THÈSE DE DOCTORAT

ÉCOLE DOCTORALE EDITE
UNIVERSITÉ PARIS VI - PIERRE ET MARIE CURIE

MULTIPLE FUNDAMENTAL FREQUENCY ESTIMATION OF POLYPHONIC RECORDINGS

Chunghsin YEH

pour obtenir le grade de

DOCTEUR de l'UNIVERSITÉ PARIS VI - PIERRE ET MARIE CURIE

le 26 juin 2008

devant le jury composé de

Xavier RODET	Université Paris 6	Directeur de thèse
Yves GRENIER	ENST Télécom Paris	Rapporteur
Anssi KLAPURI	Tampere University of Technology	Rapporteur
Alain de CHEVEIGNÉ	Université Paris 5	Examinateur
Gaël RICHARD	ENST Télécom Paris	Examinateur
Manuel DAVY	Université Lille 1	Examinateur

To my parents,

ACKNOWLEDGEMENTS

Carrying out a Ph.D. research at IRCAM, right in the center of Paris, has been an extraordinary experience. The accomplishment of this thesis would not have been possible without the support of many people to whom I would like to express my gratitude.

First of all, I would like to thank my thesis supervisor, Prof. Xavier Rodet, for his generosity and kindness to offer me an opportunity to take on a Ph.D. research in the Analysis/Synthesis team. I thank him also for his unfailing encouragement and his sharing of knowledge and experience in signal processing.

I owe the quality of this work especially to my project adviser, Dr. Axel Roöbel. With his invaluable advice and continuous support, my research has been guided in learning the approaches to theoretical studies for solving the problems, as well as practical skills for implementing the algorithms. I thank him for his tremendous patience and for spending endless hours with me discussing the problems and revising the articles and this thesis.

The long and hard work of Ph.D. research would not have been of so much fun without Niels Bogaards working in the same office. I thank for his spontaneous feedback about my various requests and numerous exchanges of information for living a good life.

Many thanks also go to my colleagues and friends at IRCAM for the many good times together in Paris.

I feel very fortunate to have several good Taiwanese friends in Paris. Sincerest thanks to Chih-Chung and Vivianne for guiding me to live a better life, and to Meihua for encouraging me with unceasing care.

I would like to thank particularly Prof. Ricardo Canzio for motivating me and helping me to persuade further studies at IRCAM.

Most of all, I would like to express my deepest gratitude to my parents for their unwavering love.

RÉSUMÉ

La fréquence fondamentale, dite F0, est un descripteur essentiel des signaux audio de musique. Bien que les algorithmes d'estimation de F0 unique aient considérablement progressé, leur application aux signaux de musique reste limitée parce que la plupart d'entre eux contiennent non pas une, mais plusieurs sources harmoniques en même temps. Par conséquent, l'estimation des F0s multiples est une analyse plus appropriée, et qui permet d'élargir le champ d'application à des tâches telles que la séparation de sources, l'extraction d'information de musique ou la transcription automatique de la musique.

La difficulté d'estimer des F0s multiples d'un signal audio réside dans le fait que les sources sonores se superposent souvent dans le domaine temporel ainsi que dans le domaine fréquentiel. Les informations extraites sont en partie ambiguës. En particulier, lorsque des notes de musique en relation harmonique sont jouées en même temps, les partiels des notes aiguës peuvent recouvrir les partiels des notes graves. D'ailleurs, les caractéristiques spectrales des instruments de musique sont variées, ce qui augmente l'incertitude des amplitudes estimées des partiels des sources sonores. La complexité qui en résulte génère aussi une ambiguïté d'octave et il est d'autre part difficile d'estimer le nombre de sources. Cette thèse traite ces problèmes en trois étapes: l' estimation du bruit, l'évaluation conjointe des F0 hypothèses, et l'inférence de la polyphonie.

Le signal observé est modélisé par la somme de plusieurs sources harmoniques et du bruit, où chaque source harmonique est modélisée par une somme de sinusoïdes. Dans le cas de l'estimation des F0s, le nombre de sources est à estimer également. Si la partie bruit n'est pas estimée à l'avance, le nombre de sources risque d'être surestimé, les sources supplémentaires servant à expliquer la partie bruit. Un algorithme d'estimation du niveau de bruit est donc développé afin de distinguer les pics relatifs au bruit des pics sinusoïdaux qui correspondent aux partiels des sources harmoniques.

Une fois les composantes spectrales identifiées comme étant des sinusoïdes ou du bruit, les partiels d'un ensemble de sources hypothétiques devraient s'ajuster à la plupart des pics sinusoïdaux. Afin d'évaluer leur plausibilité, un algorithme d'estimation conjointe est proposé, ce qui permet de traiter le problème des partiels superposés. L'algorithme d'estimation conjointe proposé est fondé sur trois hypothèses liées aux caractéristiques des instruments de musique: l'harmonicité, la douceur de l'enveloppe spectrale, et l'évolution synchrone des amplitudes des partiels. Lorsque le nombre de sources est connu, les F0s estiméees sont déterminés par la combinaison la plus

probable. Dans ce cas, l'algorithme proposé donne un résultat prometteur qui se compare favorablement à l´état de l'art.

L'estimation conjointe des F0s multiples permet de traiter de manière satisfaisante le problème des partiels superposés. Cependant, le temps de calcul de cette approche est élevé, parce que le nombre de combinaisons hypothétiques s'accroît exponentiellement avec le nombre de F0s candidats. Au contraire, l'approche basée sur une estimation itérative est plus rapide mais elle est moins optimale pour traiter le problème des partiels superposés. Dans l'espoir d'obtenir d'une part efficacité et d'autre part robustesse, ces deux approches sont combinées. Un algorithme itératif de sélection des F0s candidats, visant à en diminuer le nombre, est proposé. Comparé à deux fonctions de saillance polyphonique, cet algorithme itératif réduit de cents fois le nombre de candidats en perdant seulement 1 à 2% de la précision d'estimation des F0s multiples. Le résultat montre d'ailleurs qu'une augmentation du nombre des F0s candidats ne garantit pas une meilleure performance de l'algorithme d'estimation conjointe.

