave rectification (HWR) is a nonlinear signal processing operation introducing new frequency components to the original signal. While analytically difficult to track, the qualitative effect of HWR on the output signals of the auditory filters can be easily observed from the comparison of HWR-processed and unprocessed magnitude spectra. In Fig (b) & (c) the spectrum of the output signal of the 54th auditory filter (center freq. ~ 2.600 Hz) and its half-wave rectified counterpart are displayed for a synthetic harmonic tone complex of 250 Hz.

Frequency components are introduced in the base band and at multiples of the channels center frequency (see Fig. 3.5 – c). These arise due to beating components corresponding to the frequency intervals between the input partial tones. The most prominent interval usually corresponds to the F_0 since harmonic sounds exhibit a partial tone series that shows a constant spacing between consecutive partials tones. Fig. 3.5 demonstrates the effect of HWR.

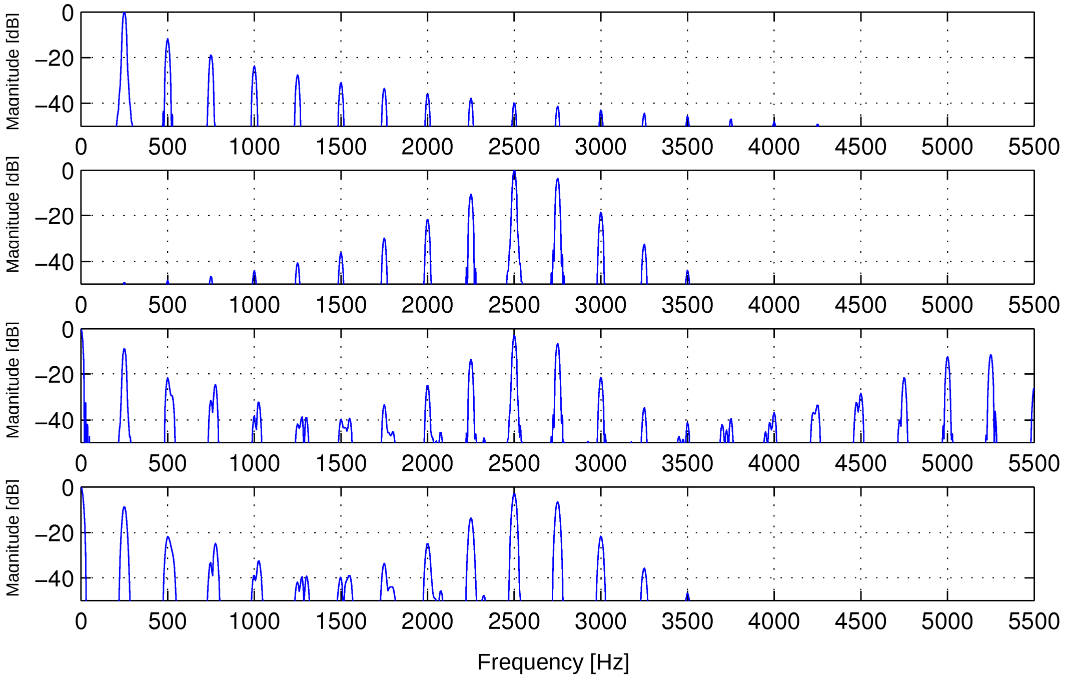


Fig. 3.5.: Effects of half wave rectification of sub band signals: a) Wideband spectrum b) sub-band spectrum of output signal of auditory channel no. 54 (fc~2.6 kHz), c) spectrum after half wave rectification, d) spec after HWR & LP-filtering. Analyzed signal has been an artificial harmonic tone complex with F0=250 Hz, fs = 11025 Hz

3.1.2.3. Low-pass filtering

The components generated at twice the center frequency (Fig 3.5 - c) due to HWR of the output signals of the auditory filter bank have not been reported to be of use [Klapuri08] since they are not guaranteed to match the harmonic series of the sound due to imperfect harmonicity. Therefore the frequency components at twice the channels center frequency are rejected by low-pass filtering the individual signals with cut-off frequencies according to the channel's center frequency using FIR filters of order 64. The qualitative effects of the described processing (1) Auditory filtering, (2) compression, (3) HWR, (4) LP-filtering have been demonstrated using synthetic harmonic tone complexes.

Finally the ability of the processing strategy to reinforce missing fundamental frequency components is demonstrated (Fig.3.6) using a synthetic harmonic tone complex of 250 Hz which does not exhibit energy at F0 location (Fig.3.6 (a)). Further spectra of the output signals of several auditory filters after auditory processing (b,c,d,e) are displayed. At last the summary spectrum (f) is shown which results from summing up the spectra from individual auditory channels.

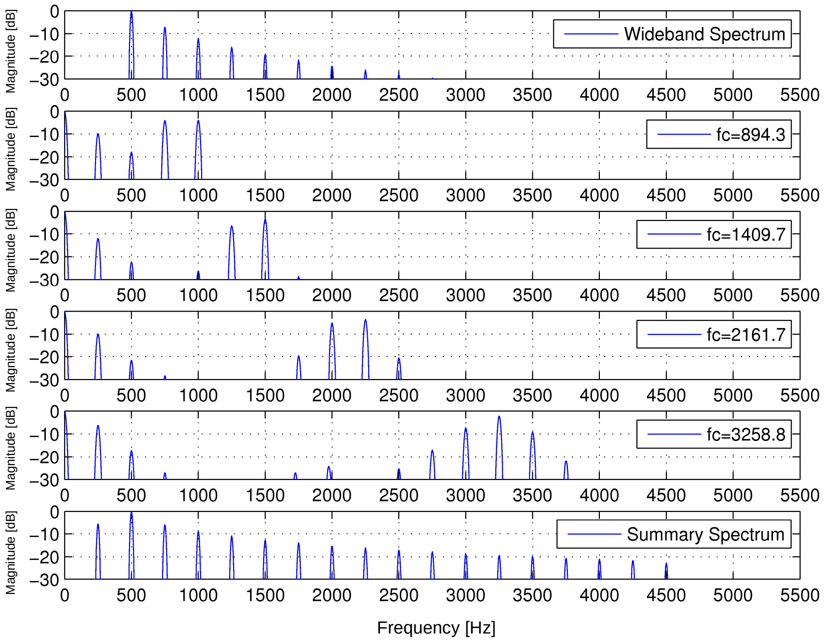


Fig. 3.6.: Reinforcement of the missing fundamental frequency component due to auditory motivated pre-processing exemplified for an artificial harmonic tone complex of 250 Hz with 0 amplitude F_0 component: (a) Magnitude Input Spectrum, (b)(c)(d)(e) Magnitude Spectra of output signals of individual auditory filters after channel normalization, HWR and LP-filtering. (f) Summary Spectrum showing reinforced F0 component

3.1.3. FFT based Periodicity analysis

Though it is not well understood how pitch is represented at higher cognitive levels, there is large evidence that some kind of periodicity analysis takes place in the individual auditory channels and that information is combined across sub-bands to yield the pitch percept [Meddis91].

A lot of pitch perception models assume this periodicity analysis to be based on some kind of auto-correlation function derived from sub-bands signals. Moreover experimental evidence supports this point of view. [Cariani96] studied the signals in the auditory nerve of cats in response to complex sounds. They recorded the responses in 507 nerve fibers and computed histograms of successive and non-successive neural spikes and combined the histograms. They found that pitch correlated strongly with the most prominent inter-spike intervals which suggests that the brain analyzes inter-spike patterns to form a pitch percept.

While autocorrelation based pitch estimation methods are able to predict well the pitch of individual sounds they fail sometimes in predicting the F_0s of multiple concurrent sounds. Even the highest maximum of the summary auto-correlation function (SACF, which is attained by summation of ACF's of individual channels) does not necessarily correspond to any of the actual pitches. Certain pitch relationships can confuse these models as it is the case for the major triad as for the interval of a perfect fifth. In these cases the constituent notes match the overtone structure of a non-existing chord root, leading to a maximum in the SACF corresponding to the virtual root note instead of the corresponding F0s.

Major Triad: Root note A = 440 Hz		**Interval of a perfect fifth: on A = 440 Hz**	
Constituent notes:	A C# E	Constituent notes:	A E
Corresponding F0s:	440 550 660	Corresponding F0s:	440 660
Virtual Root Note:	110	Virtual Root Note:	220

Fig. 3.7.: Virtual Root notes due to specific relation between F0's

Moreover autocorrelation based pitch models do not provide good robustness against additive noise. Especially in music the harmonic content is often accompanied by drums polluting the pitch information. For the mentioned reasons we apply FFT analysis to the sub-band signals instead of correlating them. Information across channels is integrated by summing up individual spectra yielding the summary spectrum which will be referred to as U_{Σ}. It should be noted that FFT and the ACF are closely related to each other over the power spectral density (PSD) [Oppenheim04]. They can be thought of as two different representations of the periodicity information contained in a signal.

3.1.4. Frequency discriminability of the periodicity analysis

Complex polyphonic audio signals can be seen as non-stationary signals, meaning that the signal content varies rapidly. Therefore they are usually analyzed using block–processing where a signal is cut into overlapping signal segments and each segment is analyzed by itself. In this way quasi-stationarity can be assumed, meaning that statistical properties don't vary much at least during one segment. This is a necessary condition for FFT analysis to correctly represent the signal content. The segments are also referred to as signal frames and the segment length as frame size measured in samples. There are two parameters determining the frequency discriminability ΔF_{FFT} between spectral components in the FFT of a signal frame, namely the frame size and the sampling frequency. The relation between the two is given by:

$$\Delta F_{FFT} = \frac{sampling\, frequency}{frame\ size} \quad [Hz] \qquad \text{(Eq. 3.5)}$$

Thus a large frame size is needed to yield high frequency discriminability. Since the spectral content of complex audio signals varies rapidly the frame size cannot be made arbitrarily large because as mentioned the signal content is only represented accurately by the Fourier Transform for stationary or at least quasi-stationary signals. Therefore the frame size has to be chosen as large as possible to yield high frequency discriminability and as small as necessary to guarantee stationarity of the signal during one analysis frame.

Usually the frame size is chosen according to the smallest frequency difference that has to be resolved for a given application. In our case we want to resolve frequencies in the range of 98 – 784 Hz corresponding to notes G2 – G5 which is the expected frequency range of singing voice fundamental frequencies. With respect to resolving harmonically related F0's like in our case the FFT resolution should be at least $\Delta F_{FFT} = 2\, semitones$. This is reasonable since two notes with a spacing of less than 2 semitones played together are perceived as particularly dissonant and are therefore rarely used in popular music. The smallest frequency difference that needs to be resolved occurs for the lowest pitch of interest being

$$\Delta F_{FFT} = \Delta F_{smallest} = F1\text{-}F2 = 98 - 110 = 12\ Hz. \quad \text{(Eq. 3.6)}$$

This would require a frame size of roughly 83ms corresponding to 918 samples at a sampling frequency of 11025 Hz. Rounding the number of samples to the closest power of two results in 1024 samples per frame at a sampling frequency of 11,025 kHz corresponding to 93ms frame length, a common value used for multi pitch estimation. This results in a frequency resolution of $\Delta F_{FFT} = 10{,}76$ Hz.

To avoid the well known leakage effect due to discontinuities at frame boundaries each signal frame is multiplied with a Hann window of the same size. The multiplication in the time domain corresponds to a convolution in the frequency domain which means that every bin of the FFT of the signal frame is convolved with the FFT of the window function. The consequence is spectral smearing which degrades the frequency resolution. The resulting frequency resolution depends on the main lobe width of the window function. By convention different window functions are compared based on the main lobe width at -6dB measured in bins. The Hann window has a main lobe width of 2 bins, so frequency resolution is degraded by a factor of 2 and the effective resolution of the FFT analysis is

$$\Delta F_{FFT\text{-}eff} = 21{,}52\ Hz \quad \text{(Eq. 3.7)}$$

Thus, the desired frequency resolution of 2 semi tones is achieved only for pitches higher than the note F3 corresponding to a frequency of 174,6Hz. This is not so critic since at low frequencies such small intervals result in harsh sounds and therefore are usually not used. This harshness or dissonance is perceived when simultaneous sinusoidal components are separated less than a critical bandwidth resulting in an interaction of the excitation patterns on the basilar membrane [Cook99]. The relative critical bandwidth (with respect to the center frequency) is larger at low frequencies and so small intervals result in larger harshness at low frequencies than at higher ones.

Finally the discriminability of the periodicity analysis in terms of resolvability of spectrally close frequency components is demonstrated in Fig. 3.8 for a mixture of two sinusoids with frequencies F1-F2 larger than $\Delta F_{FFT\text{-}eff}$ and in Fig. 3.9 for two sinusoids with frequencies F1-F2 smaller than $\Delta F_{FFT\text{-}eff.}$

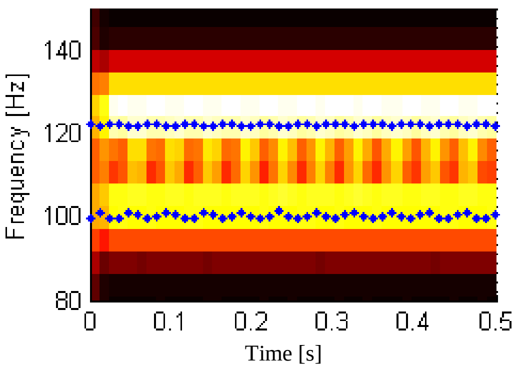


Fig. 3.8.: F0 Discriminability of the periodicity analysis for spectrally close sinusoids: Spectrogram and derived pitch estimates (blue) for 2 sinusoids of 0.5s length, with $\Delta F=|F1-F2|>\Delta F_{FFT\text{-}eff}$ (F1=100 Hz / F2 = 122 Hz), displayed between 80 and 150 Hz, spectrogram settings - frame size 92.9ms (1024samples @ fs = 11025) Hz, hop = 11.6ms

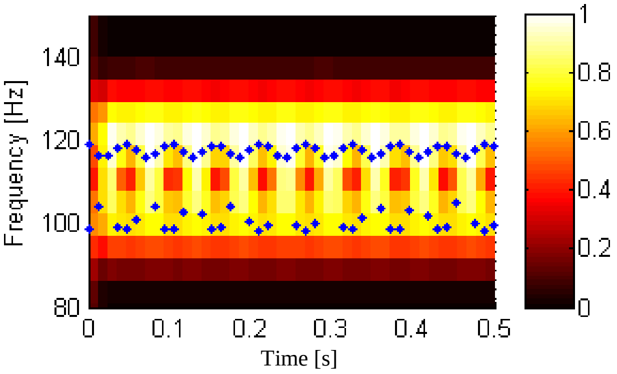


Fig. 3.9.: F0 Discriminability of the periodicity analysis for spectrally close sinusoids: Spectrogram and derived pitch estimates (blue) for 2 sinusoids with $\Delta F=|F1-F2|<\Delta F_{FFT\text{-}eff}$ (F1=100 Hz / F2 = 118 Hz). As expected correct F0 estimation breaks down for frequencies closer than $\Delta F_{FFT\text{-}eff}$

3.2. Multi Pitch estimation

It is well known that multi pitch estimation (MPE) algorithms sometimes fail in predicting every single fundamental frequency of the underlying sound sources correctly which is getting worse, the higher the number of concurrent pitches. Apart from that MPE algorithms tend to be biased towards one type of octave error, F0-doubling or F0-halving errors. We therefore apply a twofold pitch estimation strategy in order to increase the robustness and reliability of multi pitch estimates which is illustrated in the following.