L'estimation du nombre de sources, dite *inférence de la polyphonie*, est le problème le plus ardu. L'approche proposée consiste à faire une hypothèse sur le nombre de sources maximal et ensuite à sélectionner les meilleures F0s estimés. Pour cela, les F0s candidats qui se trouvent dans les meilleures combinaisons, sous l'hypothèse du nombre de sources maximal, sont retenus. L'estimation finale des F0s est obtenue en vérifiant de manière itérative les combinaisons de F0s sélectionnées selon l'ordre de probabilité de chaque F0. Une hypothèse de F0 est considérée comme valide si elle permet d'expliquer des pics d'énergie significatifs ou si elle améliore la douceur de l'enveloppe spectrale pour l'ensemble des F0s estimés.

Le système proposé est évalué en utilisant une base de données de morceaux de musique construite spécialement pour l'occasion. La précision obtenue est environ 65%. Lors de la compétition d'estimation de F0s multiples de MIREX (Music Information Retrieval Evaluation eXchange) 2007, le système proposé a été évalué comme l'un des meilleurs parmi les 16 systèmes soumis.

Mots-Clés: estimation de la fréquence fondamentale, estimation du bruit, séparation de sources, transcription automatique de la musique, analyse du signal.

ABSTRACT

The fundamental frequency, or F0, is an essential descriptor of music sound signals. Although single-F0 estimation algorithms are considerably developed, their applications to music signals remain limited, because most music signals contain concurrent harmonic sources. Therefore, multiple-F0 estimation is a more appropriate analysis, which broadens the ranges of applications to source separation, music information retrieval, automatic music transcription, amongst others.

The difficulty of multiple-F0 estimation lies in the fact that sound sources often overlap in time as well as in frequency. The extracted information is partly ambiguous. Above all, when musical notes are played in harmonic relations, the partials of higher notes may overlap completely with those of lower notes. Besides, spectral characteristics of musical instrument sounds are diverse, which increases the ambiguity in the estimation of partial amplitudes of sound sources. The resulting complexity causes not only octave ambiguity but also the ambiguity in the estimation of the number of sources. This thesis addresses these problems in three stages: noise estimation, joint evaluation of F0 hypotheses, and polyphony inference.

The observed sound signal is modeled as a sum of several harmonic sources and noise, where each harmonic source is modeled as a sum of sinusoids. To estimate multiple F0s, the number of sources is to be inferred. If the noise part is not estimated beforehand, the number of sources can be overestimated when unnecessary sources are simply used to explain the noise. A noise level estimation algorithm is therefore developed to distinguish sinusoidal peaks, considered to be the partials of harmonic sources, from noise peaks.

Once the spectral peaks are classified according to the estimated noise level, the partials of a set of hypothetical sources should match most of the sinusoidal peaks. To evaluate the plausibility of a set of hypothetical sources, a joint estimation algorithm is proposed which makes the most of the handling of overlapping partials. The joint estimation algorithm is based on three assumptions of the characteristics of harmonic instrument sounds: harmonicity, the smoothness of spectral envelope and synchronous evolution of partial amplitudes. When the number of sources is known, the estimated F0s are determined by the hypothetical combination with the best *score* of plausibility. In this case, the proposed algorithm demonstrates a promising result which is competitive to those of existing methods.

Joint estimation of multiple F0s allows a correct handling of ambiguous partial amplitudes. However, the downside is its computational demand because the number

of hypothetical combinations grows exponentially with the number of F0 candidates. Alternatively, the iterative estimation approach has an advantage in efficiency but is less optimal in the handling of overlapping partials. The combination of the two approaches can thus be expected to achieve both efficiency and robustness. An iterative estimation algorithm is proposed for candidate selection, aiming at the reduction of the number of candidates. Compared with two polyphonic salience functions, the iterative algorithm reduces hundred times of the number of candidates while losing only 1 to 2% of accuracy for multiple-F0 estimation. The result also demonstrates that more candidates does not guarantee a better performance for the joint estimation algorithm.

The estimation of the number of sources, or polyphony inference, is the most challenging problem. The proposed approach is to first estimate the maximal polyphony and then to consolidate the F0 estimates. The F0 candidates kept in the best combinations of the maximal polyphony are combined and verified iteratively, in order of their respective probabilities, to yield the final estimates. An F0 hypothesis is considered a valid estimate if it either explains significant energy or improves the spectral smoothness of the set of the valid F0 estimates.

The proposed system is evaluated by a specially constructed music database. The average accuracy rate is about 65%. In the multiple-F0 estimation contest of MIREX (Music Information Retrieval Evaluation eXchange) 2007, it has been evaluated as one of the best among 16 submitted systems.

Keywords: fundamental frequency estimation, noise estimation, source separation, automatic music transcription, signal analysis.

CONTENTS

OVERVIEW

The scope of this thesis is the estimation of multiple fundamental frequencies (F0s) of polyphonic music recordings. The objective is to develop a frame-based multiple-F0 estimation system. The thesis begins with an introduction of the fundamentals of F0 estimation, followed by a review of the state-of-the-art methods for the problem of multiple-F0 estimation. Then, the algorithm for each part of the proposed system is presented and evaluated. The thesis is laid out as follows:

In Chapter 1, the problem of single-F0 estimation is defined, which leads to the problem definition of multiple-F0 estimation. Several single-F0 estimation algorithms are reviewed, which forms the basis of the study of the multiple-F0 estimation problem. Because harmonic instrument sounds are the main concern of this thesis, their physical properties are surveyed and summarized. The difficulties of estimating multiple F0s in polyphonic music signals are then discussed.

In Chapter 2, the state of the art for multiple-F0 estimation is reviewed. The existing methods are categorized into two groups: the iterative estimation approach and the joint estimation approach. Several criteria for the estimation of the number of sources are summarized. The common problems encountered are discussed, which lays the foundations for the development of the proposed method.

In Chapter 3, three major problems to be addressed are identified and the proposed approach is presented. The development of the system is based on a generative signal model and three physical principles of musical instrument sounds. The system overview is given, which summarizes the model assumptions and the step-by-step procedures of the multiple-F0 estimation system.

In Chapter 4, an adaptive noise estimation algorithm is presented based on a generative noise model. Following two existing methods, it describes the noise level as a cepstrally liftered curve which classifies spectral peaks into sinusoids and noise. The precision of the estimated noise level is experimentally evaluated by means of a white noise signal, with or without embedded sinusoids.

In Chapter 5, an algorithm for joint estimation of multiple F0 hypotheses is presented. This algorithm is based on the assumption that the number of F0s is given. The joint estimation algorithm focuses on the handling of the overlapping partials and the scoring of a hypothetical combination. The essence of the algorithm is a score function that jointly evaluates the plausibility of a combination of F0 hypotheses. The design purpose of the score criteria is demonstrated using monophonic samples of musical instrument sounds. Then, the joint estimation algorithm

is evaluated by artificially mixed polyphonic samples.

In Chapter 6, three candidate selection methods are proposed, aiming at reducing the number of hypothetical combinations to be evaluated by the score function. Two F0 salience functions and an iterative estimation algorithm are proposed. Both salience functions rely on a global threshold to select F0 candidates. The iterative estimation method first extracts the non-harmonically related F0s and then detects the harmonically related F0s. The candidate selection algorithms are evaluated with respect to the efficiency and the robustness. Their advantages and disadvantages are discussed.

In Chapter 7, an algorithm is presented for polyphony inference, which completes the proposed multiple-F0 estimation system. In order to evaluate the proposed system, a systematic method is proposed to construct a synthesized music database with verifiable ground truth. The multiple-F0 estimation system is evaluated by two databases: one containing sources with equal energy and the other containing sources with different energy. The results are compared and discussed. Finally, a simple application to F0 tracking for solo recordings of monodic instruments is presented.

In Chapter 8, the conclusions are drawn by summarizing the main contributions of this thesis. At the end, the perspectives concerning all the algorithms developed are given for initiating related research topics.

1

INTRODUCTION

The fundamental frequency, or F0, is an essential descriptor of harmonic sound signals. Single-F0 estimation algorithms assume that there is at most one periodic source. F0 estimation of single-speaker speech signals has many applications such as speech recognition, speech transformation and speaker identification. For the analysis of music signals, it is generally admitted that single-F0 estimation algorithms are not adequate because musical notes played by various instruments usually sound simultaneously. Multiple-F0 estimation algorithms assume that there could be more than one periodic source. It is, therefore, more appropriate to analyze polyphonic signals by a multiple-F0 estimator. For music signals, multiple-F0 estimation broadens the range of applications to source separation, music information retrieval, and automatic music transcription.

In this thesis, music signals are understood to be generated by superimposing individual notes of musical instruments. Accordingly, the investigation of the F0 estimation problem begins with the survey of the physical properties of various musical instruments. The problem of single-F0 estimation is described and single-F0 estimation algorithms are reviewed. The discussion is then extended to the problem of multiple-F0 estimation. The complexity of the problem is outlined with respect to four facts: overlapping partials, diverse spectral characteristics of musical instrument sounds, transients and reverberation.

1.1 Fundamental Frequency (F0) of a Periodic Signal

1.1.1 Period and fundamental frequency

An event is said to be **periodic** if it repeats itself at a regular time interval that is called its **period**. A periodic continuous time signal $\tilde{x}(t)$ has the property that there exists a $T > 0$ for which

$$\tilde{x}(t) = \tilde{x}(t + T) \tag{1.1}$$

for all values of t (Oppenheim *et al.*, 1997). If there exists a T that satisfies eq. (1.1), there exist an infinite number of T's. It can be derived from eq. (1.1) that

$$\tilde{x}(t) = \tilde{x}(t + mT_0) \tag{1.2}$$

for all ts and any integer m. The fundamental period T_0, or simply period, is the smallest positive value of T for which eq.(1.1) holds. The period T_0 is defined by de Cheveigné and Kawahara (2002) as "the smallest positive member of the infinite set of time shifts that leave the signal invariant". The **fundamental frequency** F0 is defined as the reciprocal of the period:

$$F_0 = \frac{1}{T_0} \tag{1.3}$$

which measures the repetition rate.

Owing to the fact that the F0 of a speech or music sound source varies with time, it is usually assumed that the signal is stationary in a very short time duration. The F0 of a non-stationary signal can thus be determined through the approximation $\tilde{x}(t) \approx \tilde{x}(t + T_0)$ for the concerned duration. If a signal can be approximated in this way, it is called a **quasi-periodic** signal. Figure 1.2(a) illustrates an example of quasi-periodic signal with a distinct period between consecutive sharp dips. Due to the non-stationarity, several periods may have competitive fits to the signal, which results in the ambiguity in the determination of the F0.

1.1.2 Fourier series representation of periodic signals

Sinusoids are probably the most important periodic signals of all. A sinusoid can be represented by a cosine function with a specific amplitude, frequency and initial phase. Jean Baptiste Joseph Fourier was the first scholar that has the clear insight to see the potential for representing a signal with a *sum of harmonically related sinusoids*. Each component is called a **harmonic** that has a frequency that is a multiple of the fundamental frequency. He claimed that any periodic signal can be represented by Fourier series, which is justified by a mathematical theorem completed later by P.L. Dirichlet (Oppenheim *et al.*, 1997).