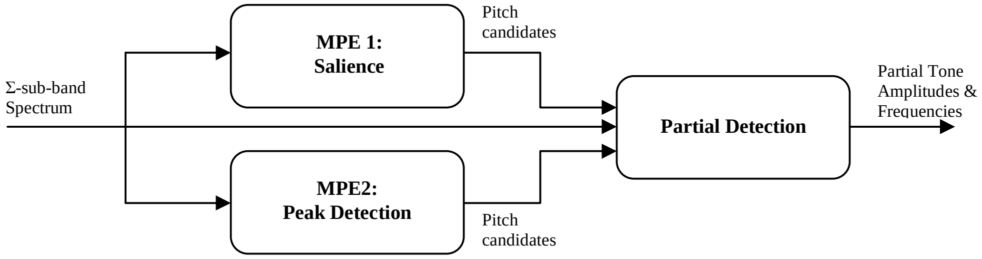


Fig. 3.10.: Block Diagram of the proposed multi pitch estimation strategy

First we apply two different MPEs to the summary sub-band spectrum (U_{Σ}). Based on the pitch candidates from both MPEs the partial detection stage tries to locate the actual peaks of the corresponding partial tones in the FFT-spectrum of one analysis frame. This shall improve reliability of pitch candidates since musical sounds and especially voice sounds tend to have harmonic spectra. Moreover this allows multiple spectral peaks to be assigned to an underlying sound source.

3.2.1. MPE1 – Salience Function

The MPE 1 makes use of a salience function to determine the harmonic strength of concurrent pitch candidates. The salience function is a function of fundamental period τ in seconds equivalent to 1/F_0. For each period candidate τ the salience s(τ) is calculated as the weighted sum of the amplitudes of harmonic partials derived from U_Σ(k). Salience is calculated corresponding to equation 3.7.:

$$s(\tau) = \sum_{m=1}^{M} w(\tau,m)U_\Sigma(k_m) \quad with\ k_m = Km/\tau\ and\ K = N/fs \qquad \text{(Eq. 3.7)}$$

with k_m corresponding to the partial tone location in the U_Σ and w(τ,m) being a weighting function (Eq. 3.8) determining to which degree individual partial tone amplitudes contribute to the salience of a specific period candidate. The largest peaks in the salience function usually correspond to the most salient pitch candidates.

$$w(\tau,m) = \frac{fs/\tau + \xi_1}{mfs/\tau + \xi_2} \qquad \text{(Eq. 3.8)}$$

The individual weights depend on the fundamental period of a pitch candidate and on the partial tone number and are given by equation 3.8 and displayed in Fig. 3.11for the first 5 partial tones for various frequencies. To the right the weighting function is illustrated for increasing frequency for the first 5 partial tones.

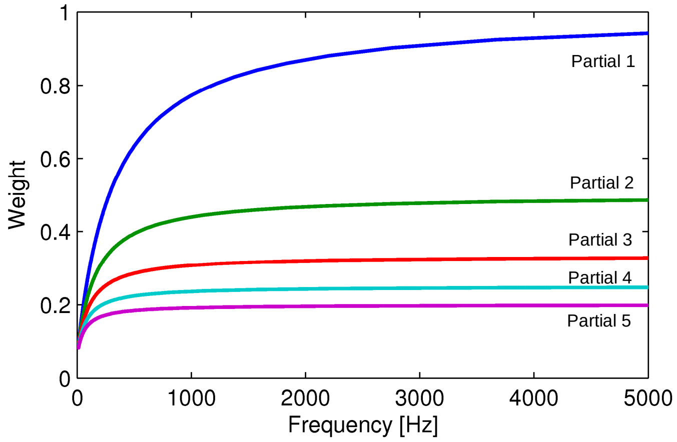


Fig. 3.11.: Weight function for the first 5 partial tones with increasing frequency

The weighting function has been evaluated using 4000 sound mixtures with different number of polyphony (Polyphony = 1,2,4 and 6, 1000 samples each) and the above form has been found to result in the most reliable peaks in the salience function [Klapuri06]. ξ_1 and ξ_2 in Eq. 3.8 are moderating terms important for low frequency F0's. The exact values have been adopted being ξ_1=20 Hz and ξ_2=320 Hz for analysis frames of approximately 93ms. Without these terms the weighting function would reduce to 1/m.

In the following (Fig. 3.12) the salience function is displayed for a mixture of 3 synthetic tones (20 partials each, with equal exponentially decreasing partial amplitudes) with fundamental frequencies of F_{0-1}=440, F_{0-2}=550 and F_{0-3}=660 Hz.

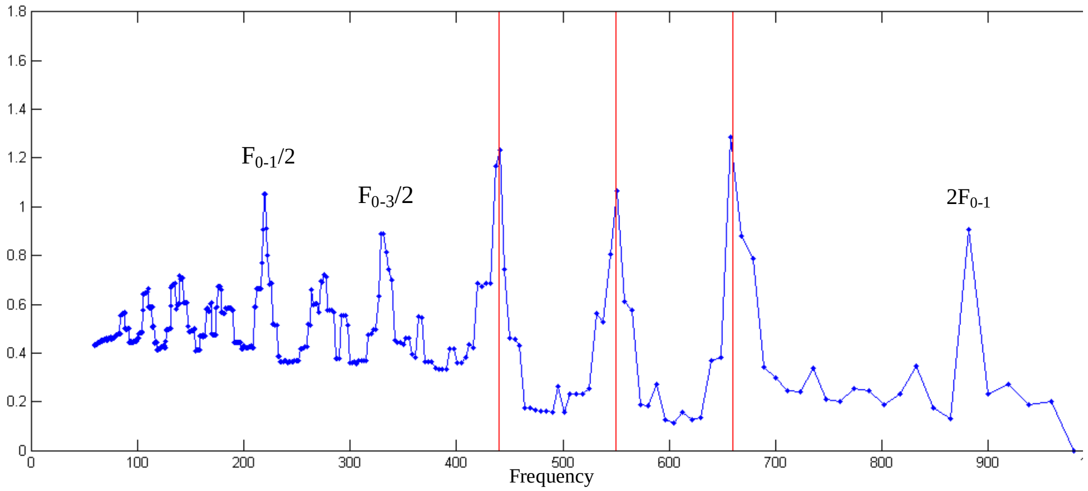


Fig. 3.12.: Salience function (blue) for a mixture of 3 synthetic tones with fundamental frequencies of $F0_1$=440, $F0_2$=550 and $F0_3$=660 Hz, red vertical lines indicate the exact F0 positions

The evaluation of the salience function s(τ) at equally spaced fundamental periods τ results in a nonlinear frequency resolution which is decreasing for increasing F_0. This can be observed in Fig. 3.12. The spacing between the blue dots, indicating consecutive values of s(τ), is getting larger for higher frequencies. Instead of equal frequency resolution over the whole frequency range, there is high resolution at low frequencies where F_0s tend to be very close.

Apart from that it can easily be observed that the salience function also exhibits peaks at double and half F_0 of the underlying pitches. This is due to the fact that pitches in octave relationship share many partial tones which support putative pitch candidates at $F_0/2$ and $2xF_0$. Therefore instead of assuming the N highest peaks in the salience function to correspond to the real pitches, a technique called "*iterative estimation and calculation*" proposed by [Klapuri08] is applied. There, after detecting the highest peak in the salience function the corresponding partial tone series is partly removed from the U_{Σ}. Then the salience function is re-estimated and again the highest peak is detected and so forth.

Since partial tone frequencies of different harmonically related pitches often overlap, they will not be removed completely which could negatively affect the detection of the remaining pitches. Instead, the partial tone amplitudes will be weighted using the weight function introduced earlier, before removal. As can be seen from the weighting function in Fig.3.11 low F_0's and low partials are removed less than higher ones. This accounts for the fact that F_0's and corresponding harmonics are close and usually overlap significantly at low frequencies while they tend to be better separated the higher the F_0.

3.2.2. MPE2 – Peak Detection

In addition to the MPE 1 proposed by [Klapuri08] we apply a simple peak picking routine to U_{Σ} in order to reveal unrecognized peaks. This is necessary since informal tests showed that increasing the number of estimated pitches in the MPE 1 doesn't help finding all the remaining pitch candidates due to the nature of the estimation procedure. In Fig. 3.13 and Fig. 3.14 this is exemplified for the frame-wise pitch candidates of MPE1 and MPE2 of a musical excerpt are plotted over its spectrogram. A total number of 8 pitches have been estimated. As can be seen from Fig. 3.14, MPE2 is capable of detecting all relevant peaks in the spectrogram.

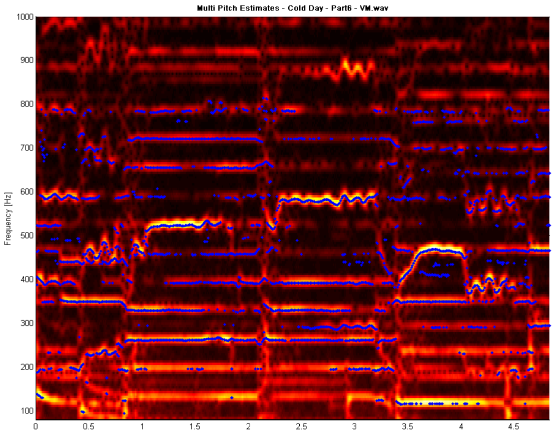


Fig. 3.13.: Frame wise pitch candidates (blue) from MPE1, # estimated voices = 8

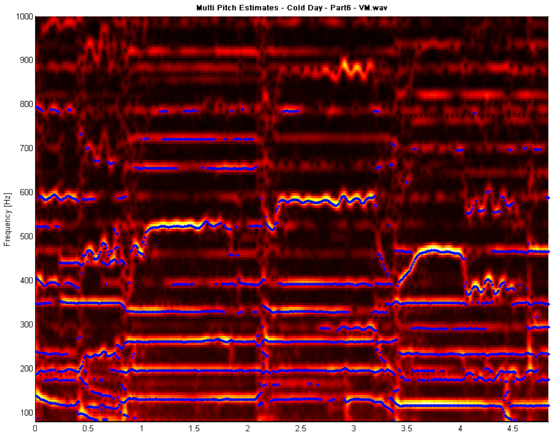


Fig. 3.14.: Frame wise pitch candidates (blue) from MPE2, # estimated voices = 8

3.3.3. Guided partial tone detection

Up to now pitch candidates have been derived from the summary spectrum U_Σ by comparing spectral energy at harmonic locations and by simple peak picking, looking for strong F0 components. It is well known that accuracy of pitch candidates rapidly degrades in the presence of additive noise usually originating from percussive instruments. Therefore in order to increase reliability of the pitch candidates the actual partial tone series of each pitch candidate is trying to be located.

If a peak to a given pitch candidate is found in the spectrum and peak amplitude is higher than a certain threshold the peak position is refined using parabolic interpolation. For each refined pitch candidate a total number of P partial tones are trying to be located at frequency positions being integer multiples of the fundamental frequency as given in Eq. 3.9.

$$F_p = F_0 \times p \; [Hz] \qquad p = 2,3,4 \ldots P \qquad \text{(Eq. 3.9)}$$

Not all instruments generate perfectly harmonic spectra and the partial locations tend to deviate from the ideal positions. In particular string instruments like piano and guitar often used in popular music show this characteristic [Järveläinen99]. Moreover if pitch estimates are inexact, the predicted partial tone frequencies are inexact too and the error increases with partial tone number. Therefore a deviation of 1% F_p from the ideal harmonic partial tone frequency is tolerated to account for that which corresponds to +/-17 cent or a (hypothetical) sixteenth tone. As for the F_0s, the located partial tone frequency- and amplitude estimates are improved using parabolic interpolation. Parabolic interpolation uses only three neighboring FFT bins to estimate the true peak location achieving satisfactory results which drastically reduces the computational load compared to zero padding.

In this way the partial tone series of each pitch candidate is trying to be located in the U_Σ. The estimated harmonic series represented by the corresponding partial tone frequencies and amplitudes are passed to the partial tracking stage. The partial tracker is responsible for the grouping of these estimates over time, connecting individual frame wise estimates to form pitch tracks representing individual note events.

3.3. Pitch Tracking

We consider the F_0 trajectory being of high informative character in the discrimination between vocal and instrumental sounds. Therefore pitch candidates of consecutive analysis frames have to be connected in order to get continuous pitch tracks enabling us to analyze the temporal evolution of pitches over time. Consequently it is essential that pitch tracks correspond to the actual F0 trajectories of the underlying sound sources.

In western music it is common that instruments and singing voice often share the same notes thus pitch tracks might be relatively close in frequency or even cross. This can lead to ambiguous situations for the pitch tracker when pitch candidates of frame N are equally close in frequency to a pitch track that existed in frame N-1. Even in less specific situations as the above mentioned the continuation based on the smallest difference in frequency between consecutive pitch candidates which seems intuitive, might lead to erroneous connections due to the large pitch variability typical to vocal sounds. An example of an ambiguous situation is illustrated in Fig. 3.15.

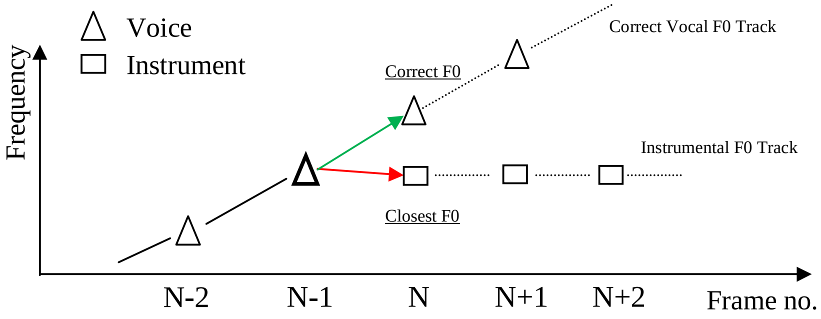


Fig. 3.15.: Simple Partial Tracking: Tracking errors due to large F0 variability of singing voice, if pitch track continuation is based on the closest distance between consecutive pitch candidates

3.3.1. Tracking based on cubic interpolation

To overcome the mentioned possible tracking ambiguity we make use of the history of pitch tracks. More specifically we calculate the expected $\overline{F0}_{exp.}^{Track\,M}(N)$ for each active pitch track in frame N out of the last three F_0's (N-3…N-1) of the corresponding pitch track applying cubic interpolation. That pitch candidate that is closest to the expected $\overline{F0}_{exp.}^{Track\,M}(N)$ is considered the correct pitch for the continuation of track M. This is illustrated schematically in Fig. 3.16.