A periodic signal that is real can be represented as a linear combination of harmonically

related complex exponentials

$$\tilde{x}(t) = \sum_{-\infty}^{+\infty} a_h e^{jh\omega_0 t}, h = 0, \pm 1, \pm 2, ... \tag{1.4}$$

where $\omega_0 = 2\pi F_0 = 2\pi/T_0$. An important property of Fourier series representation is that the sinusoidal functions form an orthonormal basis. The Fourier series coefficients can thus be efficiently computed as follows:

$$a_h = \frac{1}{T_0} \int_{T_0} \tilde{x}(t) e^{-jh\omega_0 t} dt \tag{1.5}$$

eq.(1.4) is referred to as the synthesis equation and eq.(1.5) as the analysis equation. Rearrange eq.(1.4),

$$\tilde{x}(t) = a_0 + \sum_{h=1}^{+\infty} (a_h e^{jh\omega_0 t} + a_{-h} e^{-jh\omega_0 t}) = a_0 + \sum_{h=1}^{+\infty} (a_h e^{jh\omega_0 t} + a_h^* e^{-jh\omega_0 t}) \tag{1.6}$$

requiring $a_h^* = a_{-h}$, and further express a_h in polar form as $a_h = \frac{A_h}{2} e^{j\phi_h}$, we have

$$\tilde{x}(t) = a_0 + \sum_{h=1}^{+\infty} A_h \cos(h\omega_0 t + \phi_h) \tag{1.7}$$

This is one commonly used term for the Fourier series representation of periodic signals. A sound that can be represented by eq.(1.7) is called **harmonic sound**. Because the harmonics of a quasi-periodic signal do not have frequencies that are exactly multiples of its F0, they are often called **partials**. For the practical use of eq.(1.7), a finite and small number of sinusoids H is usually used to approximate a quasi-periodic signal:

$$\tilde{x}(t) \approx a_0 + \sum_{h=1}^{H} A_h \cos(h\omega_0 t + \phi_h) \tag{1.8}$$

1.1.3 Physical properties of harmonic instrument sounds

In this thesis, the sounds of which the partials are nearly harmonically related are generally considered harmonic sounds. Non-harmonic sounds produced by the musical instruments like mallet percussion instruments and bells are not the main concern in this work. There exist several challenges when the Fourier series representation is used as the signal model to extract the F0 of a harmonic instrument sound. The approximation in eq.(1.8) does cause estimation errors, however, this is of minor concern. The major difficulties lie in the fact that the generation of musical instrument sounds involves complex physics, producing diverse spectral characteristics. The interaction of a musical instrument with the room makes the resulting sound even more complicated. The generation of musical instrument sounds involves four parts: excitation generator, resonator, radiation and room acoustics (see Figure 1.1). In the following, physical properties of harmonic instrument sounds are described for each part of sound generation (Fletcher and

Rossing, 1998; Benade, 1976; Kuttruff, 1991).

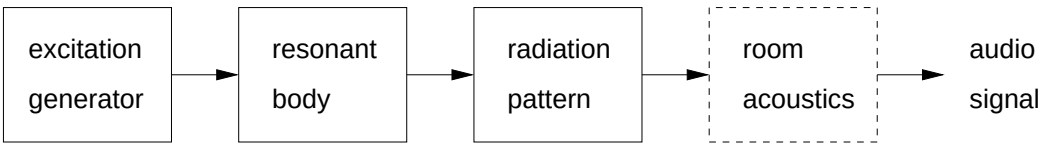


Figure 1.1: Generation of musical instrument sounds.

Excitation

For musical instruments, the sources of excitation are mechanical or acoustical vibrators. This excitation, alone or coupled with a part of the resonator, introduces a source spectrum, also called the **normal mode spectrum**. The normal mode frequencies of harmonic instruments are very nearly in integer ratios.

Woodwind instruments are made to sound by blowing an air jet across pressure-controlled valves such as single reeds or double reeds. Brass instruments are made to sound in a similar way but with the vibrating lips as valves. For flute-like instruments, an air jet is blown to strike a sharp edge for the acoustical excitation.

The excitation of string instruments is generated by plucking, bowing, or striking the strings. If a string is excited at $1/h$ of its length from one end, the every hth harmonic in the normal mode spectrum is suppressed. When a string is struck by a hammer, such as the pianos, reflected pulses return from both ends of the string and interact with moving hammer in a complicated manner, causing the vibrating spectrum to fall off more rapidly than that of a plucked string.

The stiffness of strings makes the partials stretch toward higher frequencies. The end supports of a string further influence the stretching partials. When a string is supported by pinned ends, the frequency of the hth partial can be estimated by

$$F_h = hF_0\sqrt{1 + B_s h^2}\ , \quad \text{where} \quad B_s = \frac{\pi^3 E d_s^4}{64 l_s^2 T_s} \tag{1.9}$$

where B_s is the inharmonicity factor, E is Young's modulus, d_s is the diameter of the string, l is the length of the string and T_s is tension. Clamping the ends reduces slightly the stretching of partials. For piano strings, the end support is between the pinned and the clamped but a general use of eq.(1.9) is accepted. Compared with the treble strings that are solid wire, the piano strings at lower registers are of less inharmonicity because they consist of a solid core wound with one or two layers of wires. The greater mass reduces the inharmonicity. The playing techniques also affects the inharmonicity. The inharmonicity is attenuated when the strings are bowed, but is more noticeable when they are plucked (*Pizzicato*) or struck (*Col Legno*).

Resonance

The vibration from the excitation generator causes the resonator to vibrate with it. The result of this coupled vibrating system is the setting-up of the **regimes of oscillation**. A regime

of oscillation is the state in which the coupled vibration system maintains a steady oscillation containing several harmonically related frequency components. The analysis of the resonances of musical instruments is often carried out through measuring or calculating the acoustic input impedance or the mechanical admittances.

For lip-driven brass instruments, the resonance frequencies and the corresponding amplitudes are a combinatorial effect of the mouthpiece coupled to the horn. The horn profile determines the resonance frequencies; the mouthpiece cup determines the amplitude envelope of local resonance maxima. For a note to be playable, either its fundamental frequency must coincide with an impedance maximum of the horn, or at least one of its low harmonics must coincide with such a resonance. In a normal situation, the fundamental frequency and several of its harmonics lie close in frequency to prominent horn resonances. If a particular harmonic does not lie close to an impedance maximum, the resulting oscillation amplitude is very small. Within the small-amplitude nonlinear approximation, the amplitude of the hth harmonic in general approximates the hth power of that of the fundamental. At a high playing level, the upper harmonics become more significant as their amplitudes relative to the fundamental increase.

Woodwind instruments have the common characteristic of changing the fundamental frequency of the note being played by opening one or more finger holes, thus changing the acoustic length of the air column. The impedance curves of lower registers are different from those of higher registers. For the clarinets, the resonance peaks of lower register notes coincide with an odd harmonic series and only the first few harmonics are well aligned with the impedance maxima. For higher register notes, the first impedance maxima might be lower than the frequency of the first harmonic. The similar characteristics are found also in the oboe but the resonance peaks of the oboe form a nearly harmonic series. Due to the leakage of high-frequency sound through the open tone holes, the strength of resonance gradually decreases from a cutoff frequency.

For flute-like instruments, the major cause of nonlinearity arises from the shape of the velocity profile of the air jet, which leads to considerable harmonic development in the tone emitted by a resonator. The pipe resonators of simple flute-like instruments like panpipes are stopped at their lower ends. Their resonance frequencies thus follow an approximately odd-harmonic series like those of the clarinets.

The body resonance of bowed string instruments shapes the spectrum of vibrating strings, providing cues to identify complex tones of string instruments. The bridge transforms the motion of the vibrating strings into periodic driving forces applied by its two feet to the top plate of the instrument. For bowed string instruments, the normal modes of vibration are mainly determined by the coupled motions of the top plate, back plate and enclosed air. The three consistent modes of resonance are: air modes, top modes and body modes. The resonance frequencies of the three modes vary from instrument to instrument.

Radiation

In the coupled system of excitation generator and resonator, the sound waves traveling in the instrument bodies build up a variety of normal modes. They are then radiated by the instrument

bodies in different ways to reach our ears. The research of instrument radiation is based on the study of the simplest types of sources: monopoles, dipoles and multiple point sources. The radiation of instrument bodies can therefore be approximated by a combination of simple vibrating sources. For example, the radiation from the open finger holes of a woodwind instrument or the piano soundboards can be seen as an array of point sources. The components of sounds with different frequencies radiate with different efficiencies, resulting in various directivity patterns.

In general, the sound radiation becomes more directional at higher frequencies. The low-frequency partials of brass instruments spread uniformly around the bell in all directions, whereas the high-frequency partials form a more progressively directed beam out along the axis of the horn.

Although the resonance curve of woodwind instruments are progressively weaker for the higher partials, their greater efficiency for high-frequency radiation compensates the energy of higher partials. The even-number partials of a clarinet tone are radiated more strongly than are the odd-number partials, which compensates the weak resonance at the odd-number harmonics.

The directivity patterns of bowed strings have been observed by Meyer (1972), based on whose suggestions seating arrangements of bowed strings in a concert hall were made. The directivity patterns of the violins, for example, change dramatically with frequencies, which results in the illusion that different notes come from different directions. This has been described as the "directional tone color" (Weinreich, 1997) and considered important consequences for the perception of vibratos and solo violins in concertos.

Room acoustics

In an enclosed space, certain frequency ranges might be reinforced by the modes of the room. The frequencies of the free undamped vibrations of a rectangular volume of air bounded by reflective surfaces can be expressed by the following equation (Rayleigh, 1945)

$$f_r = \frac{c}{2}\sqrt{\frac{n_x^2}{L_x^2} + \frac{n_y^2}{L_y^2} + \frac{n_z^2}{L_z^2}} \tag{1.10}$$

where c is the speed of the sound, L_x, L_y and L_z are the dimensions of the rectangular room and n_x, n_y and n_z determine the vibrating modes: axial mode (two surfaces), tangential mode (four surfaces) and oblique mode (six surfaces).

When a sound source produces sounds in an enclosed space, the direct sound is followed by early reflected sounds and later, by a collection of dense reflections called **reverberation**. The behavior of room modes and reverberation is characterized by Schroeder frequency (Kuttruff, 1991):

$$f_s \approx 2000\sqrt{\frac{T_R}{V}} \tag{1.11}$$

where T_R is the reverberation time and V is the volume of the enclosed space. When f_r is below the Schroeder frequency, the discrete room modes are important characteristics of the room

acoustics. Above the Schroeder frequency, reverberation is in fact a consequence of the overlapping of simultaneously excited room modes that are no longer individually distinguishable (Jot *et al.*, 1997). Compared to the relatively deterministic nature of early reflections, reverberation can be seen as a stochastic process.

1.1.4 Fundamental frequency and pitch

The partials of a harmonic instrument sound evokes the perception of a **pitch** that is closely related to the fundamental frequency. A simple description of the process of pitch perception can be stated as: the inner ear (cochlea) converts a vibration pattern in time (that of the eardrum) into a vibration pattern in space (along the basilar membrane) and, in turn, into a spatial pattern of neural activity which can be interpreted by human brain as a pitch. The American National Standard Institute (ANSI) defines pitch as "that auditory attribute of sound according to which sounds can be ordered on a scale extending from low to high". The French standards organization (Association Française de Normalisation, AFNOR) defines: "pitch is associated with frequency and is low or high according to whether this frequency is smaller or greater". Both verbal definitions are rather abstract. An operational definition is given in (Hartmann, 1998): "sound has certain pitch if it can be reliably matched by adjusting the frequency of a sine wave of arbitrary amplitude".

The theories of auditory pitch analysis tend to differ in two dimensions (Bregman, 1990): whether they see pitch analysis as based primarily on place information – **spectral pitch** or on periodicity information – **periodicity pitch**, and what method is used to derive the pitch from the type of information that is used. The human auditory system seems to perceive a pitch through pattern matching. We recognize a sound by its spectral pattern composed of a series of partials that characterize it. Even when some partials are too weak to be detected, the human auditory system tends to reconstruct the missing partials and complete the pattern matching task.

Although it is widely accepted that the term **pitch estimation** is equivalent to **F0 estimation**, the latter is used in this thesis for the reason that the goal of this work is not to extract what is perceived as a pitch but to extract the F0 as a parameter of the signal model.