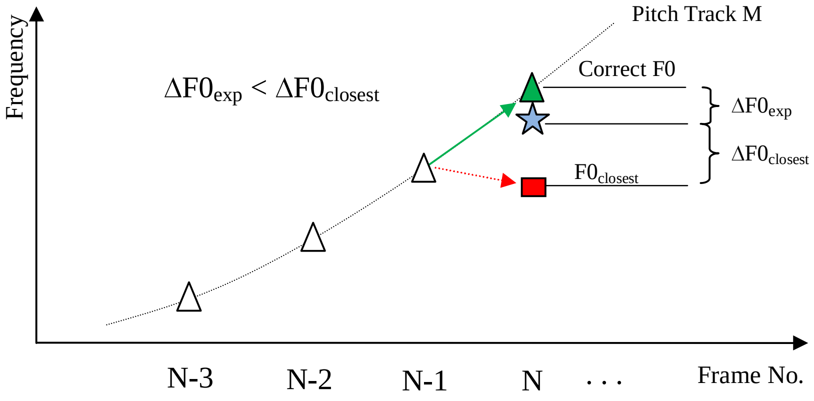


Fig. 3.16.: Correct pitch track continuation based on closest distance $\Delta F0_{\text{-exp.}}$ between the extracted pitch candidate in frame N and the expected $F0_{\text{-exp.}}$ predicted from the last 3 F0 estimates using cubic interpolation

3.4. Post processing

The pitch tracks that have been derived in the described manner are finally restricted to a minimum duration of 50ms and tracks with low F0 salience are discarded. F0 salience is simply calculated as the mean amplitude value for each pitch track and tracks showing values lower than the 50% of the local mean amplitude value are discarded. The local mean is calculated as the mean of mean amplitude values of tracks surrounding (+/-1 s.) the track under test. As can be seen from Fig 3.17 the proposed pitch tracking method is able of capturing the constituent parts of the parallel harmonic sound sources in a polyphonic audio signal from the auditory motivated spectral representation of it.

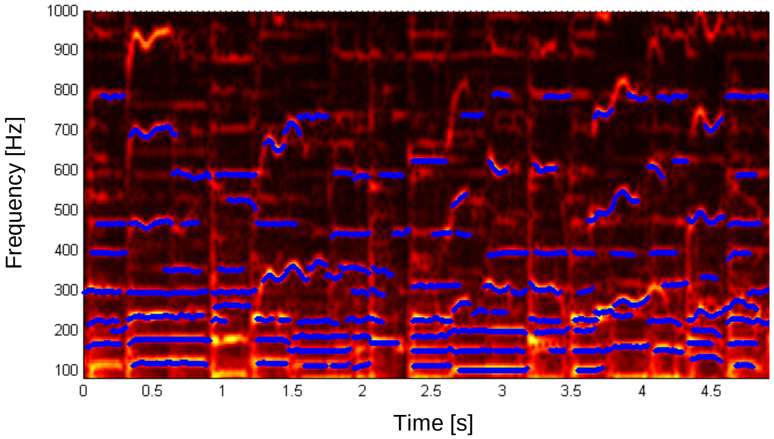


Fig. 3.17.: Pitch Tracks (blue) after post-processing plotted over the spectrogram representation of the corresponding audio excerpt

3.5. Feature Extraction

Since an audio waveform doesn't allow the direct derivation of the note sequence corresponding to the vocal melody, different kinds of signal processing are applied to the data in order to reveal the desired information. In this context the result of such processing is called a feature being a measure of a particular signal property. Features may be calculated for individual signal segments, for the audio track as a whole or from some other kind of representation (e.g.: FFT, Cepstrum, Correlogram) of the data.

The feature extraction stage plays a key role for the succeeding classification process. Often single features fail yielding satisfactory classification accuracy. Usually the use of a feature set (different features in conjunction) results in better performance. The challenge lies in the determination of the optimal feature set being the specific number and kind of features which minimize a previously defined error criterion. In our case that would be to minimize the overall error being the number of frames of vocal pitch tracks erroneously classified as instrumental tracks and vice versa. The optimal feature set is usually found by means of simulation.

In our case all features are derived from the previously extracted pitch tracks. These constitute a mid-level representation of the harmonic content of the signal, holding frequencies and amplitudes of different F0 trajectories of the corresponding partial tones. The features that will be described in the following aim at representing the characteristics of vocal and instrumental sounds that have been explained before (p.17 – p.20 – 1.5 *Singing voice characteristics*).

3.5.1. Feature Description

Feature 1: Salience

This feature aims at representing the absolute strength of the harmonic series and has been used by [Rao08] in the context of main melody estimation. It is calculated for every pitch track as the sum of partial amplitudes of the corresponding harmonic series averaged over its length which can be written as

$$salience = \frac{1}{N_{end} - N_{start}} \sum_{n=N_{start}}^{N_{end}} \sum_{p=1}^{P_{partials}} X\left(k_p, n\right) \qquad \text{(Eq. 3.10)}$$

where n refers to the frame number and k to the corresponding FFT-bin of partial tone p.

Feature 2: Mean Relative Salience (MRS)

Relative spectral energy concentrations at sub-bands have already been exploited in [Tzanetakis04] among other features to discriminate between song segments where the singing voice is present or not. Inspired from that we propose a feature that represents the relative salience of each pitch track compared to the strength of the accompaniment in the frequency range of 300 Hz – 2.500 Hz where the voice is expected to be dominant. It is calculated as the ratio of the spectral energy of a partial tone series and the remaining energy in the mentioned frequency band. Spectral energy is calculated for the partial tone series p=1…P as the sum of squared magnitude values $\left|X(k_p)\right|^2$ of the FFT bin k including the neighboring bins k-1 and k+1 for each partial tone p referred to as $X(\underline{k_P})$ to account for the spectral energy spread due to windowing. Equation 3.11 describes the calculation of the relative salience value for one single frame n. The strength of the accompaniment is calculated as the sum of magnitudes of bins $\underline{K}$ corresponding to the remaining FFT bins which do not correspond to any of the bins related to the partial tones series of track T. The absolute range of $\underline{k}_p$ and $\underline{K}$ is limited as mentioned to the corresponding frequency range of 300 Hz - 2.500 Hz.

$$S_{rel.}(n) = \frac{\sum^{P} X\left(\underline{k_p}\right)^2}{\sum X\left(\underline{K}\right)^2} \; \ldots k_p = \{k-1, k, k+1\}, p = 1 \ldots P \qquad \text{(Eq. 3.11)}$$

For each track the mean relative salience is calculated as the mean value of the frame wise salience values

$$MRS_T = \frac{1}{N_{frames}} \sum_{n=1}^{N_{frames}} S_{rel.}(n) \qquad \text{(Eq. 3.12)}$$

Feature 3: Mean F0

As mentioned, vocal F0s in popular music mainly concentrate on a limited frequency range lying roughly between 100 – 800 Hz. Therefore the mean F0 value for each pitch track is calculated, given as follows:

$$Mean(F_0) = \frac{1}{N_{frames}} \sum_{n=1}^{N_{frames}} F_0(n) \qquad \text{(Eq. 3.13)}$$

Feature 4: SPSD – Summary partial standard deviation

SPSD aims at capturing the variability of partial tones around their mean value which is inherent to singing voice signals. The standard deviation is calculated for the F0 trajectory of every partial tone of a harmonic series and finally summed up which can be written as:

$$SPSD = \sum_{p=1}^{N_{partials}} STD(F_v(p)) \qquad \text{(Eq. 3.14)}$$

With $F_v(p)$ being the frequency trajectory of partial tone p, $N_{partials}$ being the number of estimated partial tones and STD representing the standard deviation.

Feature 5: PSD – Partial standard deviation

Moreover this score has been calculated for individual partial tones (p=1…5) to study the advantage or disadvantage of considering the whole partial tone series in contrast to considering individual partial tone trajectories. The score reduces to Partial Standard Deviation (PSD) which is calculated as:

$$PSD_P = STD(F_v(p)) \qquad \text{(Eq. 3.15)}$$

With $F_v(p)$ being the frequency trajectory of partial tone p and STD representing the standard deviation.

Feature 6: Δ-F0

This feature also aims at capturing the dynamics of the frequency trajectory which is usually lower for instrumental sounds than for singing voice. It has been used in [Goto06] as one of two features for singing voice discrimination. It is calculated as the frequency difference between F0s of consecutive analysis frames n for each pitch track as follows:

$$\Delta F_0(n)_{TrackT} = F_{0_{TrackT}}(n) - F_{0_{TrackT}}(n-1) \qquad \text{(Eq. 3.16)}$$

From the Δ-F0$_v$ vector several features are derived comprising the mean Δ-F0, the standard deviation, variance and maximum calculated for each pitch track.

Feature 7: SDPF – Summary delta-partial tone frequency

This feature is closely related to the previous feature with the difference that ΔF is calculated for all partial tones of the pitch track. In detail, it is calculated as the sum of absolute frequency difference between consecutive analysis frames for every partial tone of a harmonic series which is finally summed up for all partials.

$$SDPF(F_v) = \frac{1}{N_{frames}} \sum_{p=1}^{N_{partials}} \sum_{n=2}^{N_{frames}} abs(\Delta F_v(p,n)) \qquad \text{(Eq. 3.17)}$$

$N_{partials}$ corresponds to the estimated number of partials, N_{frames} is the duration of a tone in frames, $F_v(p,n)$ is the frequency value of partial tone p in frame n of pitch track F_v, and $\Delta F_v(p,n)$ refers to the difference between the frequency value of frames n and n-1.

Feature 8: F0 range(absolute)

To capture the evolution of pitch tracks across frequencies the absolute difference between the maximal and minimal F_0 is calculated for each extracted track being a measure for the range of frequencies that is passed by a track.

$$F0_{Range-absolute}^{Track\ N} = F0_{max}^{N} - F0_{min}^{N} \qquad [Hz] \qquad \text{(Eq. 3.18)}$$

Feature 9: F0 range (relative)

In order to transform this measure to a musical scale the absolute range in semitones is calculated. The cent scale is a relative scale and therefore the frequency difference of the above equation between minimal and maximal F0 becomes a ratio between the corresponding values which equals a difference in the log domain.

$$F0_{Range-relative}^{Track\ N} = 1200 \log_2\left(\frac{F0_{max}^{N}}{F0_{min}^{N}}\right) \quad [cent] \qquad \text{(Eq. 3.19)}$$

The above described features will be the basis for the following vocal classification process. Classifiers usually need reference data which is representative for the given number of different classes and which allows generalization in terms of statistical modeling of the data. New data is subsequently classified based on the training data. Therefore a representative training data set is the key for successful classification. In the following the derivation of the training dataset will be explained in detail and the discriminative power of features derived from it will be investigated by information theoretic means (Fisher’s Ratio, LDA) and by means of simulation (N-fold-cross validation, *Chapter 4 - Evaluation*).

3.5.2. Derivation of the training data set

The discriminative power of individual features or a feature set is not known in advance and has to be verified. First informal tests on solo instrument and vocal recordings showed promising results. So the question was if features derived from the polyphonic mixture audio signal were as informative as the ones derived from the solo recordings.

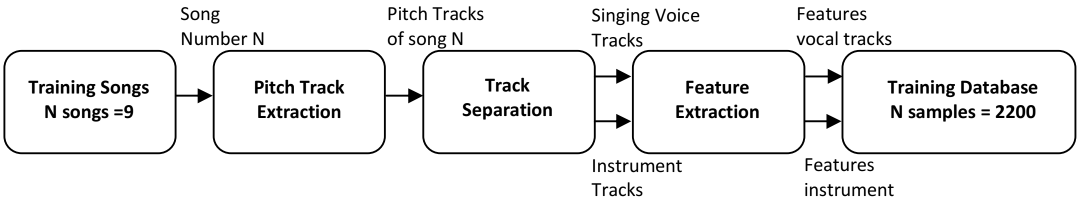


Fig.: 3.18.: Block Diagram - Derivation of the training data set

Therefore we had to generate training data on which a classifier could base its decision on. The MIREX 2005 – Training data base was used to extract the desired training instances. The data base comprises 13 polyphonic song excerpts of which 9 contain singing voice. For the derivation of the training data only the 9 songs containing a male or female singing voice have been considered. The song excerpts are approximately 30 sec each and they come with a manually annotated reference transcription of the singing voice F0 trajectory in Hz with a spacing of 10ms between consecutive analysis instants. This ground truth served as a basis for the separation between vocal and instrumental sounds. Each song excerpt has been analyzed by the developed algorithm and pitch tracks corresponding to the F0 trajectory (or integer multiples of the same) of various concurrent sounds have been extracted as described in Chapter 3.1 – 3.4. On the left hand of the Fig. 3.19 an example of the extracted pitch tracks is given, plotted over the spectrogram of a 5 second sound excerpt. Pitch tracks have been estimated for a frequency range of 100 – 800 Hz and tracks have been restricted to a minimum duration of 100ms.

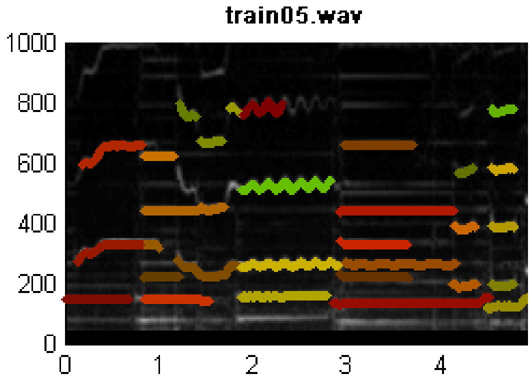


Fig.: 3.19.: Pitch Tracks derived from a song excerpt (5s) of the training data base plotted over the spectrogram representation of the corresponding audio waveform. Pitch tracks have been estimated for a frequency range of 100 – 800 Hz and tracks have been restricted to a minimum duration of 100ms.