1.2 Single-F0 Estimation

Single-F0 estimation algorithms assume that there is at most one harmonic source in the observed short-time signal. Without loss of generality, the observed signal can be expressed as a sum of a quasi-periodic part $\tilde{x}(t)$ and the residual $z(t)$:

$$\begin{aligned} x(t) &= \tilde{x}(t) + z(t) \\ &\approx \sum_{h=1}^{H} A_h \cos(h\omega_0 t + \phi_h) + z(t) \end{aligned} \tag{1.12}$$

where eq.(1.8) is used for the approximation. The single-F0 estimation problem is to extract the period or the fundamental frequency of $\tilde{x}(t)$. Notice that the goal is not to minimize the residual $z(t)$, but to extract the quasi-periodic part $\tilde{x}(t)$ with high periodicity/harmonicity. The common errors made are **subharmonic errors** and **super-harmonic errors**, both of whose estimated F0s are harmonically related to the correct F0. Subharmonic errors correspond to F0s that are *unit fractions* of the correct F0 and super-harmonic errors correspond to those which are *multiples* of the correct F0.

The mathematical formulation for the problem of single-F0 estimation depends on the way the periodicity in $\tilde{x}(t)$ is extracted. Single-F0 estimation algorithms are often categorized into two groups: the time domain approach and the spectral domain approach. Most algorithms do not use an explicit source model as expressed in the approximation of eq.(1.12), but rather attempt to extract directly the periodicity in either the time domain or the spectral domain.

1.2.1 Time domain approach

By the definition of periodic signals as eq.(1.1), time domain methods look for a similar repetitive waveform in $x(t)$ through pattern matching between $x(t)$ and the delayed $x(t)$. Pattern matching in time domain can be carried out through multiplication or subtraction between patterns.

Autocorrelation

Multiplication between patterns measures their correlations. Autocorrelation function calculates the sum of the product between a signal $x(t)$ of finite duration L and its delayed version $x(t+\tau)$ for each **lag** τ in search:

$$\mathrm{ACF}(\tau) = \frac{1}{L} \sum_{t=0}^{L-\tau-1} x(t)x(t+\tau) \tag{1.13}$$

For a quasi-periodic signal, large correlation occurs when τ equals the period or multiples of the period. Autocorrelation method selects the highest non-zero-lag peak as the estimated period. However, this simple selection technique is sensitive to formant structures in speech signals and the resonance structure in music signals. To attenuate the effect of the formant or the resonance, special treatments like center clipping, spectral flattening and nonlinear distortion were suggested by various researchers (Hess, 1983).

Magnitude difference

The Average Magnitude Difference Function (Ross *et al.*, 1974) compares the dissimilarity of $x(t)$ and $x(t+\tau)$ by the *distance* of the two patterns:

$$\mathrm{AMDF}(\tau) = \frac{1}{L} \sum_{t=0}^{L-\tau-1} |x(t) - x(t+\tau)| \tag{1.14}$$

For quasi-periodic signals, regular dominant dips can be observed when τ equals one period or multiples of the periods. The deepest non-zero-lag dip is selected as the estimated period. A similar approach is to measure the dissimilarity by the *squared distance*:

$$\mathrm{SDF}(\tau) = \frac{1}{L} \sum_{t=0}^{L-\tau-1} (x(t) - x(t+\tau))^2 \tag{1.15}$$

Squared Difference Function (SDF) is adapted for the algorithm YIN by normalizing SDF with its average over shorter-lag values (de Cheveigné and Kawahara, 2002). This is called the **cumulative mean normalized difference function**. It removes the dips at lags near zero and thus avoids super-harmonic errors. YIN has been shown to outperform several conventional methods for speech signals (de Cheveigné and Kawahara, 2001). The difference function can be generalized into any power of the distance measure and it has been investigated that a power larger than one is appropriate for weakly stationary signals. A power of two seems to be a good choice (Nguyen and Imai, 1977).

Both AMDF and SDF are related to the autocorrelation function. These methods are phase-insensitive since partials are subtracted regardless of their phases. However, they are sensitive to intensity variations, noise and low-frequency spurious signals (Hess, 1983).

Figure 1.2 shows three time-domain functions for a baritone sax signal with F0 $= 237Hz$ ($T_0 = 4.2ms$). Despite the different ways to measure the similarity between a short-time signal and its delayed versions, the common problem is the ambiguity in selecting the best period since all multiples of the period, i. e., subharmonics, are competitive candidates.

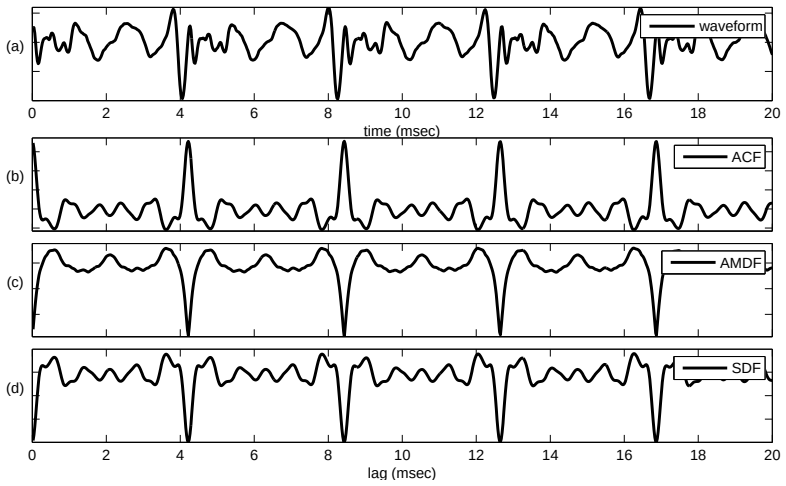


Figure 1.2: Three time-domain salience functions for a baritone sax signal of $T_0 = 4.2ms$: (a) signal waveform; (b) autocorrelation function; (c) average magnitude difference function; and (d) squared difference function.

1.2.2 Spectral domain approach

F0 estimation in the spectral domain extracts periodicity from the spectral representation based on Fourier Transform. The spectrum of a harmonic sound has dominant peaks at nearly equal spacing (see Figure 1.3(a)). Spectral domain approaches either (1) measure the regular spacing of dominant peaks as F0 or (2) formulate the salience of F0 as a function of hypothetical partials. From this point of view, fundamental frequency can also be defined as the *greatest common divisor* of the frequencies of all the harmonics.

Cepstrum

If there is periodicity of "ripples" formed by sinusoidal peaks in the spectrum, it is reasonable to apply again the Fourier analysis on the observed spectrum to analyzing the underlying periodicity. The real cepstrum of a signal is the inverse Fourier transform of the logarithm of its power spectrum. Schroeder proposed cepstrum for F0 estimation in 1962 based on the first cepstral analysis paper on seismic signals. A complete methodology was given later by Noll (Noll, 1967). The lower-**quefrency** components in the cepstrum provide the spectral envelope information and the components as sharp cepstral peaks correspond to period candidates.

Spectral autocorrelation

As autocorrelation function searches for repetitive patterns in the time domain, it can also be applied to the spectral domain (Lahat *et al.*, 1987). The periodicity search is attained by pattern matching between the spectrum and its shifted versions. When the shift is not equal to F0 or multiples of F0, the product between the spectrum and the shifted spectrum is attenuated because partial peaks are not well aligned. The shift equal to F0 should result in the maximal spectral autocorrelation coefficient.

Spectral compression

To infer the F0 from higher harmonics that are observed as spectral peaks, Schroeder proposed to sum the *frequency-warped* spectra, i. e., spectra compressed by different integer factors on the frequency axis. The **Schroeder histogram** counts equally the contribution of each spectral peak to the related F0s that are common divisors of its frequency, which is not robust against noise and spurious peaks in the spectrum. He proposed further to weight the spectral components according to their amplitudes (Schroeder, 1968): **harmonic product spectrum** uses the log power spectrum [1] and **harmonic sum spectrum** uses the linear spectrum [2]. Summing compressed spectra focuses the energy of higher partials on distinct peaks, and the maximal peak determines the related F0.

[1] Harmonic product spectrum requires to take the exponential of the summary spectrum.
[2] The linear amplitude can be exponentially compressed beforehand.

Harmonic matching

This technique often makes use of *harmonic spectral patterns* to match the observed spectrum, either by a specific spectral model or by a harmonic comb without specifying the amplitudes of the harmonics. Specific spectral models are more often used for polyphonic signals and will be discussed later. A harmonic comb is a series of spectral pulses with equal spacing defined by an F0 hypothesis. The degree of match for an F0 hypothesis can be evaluated by the correlation between the harmonic comb and the observed spectrum (Martin, 1982; Brown, 1992), or by the minimization of the distance between the frequencies of the harmonics and the frequencies of the matched peaks (Goldstein, 1973; Duifhuis and Willems, 1982). To improve the robustness of harmonic matching, several factors have been studied: the number of harmonics (Goldstein, 1973), the quality of the peaks (Sluyter *et al.*, 1982), the tolerance interval (Sreenivas and Rao, 1981), the presence of harmonics (Doval and Rodet, 1991), etc.

Spectral peak inter-spacing

As long as the partials are well separated in the spectrum, F0 can be estimated by measuring the regular spacing between each pair of partials (Harris, 1963). Each F0 hypothesis is supported by a group of spectral peaks that have frequency spacing close to the F0 hypothesis. The hypothesis of the best support is selected as the estimated F0. The measure of support is usually related to energy and harmonicity.

Using the baritone sax signal demonstrated in Section 1.2.1, the salience functions of the spectral domain methods are tested and shown in Figure 1.3. The cepstrum method [1] is based on the logarithm of the power spectrum, whereas the others are based on the linear magnitude spectrum. The harmonic matching salience function (see Section 5.1.1) uses a harmonic comb of 15 harmonics. The spectral peak inter-spacing function groups the peaks that are of similar spectral intervals, with a constraint on reasonable harmonic locations (see Section 6.1.2 and Appendix E). All the salience functions have their maxima close to the correct F0 except those of the cepstrum method.

1.3 Multiple-F0 Estimation

Multiple-F0 estimation algorithms assume that there can be more than one harmonic source in the observed short-time signal. In general, the observed signal can be expressed as a sum of harmonic sources plus the residual. Using the Fourier series, this model can be represented as

[1] The exponential operation is skipped to show how the harmonics compete with one another.

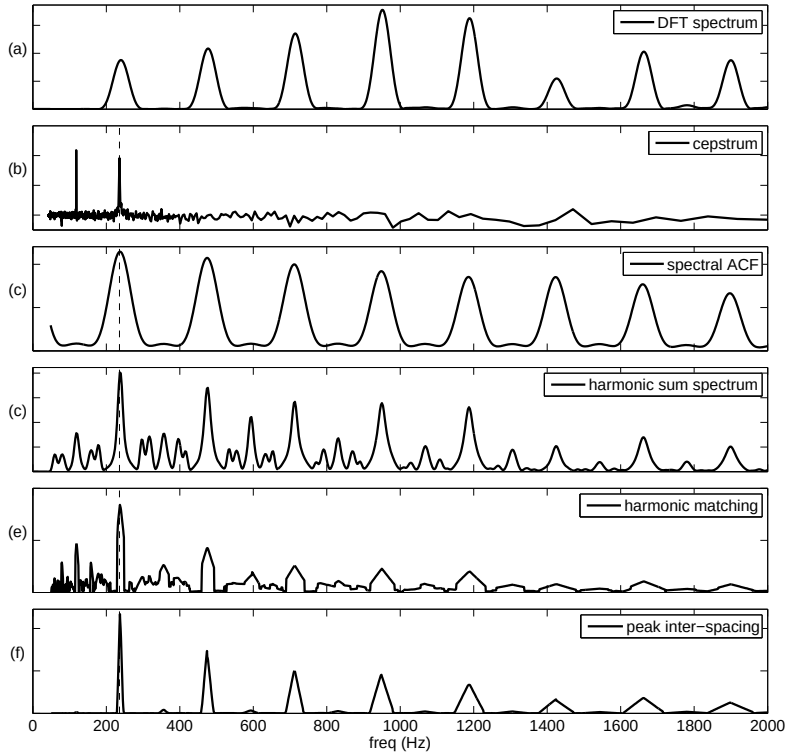


Figure 1.3: Five salience functions of the spectral domain approach tested using a baritone sax signal of F0 = $237Hz$. (a) signal spectrum; (b) cepstrum; (c) spectral autocorrelation; (d) harmonic sum spectrum; (e) harmonic matching; and (f) peak inter-spacing.