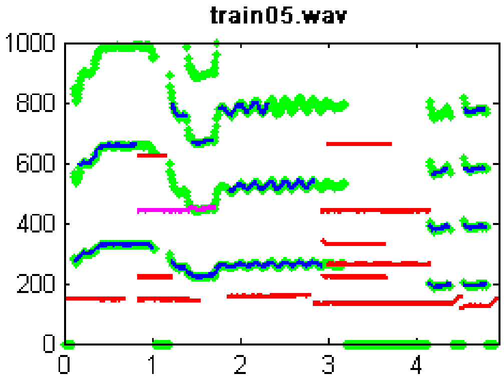


Fig.: 3.20.: Separation of pitch tracks into classes voiced (red) & instrumental (blue) based on the ground truth F0 trajectory (green). Integer multiples of the ground truth are considered as voiced too. Tracks that don’t allow a clear distinction (beteen 10% to 60% overlap with
reference) are not considered (magenta)

Based on the MIREX reference transcription of the F0 trajectory of the singing voice the extracted pitch tracks have been separated into the classes vocal and instrumental. Now tracks that do not deviate more than a quarter-tone (+/-50 cent) from the reference F_0 or an integer multiple of the same for at least 60% of the track duration are considered as vocal corresponding to the blue tracks in Fig. 3.20. Tracks that didn’t allow such a clear distinction (between 10% to 60% overlap with reference) have been excluded to avoid ambiguities in the training data set. In Fig 3.20 pitch tracks corresponding to instrumental sounds are displayed in red.

In the described manner pitch tracks have been derived for every one of the 9 songs of the vocal training data set and pitch tracks have been separated carefully into classes vocal and instrumental based on the manually annotated vocal reference F0 trajectory. Pitch tracks have been restricted to a minimum duration of 50ms. From these pitch tracks the described features have been derived which we consider as our training data base. The total number of samples in the training database is $N_{samples} = 2263$ which splits up to $N_{music} = 1319$ and $N_{voice} = 944$. Before the training data might be used to classify new data instances it has to be verified that the collected data allows discrimination between the two classes.

3.5.3. Discriminative power of the features set

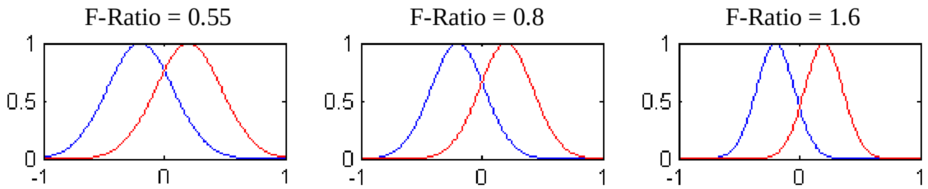
To get an idea of which features might be informative and which not, statistical testing is performed for individual features before feature combinations are tested. The discriminability of the features of the labeled training data has been studied in three different ways. On one hand statistical testing of individual features in terms of calculating Fisher's Ratio has been performed. On the other hand the discriminative power of the whole feature set is investigated applying Linear Discriminant Analysis (LDA). Finally in Chapter 4 the generality of the features set will studied by means of simulation using the leave-one-out method.

<u>Fisher's Ratio</u>

Fisher's Ratio (F_R) is calculated as the ratio between the inter-class variance and the intra-class variance and is given for the classes C_1 and C_2 as follows:

$$F_R = \frac{(\mu_{C1} - \mu_{C2})^2}{\sigma_{C1}^2 + \sigma_{C2}^2} \qquad \text{(Eq.: 3.20)}$$

It reflects the degree of overlap of the two distributions C_1 and C_2 and if the classes are separable in terms of the mean value and the variance of the distributions. To give an idea of the value range, F_R has been exemplified in Fig. 3.21 for 3 distributions showing different degrees of overlap.



$\mu_1 = 0.2 / \sigma^2_1 = 0.15$ $\mu_2 = -0.2 / \sigma^2_2 = 0.15$ overlap: 46,6%

$\mu_1 = 0.2 / \sigma^2_1 = 0.1$ $\mu_2 = -0.2 / \sigma^2_2 = 0.1$ overlap: 37,1%

$\mu_1 = 0.2 / \sigma^2_1 = 0.05$ $\mu_2 = -0.2 / \sigma^2_2 = 0.05$ overlap: 20,6%

Fig.: 3.21.: Fisher's Ratio exemplified for different distributions. Note that the mean values remain the same for all examples while the variance decreases.

Fisher's Ratio has been calculated from the labeled training data for individual features of the feature set described before. Results are given in Table 1.

Fisher's Ratio for single features			
Feature 6 (mean value)	Mean ΔF0	0.754	*
Feature 4	SPSD	0.644	*
Feature 5 (Partial 3)	PSD$_3$	0.635	
Feature 5 (Partial 4)	PSD$_4$	0.634	
Feature 5 (Partial 2)	PSD$_2$	0.631	
Feature 5 (Partial 5)	PSD$_5$	0.631	
Feature 8	Abs. F0 range (Hz)	0.624	*
Feature 5 (Partial 1=F0)	PSD$_1$	0.623	
Feature 6 (standard dev.)	Std ΔF0	0.571	*
Feature 7	SDPF	0.569	*
Feature 3	Mean F0	0.475	*
Feature 6 (maximum)	Max ΔF0	0.464	*
Feature 2	MRS	0.296	*
Feature 6 (variance)	Var ΔF0	0.253	
Feature 1	Salience	0.138	*
Feature 9	F0 range (cent)	0.095	

Table 1: Fisher's Ratio derived from the samples of the training data base (N samples = 2263; N vocal = 944; N non-vocal = 1319) for individual features. The "*" symbol indicates the features of the feature subset which is later used for evaluation

As can be seen single features of the training data base bear discriminability between the two classes to some extent but Fisher's ratios are far too low to reliably predict class affiliation based on one single feature. However, single features showing low F_R's might be valuable when used in combination with other features. Therefore we apply linear discriminant analysis to the whole feature set.

Linear Discriminant Analysis

Linear Discriminant analysis is a technique used in machine learning for dimensionality reduction of a feature space. Based on the class information the method tries to find a linear combination of the present feature set $\overline{\boldsymbol{F}}$ (comprising features $F_{1...}F_N$) of the N dimensional feature space according to

$$Y = w^T \overline{\mathbf{F}} \qquad \text{(Eq. 3.21)}$$

such that separability in the new 1 dimensional feature space Y is maximized. The criterion for maximization is the ratio of between-class scatter and within-class scatter.

Feature Selection using LDA

So by applying LDA we project the existing feature space onto a new features space where classes are in general more separable. Therefore also F_R's will be in general larger for the transformed features than for individual features. This can be used for feature selection by iteratively excluding individual features from the whole feature set, calculating F_R of the transformed feature subset and comparing it to F_R derived from the transform of the whole feature set. In this way features are subsequently excluded from the feature set as long as F_R of the transformed subset increased or at least did not degrade significantly. Those features showing the lowest individual F_R have been the first candidates for exclusion.

It turned out that the features *mean F0, mean relative salience (MRS)* and *mean ΔF0* are the features bearing the highest discriminability when used in combination with respect to the training data set since they resulted in the largest degradation in terms of Fisher's Ratio of the LDA transformed features when excluded from the feature set.

The final feature set contains 9 of the 16 proposed features. Fisher's Ratio for the LDA transformed features set is F_R=3,18 and for the feature subset $F_{R\text{-sub}}$=2,97. Feature distributions for the transformed feature subset are exemplified in Fig. 3.22 and 3.23.

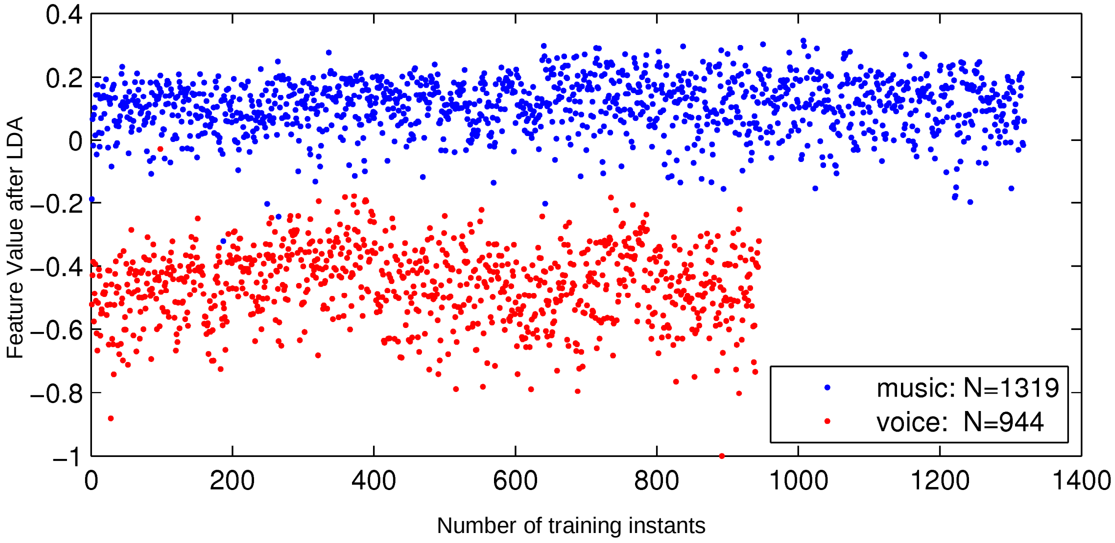


Fig. 3.22: Feature Values after LDA for the two classes voice (red) / music (blue).
Good separation of the training data using the derived feature subset is obviously given.

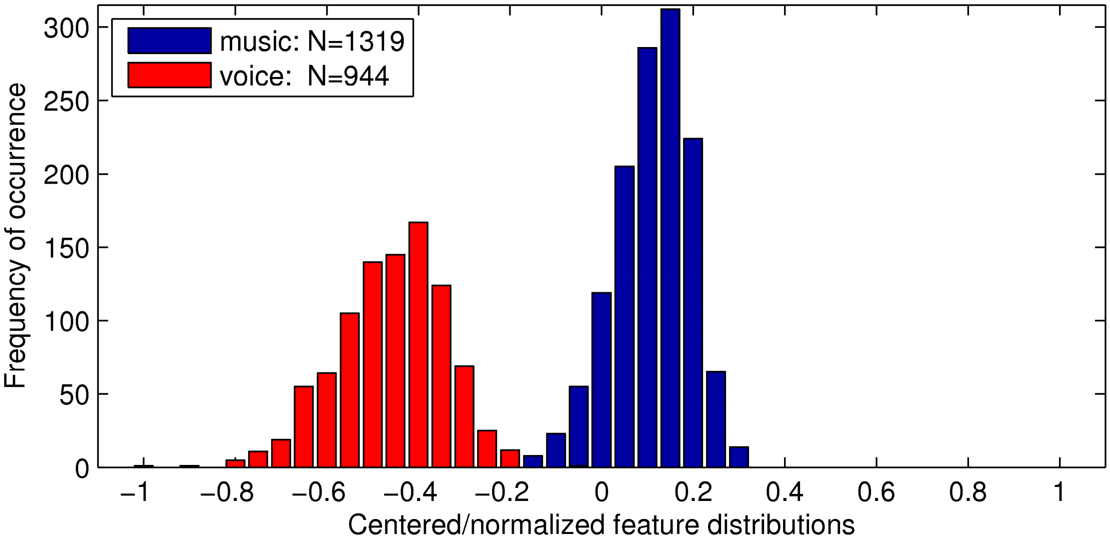


Fig. 3.23.: Distribution of feature values after LDA in terms of frequency of occurrence for the two classes voice (red) / music (blue). Good separation of the training data using the derived feature subset is obviously given.

The discriminative power of individual features of the feature set has been investigated using Fisher’s Ratio. A feature subset has been found using LDA which does not significantly degrade linear separability of the data compared to the whole feature set. Now that we know that the training data can be discriminated to a certain extent based on the derived features the generality of the training data has to be verified in terms of how well class affiliation of new data instants can be predicted correctly based on the training data. This is called validation of the training data which will be done using the N-fold-cross-validation method, described in detail in *“Chapter 4 – Evaluation and Results”*.

3.6. Classification

Classification of new data instances based on the training data is accomplished using the KNN (K-Nearest-Neighbor) classifier. The principle of KNN-classification is that new data instants are classified based on the class affiliation of the K closest training instants as illustrated in Fig. 3.24 for K=3 and K=5. Usually the Euclidean distance is used as a distance metric.

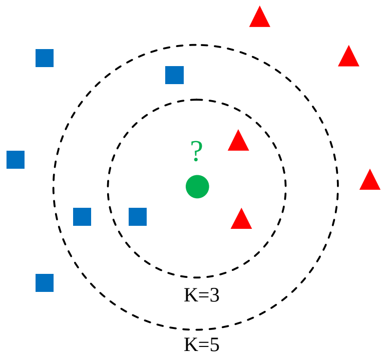


Fig. 3.24.: Principle of KNN classification: A new data instant (green) is classified based on the class affiliation of the K closest training instants (Class1: blue / Class2: red), closest neighbors shown for K=3 and K=5

The KNN classifier has been chosen for its simplicity of implementation and because it allows to easily adjust the sensitivity of the classifier to the training data by varying the number K of neighboring data instances which strongly affects the structure of class boundaries [Duda01]. The choice of K usually depends on the number of training instances N_{Train} and a general guideline is to select K according to $\sqrt{N_{Train}}$. If the classes are well separated in the feature space a smaller value for K can be selected.

Another reason for choosing the KNN classifier is that it has established as reference for pattern recognition. Compared to more elaborate classifiers the KNN might not always perform best but it sets a baseline for achievable classifier accuracy.

A visual example of the classification results for a 8s song excerpt of the training database is given in the Fig. 3.25. Pitch estimation has been performed between 100 Hz and 800 Hz and track duration has been restricted to be at least 100ms. Features of the song under test previously derived for each training song are of course excluded from the training set before classification to avoid that tracks are classified based on training data derived from the same song. Tracks classified as voice are displayed in red and instrumental tracks in blue. The reference F0 trajectory is displayed in green.

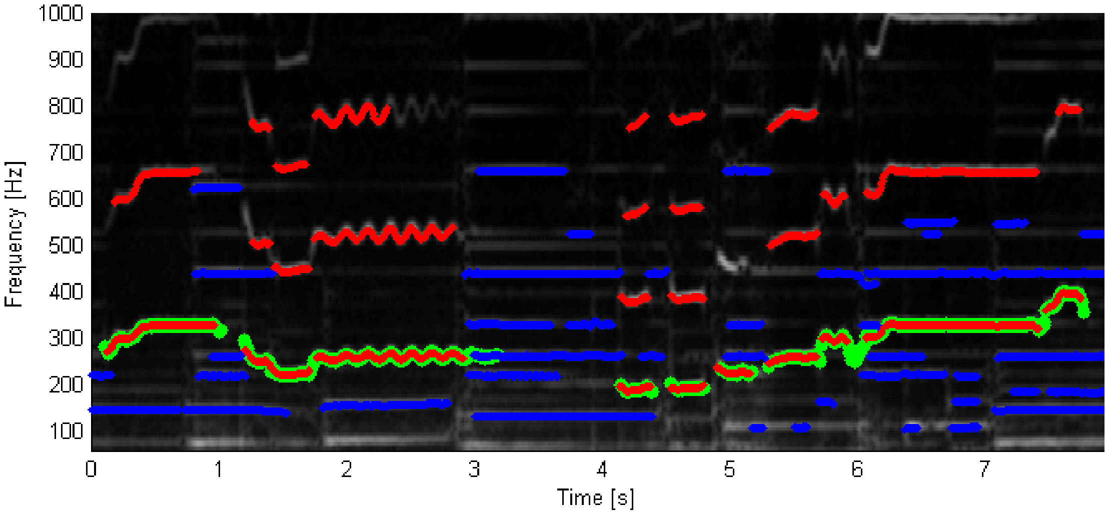


Fig. 3.25.: Successful classification of pitch tracks into classes voice (red) and instrumental (blue) based on the training data set containing N=2000 training instances. The reference F0 trajectory is displayed in green. Number of neighbors of the KNN classifier was set to K=11.

As can be seen the training data allows successful classification of pitch tracks into the classes voice and instrumental. Of course perfect discrimination as in the example above is not always achievable for different reasons. On one hand the more notes are played simultaneously the more difficult is the extraction and separation of individual pitch tracks. On the other hand not all singers show the expected vocal characteristics in such a pronounced way which can lead to misclassification.

3.7. Final Pitch streaming

As described before in *“3.5.1. Derivation of the training data set”* also tracks found at integer multiples of the reference vocal F0’s are considered as training instances for the class voice. This is the only way to maintain class separation without having to restrict the search range for F_0’s too much. Consequently there might be multiple tracks overlapping in time after classification usually corresponding to the F_0 and the first few partials that fall within the frequency range of interest as can be seen above in Fig. 3.25 . Therefore the output of the classifier has to be post processed and the tracks classified as voice have to be reduced to one final vocal track. Parallel pitch tracks are compared on the basis of the summary mean spectral amplitude (SMSA) of the first 3 partial tones. For overlapping tracks always the one showing a lower SMSA value is discarded. In Fig. 3.26 the final vocal pitch stream derived in the described manner for the above example is shown in red together with the reference F0 in green plotted over the spectrogram representation of the 8s song excerpt.

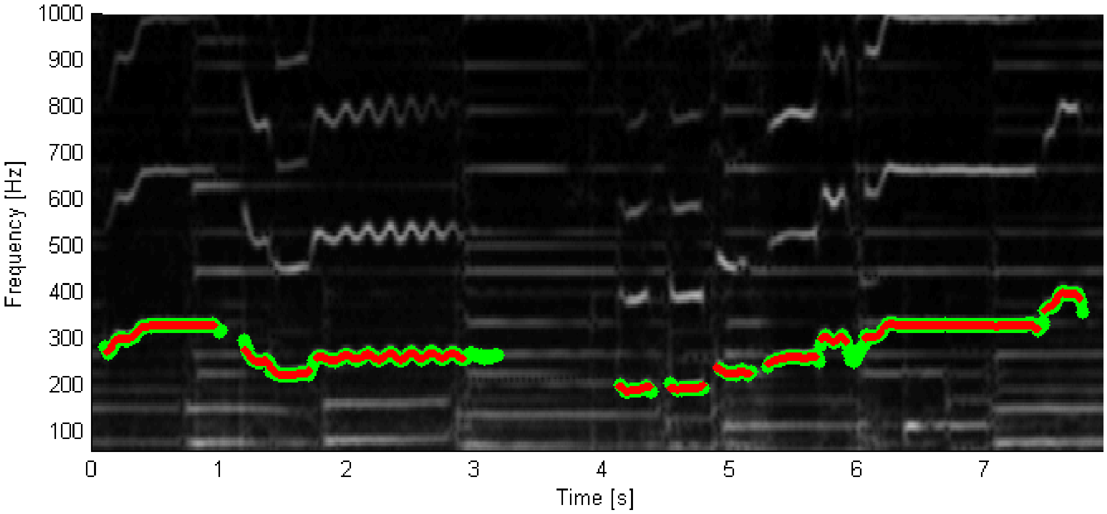


Fig. 3.26.: Final vocal pitch track (red) after reduction and reference F0 (green) plotted over the spectrogram representation of the corresponding song excerpt.

3.8. Short-time energy based track grouping

It has been observed that one source of errors with respect to classification accuracy is the misconnection of consecutive spectrally close note events to one pitch track. Therefore we implemented forced track separation based on the short time energy (STE). The STE is calculated from the audio waveform every 1.5 ms for frames of 5.8ms. Due to the small hop size and frame size the resulting function will reveal the moments of highest short time signal energy. In typical music recordings this increase would usually be caused by the base drum or the snare drum. So by forcing tracks to be separated at moments of high STE the pitch tracks are automatically synchronized with the beat of a song. To find possible split points we apply peak picking to the STE function. Inter-peak intervals are restricted to be at least 400 ms apart from each other corresponding to a maximal expected song tempo of 150 BPM. The detected peaks are further restricted to be at least higher than two times the mean value of the STE function. The derivation of split points is exemplified in Fig. Pitch tracks and STE function (blue) are plotted over the spectrogram of the corresponding song excerpt. The applied threshold corresponds to the red line and split points are indicated as green asterisks on the STE function and as red asterisks at 0 frequency position. The weakness of STE function is that it works well if there are strong percussive elements (base drum, snare drum) present in the audio signal. For songs with no or few percussion the peaks in the STE will be less pronounced and correct separation will be more difficult.

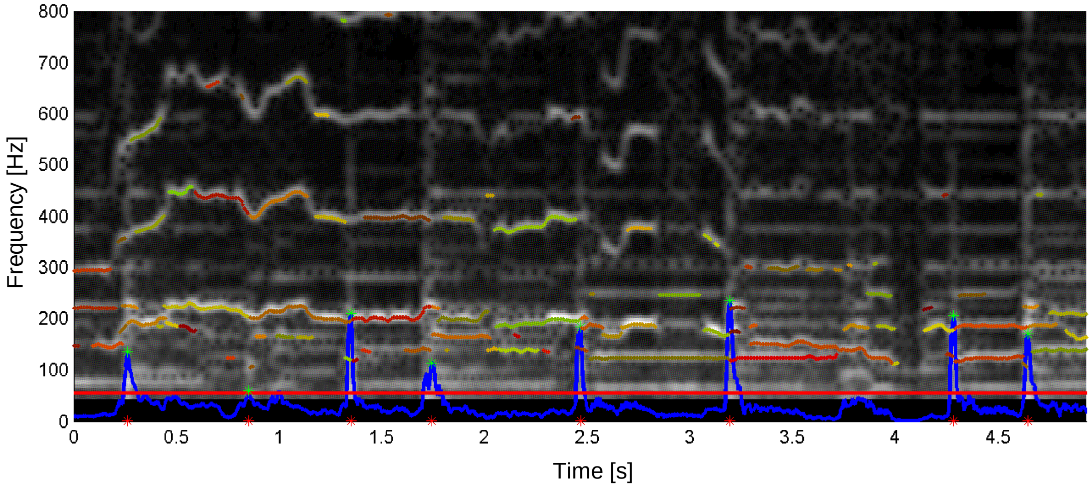


Fig. 3.27.: Forced pitch track separation based on short time energy (blue): split points are indicated as green asterisks on the short time energy (STE) function and as red asterisks at 0 frequency position, applied threshold for peak detection = 2*mean(STE) (red), actual values of the STE function have been increased for better visibility

3.8. MIDI quantization

The raw F_0 trajectory estimated in the described manner has no musical meaning and therefore has to be converted into a distinct note sequence. This may effectively be done by MIDI quantizing the F0 track according to:

$$MIDI\ Note\ Number = 69 + 12 * log_2\left(\frac{F}{440}\right) \qquad \text{(Eq. 3.22)}$$

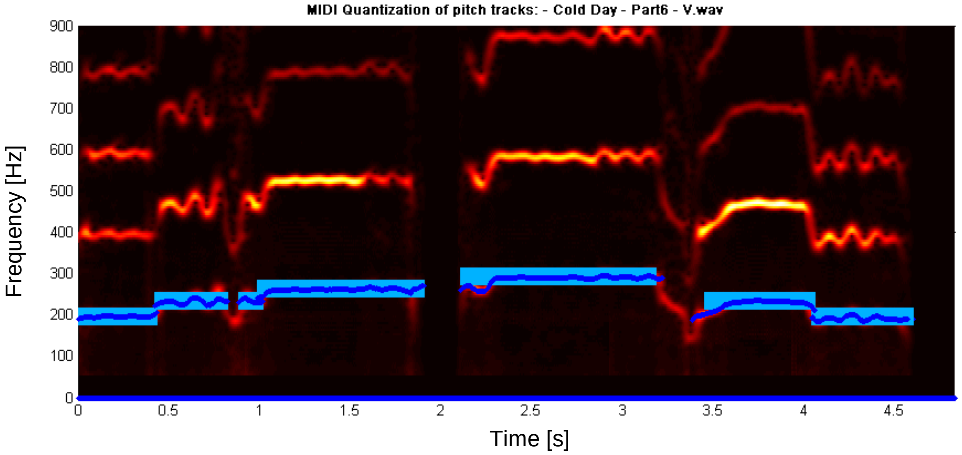


Fig. 3.28.: MIDI quantization of the pitch trajectory of a solo vocal performance (dark blue) into discrete note events (light blue)

However, direct MIDI quantization of the vocal melody doesn’t always yield satisfactory results due to large frequency modulation typical to singing (see Fig. 3.29 and Fig. 3.30).

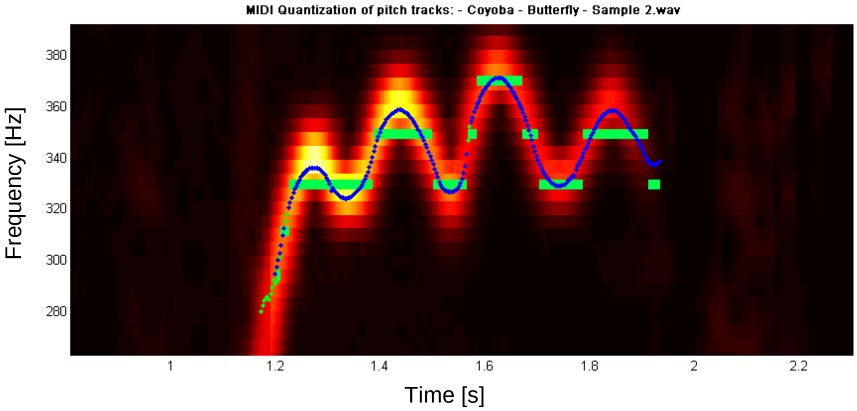


Fig. 3.29.: Erroneous MIDI note quantization (green) due to large frequency modulation caused by vibrato

Therefore F0 tracks are filtered by a 10 Hz moving average filter which corresponding to the largest frequency measured for vibrato. The smoothed pitch tracks are then subjected to MIDI quantization. Still there are spurious errors which can be reduced by restricting the duration of notes to be larger than a certain minimum length since voice sounds tend to be continuous.

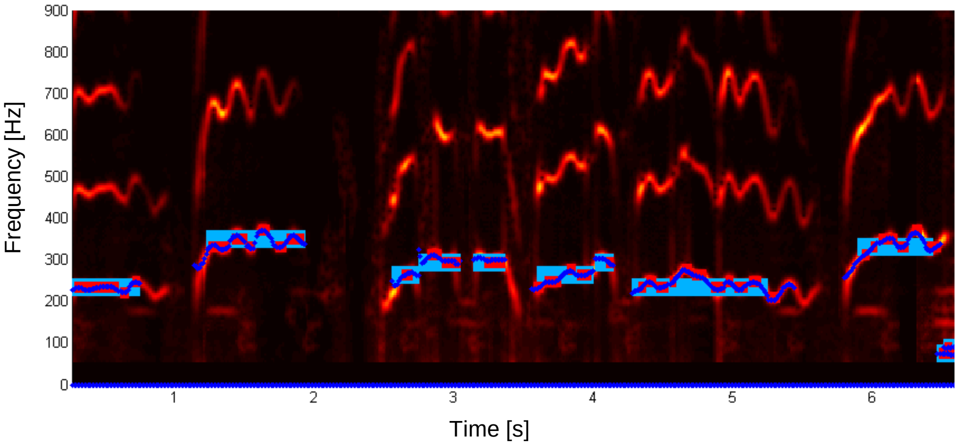


Fig. 3.30.: Erroneous MIDI note quantization (red) due to large vibrato and quantization after moving average filtering (light blue) of the F_0-trajectory (dark blue)

3.8. System output

The output of the system is a MIDI file which to our point of view represents a more general description of the musical content than a classical score. The fact that plenty of software programs related to music processing and music content analysis make use of MIDI suggests the usage of it as intercommunication file type in order to facilitate further use of the extracted musical information. Moreover one can easily listen to a MIDI file or create a musical score using music notation software. In contrast to that only trained musicians are able to read and reproduce music from a score written on paper.

In addition the exact F0 track of the vocal melody is written to a text file containing the time stamps of individual analysis instants and the corresponding F0 values.

Chapter 4 – Evaluation and Results

The proposed approach towards singing voice detection has been evaluated using the training database of the MIREX 2005 melody transcription contest comprising 13 songs excerpts of an approximate length of 30 seconds each. The songs cover different genres and 9 of them contain vocals while 4 are MIDI songs only. Since we focus on the detection of the singing voice the 4 MIDI songs have been excluded from the test data set and for evaluation only the 9 vocal song excerpts have been considered. Thus given results correspond to mean values of the mentioned vocal dataset. The database comes with a reference transcription of the F0 trajectory of the vocals in terms of a text file containing the time instants and the corresponding F0's in Hz which have been annotated manually. The reference time grid has a spacing of 10ms. According to the guidelines of the MIREX melody extraction contest F_0 estimates deviating less than a ¼ tone from the reference are considered as correct estimates. Since F_0's of musical notes are logarithmically spaced in frequency, the range of +/- ¼-tone in Hz is dependent on the actual note. Therefore all pitch estimates are converted to the cent-scale which linearizes the logarithmic nature of F_0s of musical notes. There the spacing of a semitone always corresponds to 100 cent independent of the actual note. So pitch estimates are considered correct if they deviate less than +/-50 cents from the reference.

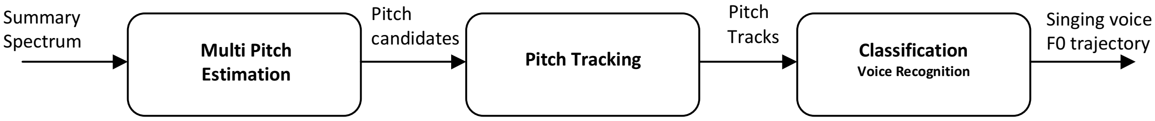


Fig. 4.1.: Main Stages of the proposed singing voice F0 estimation method

The main processing stages of the proposed method (see Fig.4.1), namely *multi pitch estimation*, *partial tracking* and *classification* (singing voice recognition) have been evaluated separately. First of all "raw singing voice F0 estimation accuracy" is calculated from the frame wise pitch estimates. It reflects the ability of the pitch estimation stage to correctly estimate the singing voice F0 trajectory in the polyphonic mixture signal.

Secondly a "pitch tracking accuracy" is calculated showing how much information is lost when individual frame wise pitch candidates are connected to form continuous pitch tracks restricting the length of pitch tracks to be at least larger than a specified minimum duration (in our case 50ms). Finally the classifier performance is evaluated in terms of correct separation between F0 tracks corresponding to vocal and instrumental F0 trajectories. Due to the hierarchical structure the performance of individual stages is strongly interdependent.

In Table 2 the algorithm settings that have been used for evaluation are summarized.

ALGORITHM SETTINGS (used for evaluation)			
Settings Pitch Estimation		**Settings Pitch Tracking**	
Sampling frequency	11025 Hz	N partials tracking	6
Frame Size	92,88 ms	N tracking tolerance	5
Hop Size	5,80 ms	Minimum track duration [s]	0,05
Z-padding factor	2	Max allowed frame-to-frame chirp rate	2,5 % F_0
Pitch estimates per frame	1 - 10	**Classification**	
Minimum Frequency	98 Hz	N samples Training	N voice = 944 / N music =1.391
Maximum Frequency	800 Hz	KNN classifier	K=11 / K=31 / K=51
		Feature subset	According to features in Table 1 indicated by "*"
		Training/Evaluation Data Set	
		No of songs.	9
		Mean duration [s]	30

Table 2: Algorithm settings used for evaluation

4.1. Raw Singing Voice F0 Estimation Accuracy

As mentioned this score reflects the ability of the pitch estimation stage to correctly estimate the singing voice F0s in the polyphonic mixture signal and is computed from the frame wise pitch estimates.

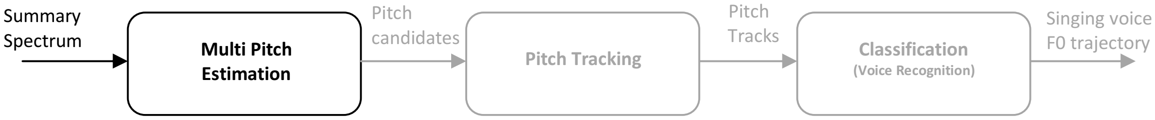


Fig. 4.2.: Main stages of the proposed singing voice F0 estimation method, highlighted block is subject of evaluation

The number of estimated pitches per frame strongly affects the ability to recognize the singing voice F0 among others and has therefore been studied for the range of $N_{MPE} = 1 \ldots 10$ $\frac{pitch\ estimates}{frame}$. Singing voice F_0 estimation accuracy ($F0_V$-acc.) is calculated as the ratio between number of correctly estimated singing voice F0s and the total number of frames containing singing voice.

$$Raw\ Vocal\ F0\ estimation\ accuracy\ = \frac{N\ correctly\ estimated\ vocal\ F0's}{N\ voiced\ frames}\ [\%]$$

A frame is considered correctly estimated if any of the N estimated pitches is within +/- ¼-note to the reference F_0.

4.1.1. Performance of the individual MPEs

As we make use of two MPEs, one based on the calculation of the salience function (referred to as MPE1) and the other based on simple peak picking (MPE2-PP), the performance of the two is evaluated individually and also for a combined approach, where estimates of both MPEs derived from the summary spectrum are considered simultaneously described by the block diagram in Fig 4.3.

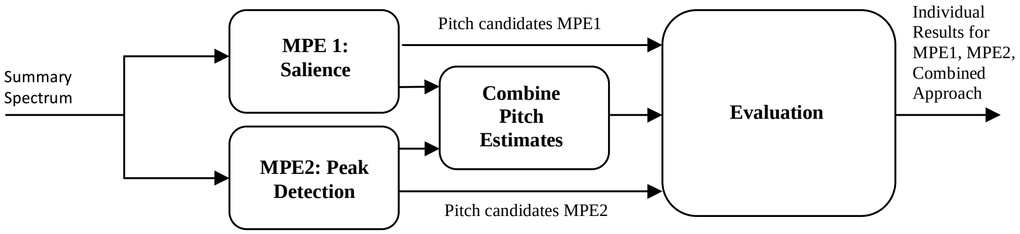


Fig. 4.3.: Block diagram of the evaluation framework for individual MPEs and the combined approach

The following diagram summarizes performance of the two individual MPEs and the combined approach. The x-axis corresponds to results for various numbers of pitch estimates (N=1…10) per frame. Each box plot was derived from the mean performance for individual songs of the training data base (N_{songs}=9).

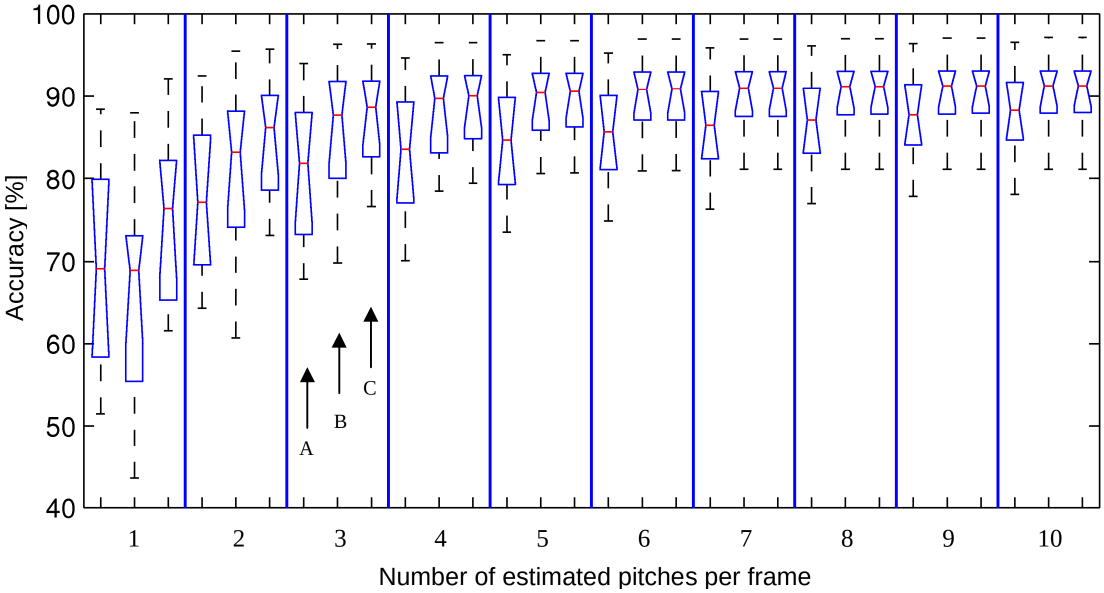


Fig. 4.4.: Variance of singing voice F_0 estimation accuracy for MPE1 (A), MPE2 (B) and the combined approach (C) displayed for different numbers of estimated pitches per frame. All pitch estimates have been derived from the summary spectrum

As can be seen from the diagrams in Fig. 4.4 for low numbers of estimated pitches (N<4) the combined use of the pitch estimates from both MPEs results only in a small performance gain. Moreover it can be seen that for nearly all numbers of estimated pitches, simple peak picking based MPE (B) outperforms the salience function based one (A). The reason for this is probably the iterative estimation and cancellation procedure applied in MPE1. So for higher numbers of estimated pitches (N>4) single use of the peak-picking based MPE (MPE2) seems to be the adequate and sufficient choice. Apart from that it can be assumed that 100 % pitch accuracy cannot be reached by the used pitch estimators, not even by a further increase of the number of estimated pitches. Maximum average singing voice F0 accuracy is 90.4% for N=10 estimated pitches per frame (see also Table 4). This is an important finding since the overall accuracy of the vocal transcription algorithm is limited by the accuracy of the pitch estimation stage. A visual inspection revealed that strong F_0 variations of the singing voice especially at the beginning of notes are hard to track when coinciding with strong instrument or percussion onsets. It has been demonstrated that performances of the combined approach and peak-picking approach (MPE2) are very similar for numbers of frame wise pitch estimates larger than 3. Therefore in the following examples results will only be given for MPE2.

4.1.2. Auditory motivated summary spectrum vs. magnitude spectrum

Since we apply auditory motivated processing to the input data before pitch estimation is performed, we wanted to study the influence of this processing on the pitch accuracy compared to the accuracy of pitch candidates that have been derived from unprocessed magnitude spectra. Therefore the performance of MPE2 applied to simple magnitude spectra has been additionally evaluated. Fig. 4.5 holds the results for accuracy of singing voice F0 estimation performed by MPE2, one time applied to the auditory motivated summary spectrum and one time applied to simple magnitude spectra. Results are shown for different numbers of estimated pitches.

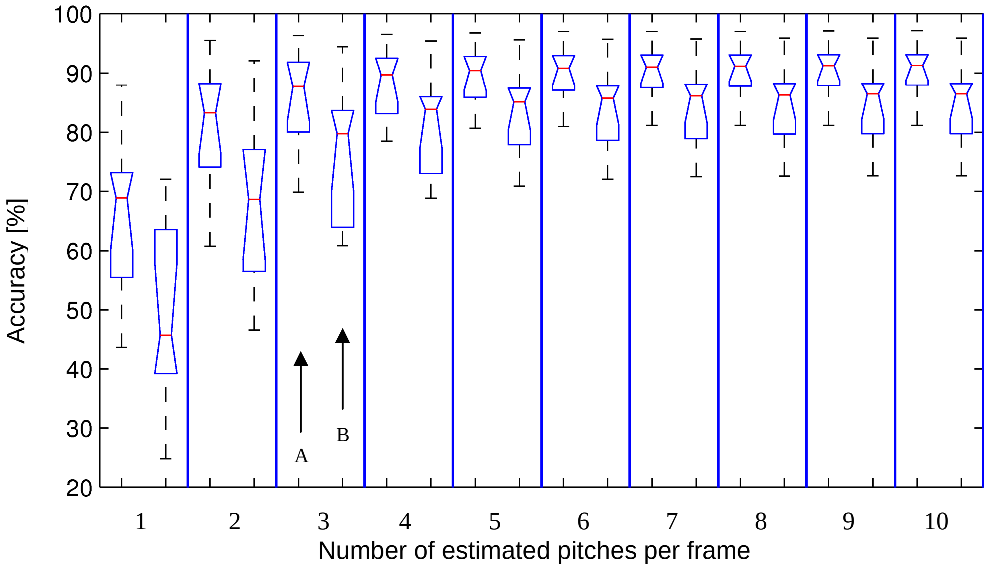


Fig. 4.5.: Influence of the auditory preprocessing on the singing voice F0 estimation: Variance of estimation accuracy for pitch estimation performed by MPE2 for different numbers of estimated pitches. For each number of MPEs two boxplots are given, the first corresponding to the use of the auditory summary spectrum (A) and the second corresponding to the use of simple magnitude spectra (B) for multi pitch estimation. Displayed metrics: median (red), quartiles (blue bottleneck), minimum / maximum (dotted)

It should be noted that displayed boxplots in Fig. 4.5 for MPE2 (A) correspond to the values in Fig. 4.4 (B) which is not apparent dur to different scaling of the y-axis. The results demonstrate that the use of the summary sub-band spectra U_Σ for pitch estimation instead of mere magnitude spectra results in a significant performance gain. The gain is about 5% for higher numbers of estimated pitches (N>4) which justifies the increase in computational complexity (70x, since FFT's have to be calculated for every auditory channel, $N_{channels}$= 70, instead of 1 FFT per frame, neglecting filtering operations to derive the sub-band signals). Moreover the variance in pitch accuracy is smaller for $U_{\Sigma.}$

4.2. Pitch tracking accuracy

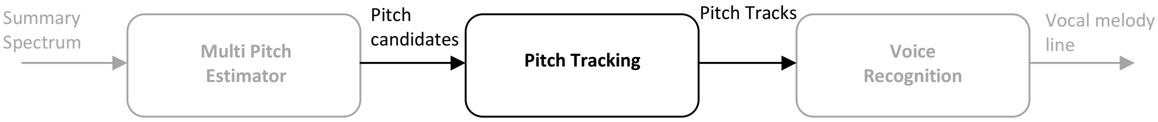


Fig. 4.6.: Main stages of the proposed singing voice F0 estimation method, highlighted block is subject of evaluation

As explained the partial tracking score reveals the average loss of information when individual frame wise pitch candidates are connected to form continuous pitch tracks restricting the length of pitch tracks to be at least larger than a specified minimum duration (in our case 50ms) and discarding low amplitude pitch candidates before tracking. In the following diagram the accuracy of the pure pitch candidates of the MPE2 compared to the accuracy after pitch track formation is displayed. The x-axis represents the performances for different numbers of estimated voices. Given results correspond to mean values of evaluation results for individual songs. In the final comparison also the variance of the pitch tracking accuracy will be given.

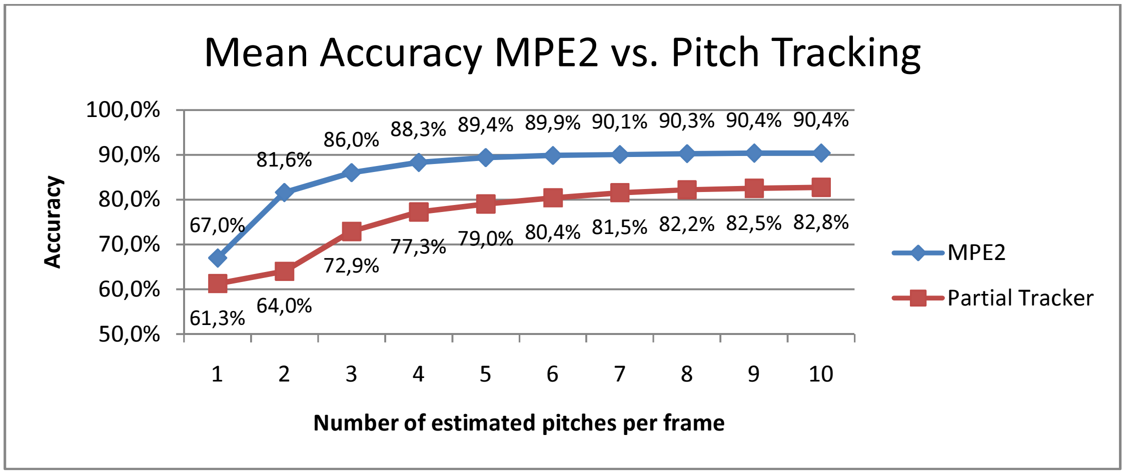


Fig. 4.6.: Comparison of mean accuracy of MPE2 and the following Pitch Tracking stage

As expected from the hierarchical structure the accuracy of pitch tracks increases with higher accuracy of the pitch estimates. It has been observed, that the accuracy of the pitch tracker can be increased in principal by allowing very short tracks too but this has been found to negatively affect the following classification process. Therefore pitch tracks have been restricted to be at least of 50ms length. The absolute maximum average accuracy of the partial tracker is 82.8 % achieved for 10 estimated pitches per frame. The relative tracking accuracy with respect to the accuracy of MPE2 is 91.6 %.

4.2.1. Tracking based on cubic interpolation

The difference between tracking based on cubic interpolation compared to simple tracking in terms of mean tracking accuracy is very low. The highest improvement compared to simple tracking is about 0.3% on average and the highest improvement for individual songs is about 0.9%, achieved when 10 pitches are estimated per frame (see Table 4, p.71). In an earlier implementation where pitch candidates have not been restricted based on the corresponding amplitude values before tracking, higher differences between the two implementations could be observed. This might be explained by the fact that the more pitch estimates per frame the higher the possibility for misconnections. Another reason for the little improvements compared to simple tracking might be the small hop size (5.8ms) used which also facilitates the correct continuation of pitch tracks. For larger hop size the improvements of the proposed tracking algorithm compared to a simpler one are expected to be higher.

4.3. Classifier Performance

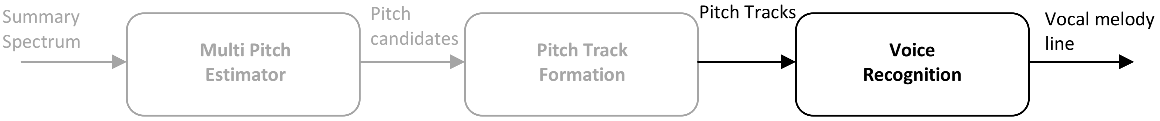


Fig. 4.7.: Main stages of the proposed singing voice F0 estimation method, highlighted block is subject of evaluation

The ability of the classifier to correctly separate pitch tracks into classes voice and instrumental based on the training data has been evaluated using N-fold-cross-validation. It is a technique to test classifier performance in cases where the same data set is used for training as for validation. The procedure is visualized in the block diagrams of Fig. 4.8 and Fig. 4.9. In a first step features are derived from the training data and training instances are separated into classes according to the provided label information (explained in 3.5.2), this is referred to as supervised learning. The generality and discriminability based on the derived training instances is assessed in a second step when new data has to be classified.

Supervised Learning

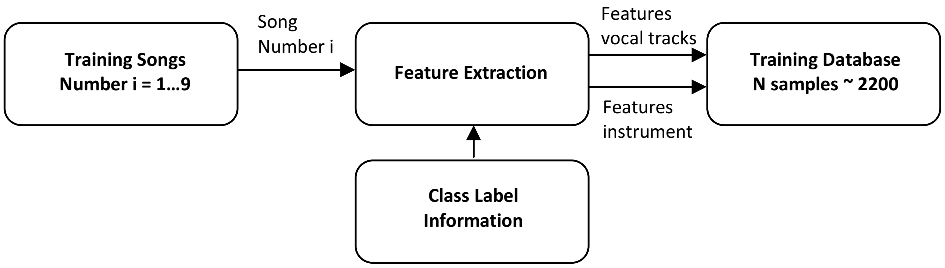


Fig. 4.8.: Derivation of labeled training data referred to as supervised learning exemplified for our case

N-fold cross validation

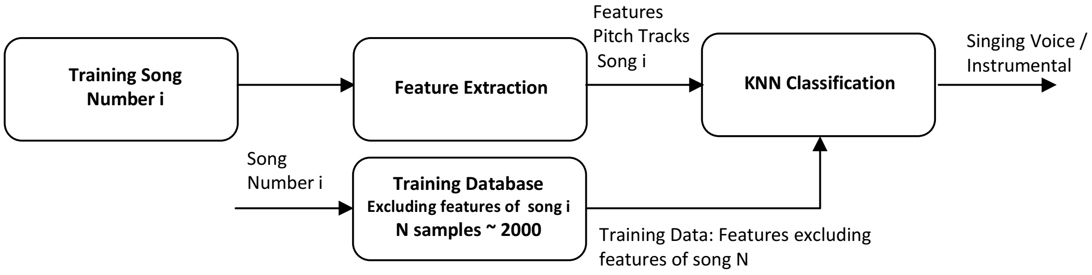

Training Song Number i
Feature Extraction
Features Pitch Tracks Song i
KNN Classification
Singing Voice / Instrumental
Song Number i
Training Database Excluding features of song i N samples ~ 2000
Training Data: Features excluding features of song N

Fig. 4.9.: Block Diagram of the N-fold-cross validation procedure

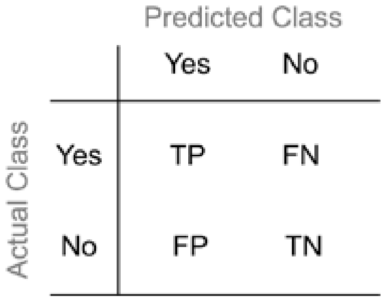
As can be seen from the block diagram pitch tracks of the song under test (1 out of 9 training songs) are classified based on the features of the training data base excluding the features derived from the song under test which is done for every song. In this way it can be avoided that the same data instances are used for training as for classification.

The results of the 9-fold cross validation are given in two ways. First the summarized results of mean classifier performance are given in terms of the confusion matrix for the best performing setting. Then variance of classifier performance for different numbers of estimated pitches per frame will be given. Since the reference vocal F0 transcription is given frame wise, evaluation will also be performed on a frame level. So each pitch track contributes to the evaluation result according to its duration in frames.

Confusion matrix

The confusion matrix summarizes classifier performance. The given metrics are the "True Positives" (TP) being the ratio between the number of correctly classified vocal frames and the total number of frames containing singing voice, the "False Negatives" (FN) being the ratio between the number of frames that have not been recognized by the classifier but actually contain singing voice and the total number of frames containing singing voice, the "False Positives" (FP) being the ratio of the number of frames that have been identified as voiced but actually don't contain voice and the total number of unvoiced frames and finally the "True negatives" being the ratio between the number of unvoiced frames that have been rejected correctly and the total number of unvoiced frames.

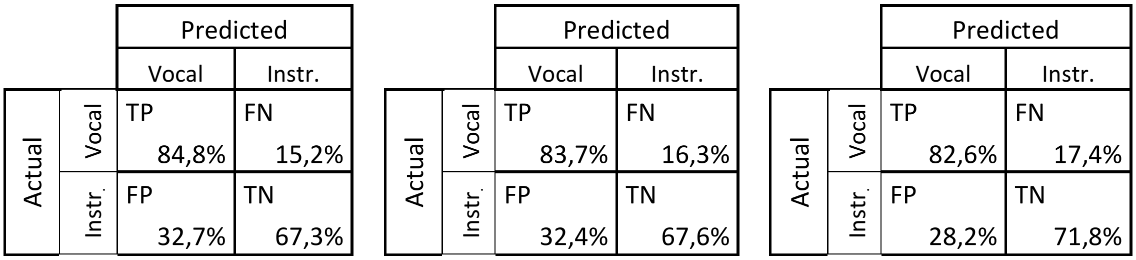
In order to demonstrate effective classifier performance frames where the singing voice has not been estimated correctly have been excluded before score calculation. In this way the TP and FN as TN and FP sum up to 100%. Later on results will also be given with respect to absolute accuracy (see Fig. 4.11, Fig. 4.12, and Table 4).



		Predicted Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

Fig. 4.10.: Confusion matrix with corresponding metrics

As discussed earlier the number of neighbors used in KNN classification strongly affects the classifier performance. Usually the number of neighbors K is chosen according to the square root of the number of training instances N. In our case that would be K~47 for N=2236 training instances. The KNN classifier has been evaluated for 3 different values of K, namely K=11, K=31 and K=51 and a number of 10 pitch estimates per frame.



		Predicted	
		Vocal	Instr.
Actual	Vocal	TP 84,8%	FN 15,2%
	Instr.	FP 32,7%	TN 67,3%

K = 11

		Predicted	
		Vocal	Instr.
Actual	Vocal	TP 83,7%	FN 16,3%
	Instr.	FP 32,4%	TN 67,6%

K=31

		Predicted	
		Vocal	Instr.
Actual	Vocal	TP 82,6%	FN 17,4%
	Instr.	FP 28,2%	TN 71,8%

K=51

Table 3: Relative Performance Values of the KNN classifier for different numbers of neighbors K

From the confusion matrices in Table 3 it can be observed that for higher numbers of K the false positive rate decreases. Simultaneously also the true positives decrease with increasing number of neighbors. Obviously there is a tradeoff between correct classification and correct rejection of samples. So maximizing only one of the two values usually degrades performance of the other. In the following we use a value of K=11 which resulted in the highest TP rates.

In order to demonstrate the interdependence between correct classification (TP) and correct rejection (TN) the performance values for individual songs are summarized in Fig. 4.11 using box plots. Each column holds the results in terms of TP (A) and TN (B) in percent for various numbers of estimated pitches.

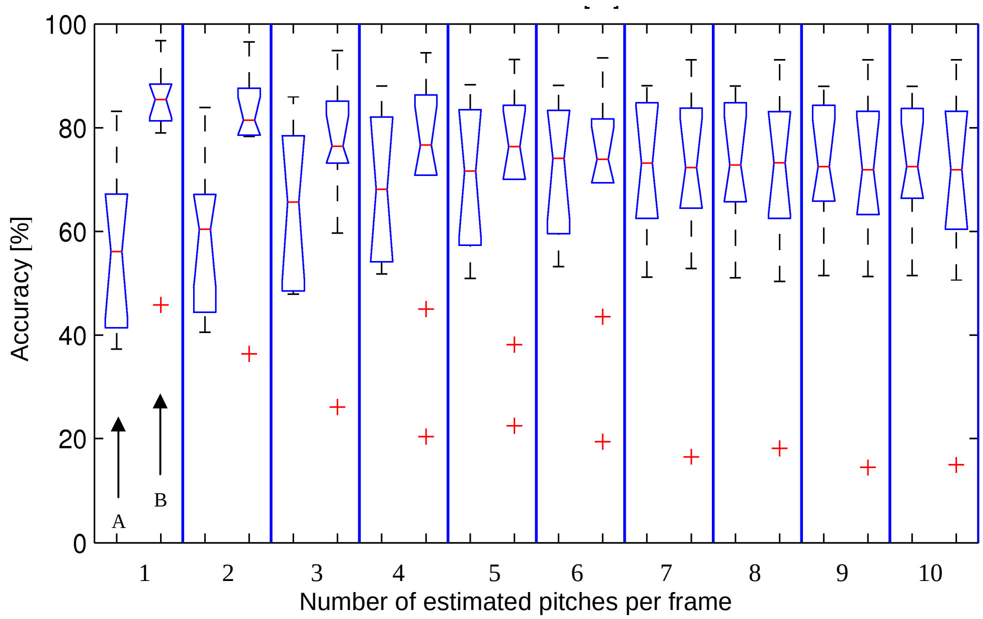


Fig. 4.11.: Absolute classifier performance in terms of True Positives (A) and True Negatives (B) for different numbers of estimated pitches, classifier settings KNN (K=11). Displayed metrics: median (red), quartiles (blue bottleneck), minimum / maximum (dotted), “+” outliers larger than 1.5 times the interquartile range

Accuracy is defined as the sum of TP and TN divided by the total number of frames. Therefore the balance between both values is important to attain high classifier accuracy. As can be seen from Fig. 4.11, balance between TP and TN is only given for higher numbers of estimated pitches. It is clear that TN decreases with increasing TP and increasing number of estimated pitches since the higher the number of found pitch tracks, the higher the probability that a tracks really capture the singing voice. Simultaneously the probability for misclassification increases the higher the number of parallel pitch tracks. There are two outliers indicated by red “+” in Fig. 4.11 for TN which can be important to detect possible weaknesses of the algorithm. From visual and acoustic inspection of the corresponding audio files it has been found that at least one of the outliers is caused by a low ratio between singing voice signal and music accompaniment which might explain the decreased ability of the classifier to correctly reject pitch tracks corresponding to instrumental sounds. The second outlier corresponds to a song with strong instrumental sounds (Piano & Guitar) but without percussion. It has been found that for the mentioned song the short time energy based track separation mechanism fails to separate the tracks of the music accompaniment thus leading to misclassification and a very low TN value.

Further it can be seen that there are no outliers for the TPs and the average correct recognition of singing voice signals for higher numbers of estimated pitches (N>4) is somewhere around 70 %. Finally the overall accuracy of the individual stages will be compared.

Final comparison of accuracies

A final comparison of the accuracies of the individual algorithm processing stages is given in Fig. 4.12 for different numbers of estimated pitches. As can be seen variance of accuracies increases from stage to stage which reflects the interdependence between the stages. The higher the variance in accuracy of the first stage (MPE) the larger the variance of the following stage (Partial Tracker). However, the proposed MPE seems to be able of correctly estimating singing voice F0s to a large extent considering (for N > 3 $\min_{acc}$ ~ 80%, $\max_{acc}$ ~ 97 %, median_{acc}~ 90%). The accuracy of the partial tracking stage decreases due to discarding tracks shorter than 50ms. Maybe this value is too restrictive and should be lowered to attain higher partial tracking accuracy at the cost of higher FPs. Finally accuracy of the classifier varies a lot for individual songs ($\min_{acc}$~40% $\max_{acc}$ ~88%). The main errors (Table 4) have been found to occur for voices that only show little F0 variability which is one of the main characteristic that the described features try to capture.

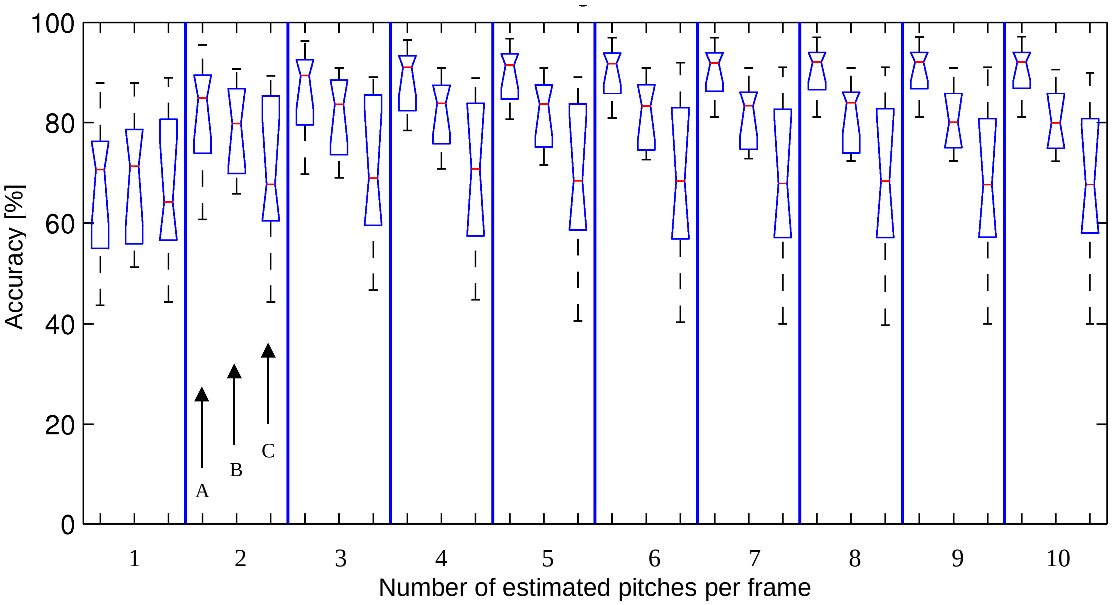


Fig. 4.12.: Comparison of absolute accuracies of individual processing stages. Displayed results: MPE2 (A) vs. Pitch Tracking (B) vs. Classifier Accuracy (C, Settings KNN: K=11)
Displayed metrics: median (red), quartiles (blue bottleneck), minimum / maximum (dotted), “+” outliers larger than 1.5 times the interquartile range

Song Number	Pitch Accuracy [%]	Tracking Accuracy (cubic) [%]	Tracking Accuracy (simple) [%]	Classifier Acc [%]	TP [%]
1	96,9	90,8	90,5	89,8	88,1
2	87,9	79,1	78,6	66,3	63,4
3	81,1	74,4	73,5	55,1	51,6
4	88,6	75,9	75,6	69,2	67,8
5	97,1	92,0	91,8	88,4	87,5
6	83,7	72,8	72,6	40,2	66,3
7	92,1	84,6	84,3	78,9	79,8
8	93,1	85,9	85,4	83,6	83,7
9	92,9	89,3	89,2	66,5	71,4
Mean / Median	**90,4 / 92,1**	**82,7 / 84,6**	**82,4 / 84,3**	**71,5 / 69,2**	**73,7 / 71,4**

Table 4: Results for individual songs of the training/test set for the different processing stages of the algorithm. Algorithm settings according to Table 2 – Number of estimated pitches per frame = 10, Classifier Settings K=11

The low classifier performance for song number 3 might be caused by a lack of training data in the corresponding frequency range. As can be seen from Fig. 1.18 (p.21) song number 3 is the song with the lowest F_0s. Song number 6 is the song with the second lowest F_0s. Moreover it has been observed from the reference annotation that this song has the lowest mean ΔF_0 which means that singing voice characteristics (e.g. vibrato) might be less pronounced which makes discrimination between vocal and instrumental sounds more difficult, which possibly is the reason for significantly lower classifier accuracy. Since absolute accuracy of the classifier is strongly dependent on the accuracy previous stages (MPE, Tracking), classifier accuracy is related to the total number of correctly estimated pitches to allow a fair comparison. This is referred to as the relative classifier accuracy calculated as the ratio of classifier accuracy and absolute pitch estimation accuracy. In this way mean relative classifier accuracy reaches 79.09%.

4.4. Comparison with the MIREX 2008 melody extraction contest

In order to compare our approach to other recent approaches results in terms of overall accuracy are compared to the results of different algorithms that entered the MIREX melody extraction contest 2008. The results are displayed in Fig. 4.13 using box plots representing the variance of algorithm performance across the evaluation song database. Unfortunately results are not directly comparable since the data base that has been used for training and evaluation of the proposed algorithm comprises only 9 of 16 songs that were originally used in the MIREX Melody Extraction Contest 2008 in the category singing voice melody extraction. In detail the original data set that is reserved for competition comprises 16 songs containing vocals and the training set which is available online(1) comprises only 9 of these 16 songs. Nevertheless comparing results gives an idea of achievable accuracy.

As can be seen, the proposed algorithm (no.9, displayed in green) shows large variance in accuracy compared to the others (except for algorithm 3 which is the worst performing algorithm). On the other hand the proposed algorithm shows the highest maximum classifier accuracy value attained for one song which is 89,8% (see Table 4). These results suggest that the proposed algorithm works better under certain conditions (number and kind of used instruments, singing style much/less vibrato, ratio between music and singing voice) which vary from song to song. This could be seen as weakness of the proposed approach. However, it should be noted that the winning algorithm (no.5) of the MIREX Melody Extraction contest 2008 did not specifically address the estimation or detection of singing voice but was rather designed to estimate the pre-dominant pitch, indifferent if the pre-dominant pitch corresponded to singing voice or instrumental sounds.

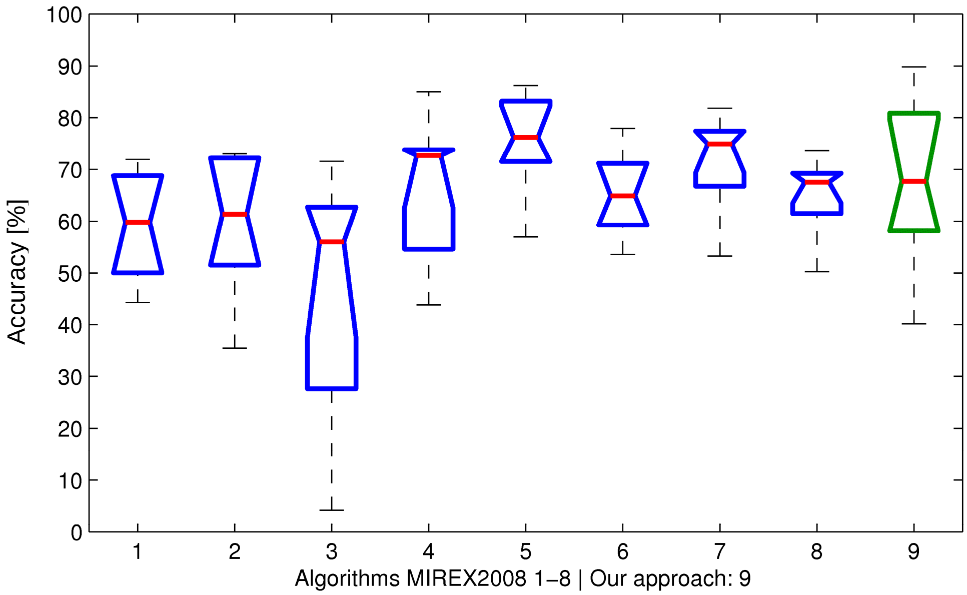


Fig. 4.13.: Variance in overall accuracy of the algorithms that entered the MIREX 2008 melody extraction contest (1-8) compared to our approach (9 – green, Settings KNN K=11). Results are unfortunately not directly comparable since the data set used for evaluation of our approach contains only 9 of the 16 vocal songs contained in the data set used for competition which has not been available

[1] LabROSA: "Laboratory for the recognition and organization of speech and audio" http://labrosa.ee.columbia.edu/projects/melody/

Chapter 5 - Conclusion

A frame work for the detection of singing voice signals in polyphonic popular music recordings has been presented. Results suggest that the representation of the musical content of song in terms of multiple parallel F0 trajectories allows discrimination between vocal and instrumental sounds.

The individual processing stages have been evaluated separately and corresponding results have been presented. The auditory motivated preprocessing of the audio signal improves the ability to detect the singing voice F0 in the polyphonic mixture signal. The average gain in vocal F0 estimation accuracy compared to estimation from simple FFT magnitude spectra for various numbers of estimated pitches is about 5%. This justifies the increase of computational complexity due to auditory preprocessing. Absolute vocal F0 accuracy of frame wise pitch estimates yields 90.4% on average for the 9 song excerpts. Pitch tracking based on cubic extrapolation has been presented yielding absolute accuracy of 82,7% and relative tracking accuracy (with respect to the pitch estimation accuracy) yields 91,5%. Tracking based ion cubic extrapolation shows only little improvements compared to a simple pitch tracking algorithm (0.3 % on average). A training data set later used for classification has been derived from polyphonic music excerpts based on reference transcriptions of the singing voice F0 trajectory. The discriminability of the two classes based on the features derived from the training data has been tested by means of statistical testing (F-ratio, LDA) and by means of simulation (N-fold cross validation). Absolute classifier accuracy reaches 71,5% and relative classifier accuracy (with respect to the pitch estimation accuracy) reaches 79,1%. Finally a method has been presented how to convert the vocal F0 trajectory into individual note events based on MIDI quantization. Weaknesses of the proposed approach towards singing voice F0 estimation have been discussed and finally an outlook for further improvements will be given in the following.

Chapter 6 - Outlook

Evaluation results demonstrate that the proposed approach towards singing voice detection is comparable to recent methods. Nevertheless there is room for improvements. In the actual implementation only monophonic audio signals are considered. In doing so, panning information, possibly bearing valuable information for source tracking, is neglected.

Moreover the context of vocal F0s to the musical accompaniment in terms of harmony could be incorporated to facilitate final note quantization. Additionally the probability of note transitions in the final note sequence could be evaluated using a musicological model.

In the current implementation the main criterion for the tracking of pitch candidates is the frequency in Hz. Since also the amplitude trajectory tends to continuous and varies slowly from frame to frame for a certain pitch track a second tracking criterion could be incorporated based on the evolution of the amplitude trajectory.

Another idea is the use of multiple parallel classifiers for voicing decisions. In that way one could possibly increase reliability of classification using decision by majority.

Apart from that it should be investigated if there is a gender specific difference for achievable classifier performance and to which degree the singer dependent absolute F0 range has an influence on that.

REFERENCES

[Bregman90] "Auditory Scene Analysis – The Perceptual Organization of Sound", Albert S. Bregman, The MIT Press, 1990

[Boer78] "On cochlear encoding: Potentials and limitations of the reverse-correlation technique", E. de Boer and H. R. de Jongh, J. Acoust. Soc. Amer.,vol. 63, no. 1, pp. 115–135, 1978.

[Borch02] "Spectral distribution of solo voice and accompaniment in pop music", Daniel Zangger Borch and J. Sundberg, 1. Stockholm Music Conservatory, 2002

[Cheveigne05] "Pitch perception models," by A. de Cheveigné, in *Pitch*, C. Plack and A. Oxenham, Springer, 2005

[Duda01] "Pattern Classification", Richard O. Duda, Peter E. Hart, David G. Stork, Wiley & Sons, 2001

[Ellis96] "Prediction-driven computational auditory scene analysis", D. P. W. Ellis, Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, 1996

[Fletcher98] "The physics of Musical Instruments", H. Fletcher and T.D. Rossing, Springer, Germany, Second Edition, 1998

[Gelfand01] "Essentials of Audiology", S.A. Gelfand, Thieme, 2001

[Goto04] "A real-time music-scene-description system: predominant-F0 estimation for detecting melody and bass lines in real-world audio signals". M. Goto, National Institute of Advanced Industrial Science and Technology (AIST), Japan, 2004

[Goto06] "F0 estimation method for singing voice in polyphonic audio signals based on statistical vocal model and Viterbi search ", Hiromasa Fujihara, Tetsuro Kitahara, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, IEEE International Conference on Acoustics, Speech, and Signal Processing, 2006

[Järveläinen99] "Audibility of Inharmonicity in string instrument sounds and implications to digital sound synthesis", Hanna Järveläinen, Vesa Välimäki, and Matti Karjalainen, Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, 1999

[Klapuri06] "Signal processing methods for music transcription", Anssi Klapuri and Manuel Davy, Springer, 2006

[Klapuri08] “Multipitch Analysis of polyphonic music and speech signals using an auditory model ”, Anssi Klapuri, IEEE Transactions on audio, speech, and language processing. Vol. 16, No. 2, February 2008

[Li05] “Detecting pitch of singing voice in polyphonic audio”, Yipeng Li and DeLiang Wan, Department of Computer Science and Engineering, & Center of Cognitive Science, Ohio State University, 2005

[Lyon95] “Automatic gain control in cochlear mechanics”, Richard F. Lyon, Advanced Technology Group, Apple Computer, Inc., Cupertino, USA, and Computer Science Department, California Institute of Technology, Pasadena, USA, 1995

[Moore95] “Frequency analysis and masking,” from Hearing—Handbook of Perception and Cognition, B. C. J. Moore, 2nd ed. San Diego, 1995, pp. 161–205.

[Moore04] “Phsychology of Hearing”, B.C.J. Moore, 5th ed., Elsevier Academic Press, 2004

[Oppenheim04] “Zeitdiskrete Signalverarbeitung”, A. V. Oppenheim, R. W. Schafer, J. R. Buck, 2nd edition, Pearson, 2004

[Patterson76] “Auditory filter shapes derived with noise stimuli”, R. D. Patterson, J. Acoust. Soc. Amer., vol. 59, no. 3, pp. 640–654, 1976

[Patterson96] “A functional model of neural activity patterns and auditory images,” R. D. Patterson and J. Holdsworth, in Advances in Speech, Hearing and Language Processing, W. A. Ainsworth, Ed. Greenwich, 1996, pp. 548–558.

[Plack95] “Loudness perception and intensity coding,” C. J. Plack and R. P. Carlyon, Ch4. in Hearing-Handbook of Perception and Cognition, B. C. J. Moore, Ed., 2nd ed. San Diego, 1995, pp. 123–160.

[Plack04] “Pitch: neural coding and perception”, Christopher J. Plack,Andrew J. Oxenham,Richard R. Fay,Arthur N. Popper, Springer, 2004

[Ryynänen04] “Probabilistic Modelling of Note Events in the Transcription of Monophonic Melodies“, M. Ryynänen, , Master of Science Thesis, Tampere University of Technology, 2004

[Pickels08] “An Introduction to the physiology of hearing”, by J.O. Pickles 3rd ed., Academic Press, 2008

[Poliner2007] “Melody Transcription from music audio – approaches and evaluation”, Graham E. Poliner, Daniel P. W. Ellis, Andreas F. Ehmann, Emilia Gómez, Sebastian Streich, and Beesuan Ong. IEEE Transactions on audio, speech, and language processing. Vol. 15, No. 4, May 2007

[Rao08] “Melody Extraction using harmonic matching”, Vishweshwara Rao, Preeti Rao, Indian Institute of Technology, Bombay, 2008

[Shenoy05] “Singing Voice Detection for Karaoke Applications”, Arun Shenoy, Yuansheng Wu, Ye Wang, in Proc. Int. Symp. on Visual Communications, and Image Processing (VCIP), 2005

[Sundberg77] “The science of the singing voice”, J. Sundberg, Northern Illinois University Press, Illinois, 1987

[Saitou02] “Extraction of F_0 dynamic characteristics and development of F0 control in singing voice”, Takeshi Saitou, Masashi Unoki, and Masato Akagi, , Proceedings of the International Conference on Auditory Display, Japan, 2002

[Selfridge-Field98] “Conceptual and representational issues in melodic comparison”, E. Selfridge-Field, Computing in Musicology, vol. 11, pages 3-64, 1998.

[Tzanetakis04] “Song specific bootstrapping of singing voice structure”, George Tzanetakis, Department of Computer Science, University of Victoria, Canada, 2004