follows:

$$\begin{aligned} x(t) &= \tilde{x}(t) + z(t) \\ &= \sum_{m=1}^{M} \tilde{x}_m(t) + z(t), \qquad M > 0 \text{ with } \tilde{x}_m(t) = \tilde{x}_m(t + T_m) \\ &= \sum_{m=1}^{M} \Big\{ a_m + \sum_{h=1}^{+\infty} A_{m,h} \cos(h\omega_m t + \phi_{m,h}) \Big\} + z(t) \\ &\approx \sum_{m=1}^{M} \sum_{h=1}^{H_m} A_{m,h} \cos(h\omega_m t + \phi_{m,h}) + \bar{z}(t) \end{aligned} \tag{1.16}$$

where H_m is the number of harmonics, and M is the number of harmonic sources. The approximation for the periodic source uses the expression of eq.(1.8), and the substitution $\bar{z}(t) = z(t) + \sum_{m=1}^{M} a_m$ is used for the ease of representation. The problem of multiple-F0 estimation is to infer the number of sources and estimate the related F0s. The residual $z(t)$ comes from the components that are not explained by the sinusoids, for instance, the background noise, the spurious components or the inharmonic partials.

1.3.1 Problem Complexity

The complexity of polyphonic music signals can be demonstrated by comparing the spectrogram of a polyphonic music recording with that of a monophonic music recording (see Figure 1.4). The difficulties of extracting multiple F0s from a music recording lie in the handling of overlapping partials, transients, and reverberation, as well as the modeling of musical instrument sounds with diverse spectral characteristics.

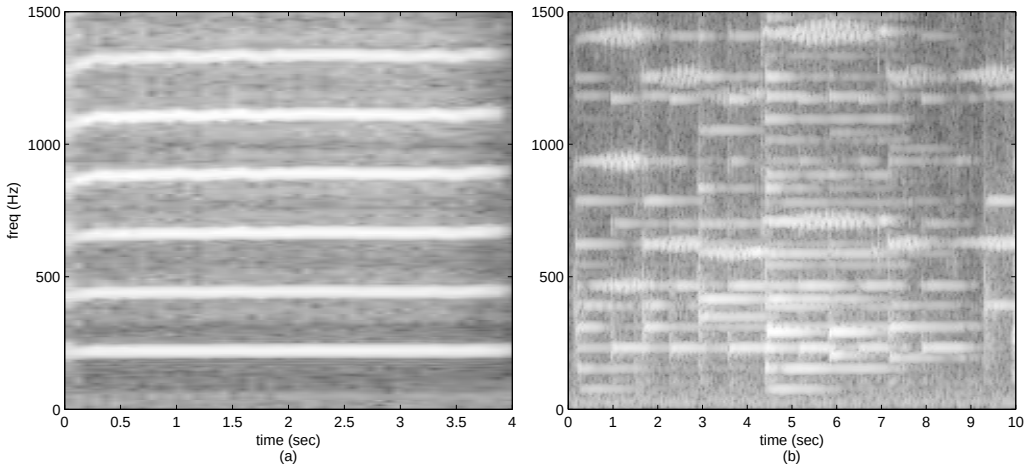


Figure 1.4: Comparison of the spectrogram of a monophonic signal with that of a polyphonic signal: (a) a trumpet note sample; (b) a piano and violin duo recording.

Overlapping partials

For polyphonic signals, different sources may interfere with one another in a way that their components overlap in time as well as in frequency. The frequencies, amplitudes and phases of the overlapping partials of harmonic sources are thus disturbed. For equal temperament, the fundamental frequencies of most musical notes are harmonically related, which results in a high probability of partial overlapping in polyphonic music signals (Klapuri, 1998). When the fundamental frequencies of two notes form integer ratios, for example, an **octave** relation, the partials of the higher note may overlap completely with those of the lower note.

Since the simultaneously sounding notes are usually unknown, it is very difficult to locate the overlapping partials. Parsons tried to detect overlapping partials with three tests: the symmetry of a spectral peak, the distance between adjacent peaks, and the well-behaved phase (Parsons, 1976). This technique relies on the sinusoidality of stationary sinusoids and is not suitable for modulated sinusoids. Moreover, the maximal number of concurrent sources is limited to two, which is not practical for the general case. Even if the number of concurrent sources is known beforehand, it still remains a challenge to decompose the overlapping partials into their original sources (Viste and Evangelista, 2002; Virtanen, 2003a; Every and Szymanski, 2004). Although the precise reallocation of the overlapping partials may not be required for multiple-F0 estimation, partial overlapping is an important issue to be addressed to achieve robust estimation

of F0s.

Diverse spectral characteristics

Music signals are mixtures of musical notes played by various instruments. The diverse spectral characteristics of musical instrument sounds add a great complexity to the problem of multiple-F0 estimation. Based on the description of instrument sound generation in Section 1.1.3, the spectral characteristics of harmonic instrument sounds are summarized in the following.

1. **Spectral envelopes**: The spectral envelope of a harmonic signal denotes a contour that passes through the prominent spectral peaks which are generally the partials. Many musical instruments produce sounds with smooth spectral envelopes [1] but differ immensely in their shapes (see Figure 1.5). Relatively weak fundamentals are often observed in the lower registers of some instruments like pianos, bassoons, oboes and guitars, resulting in *not-so-smooth* spectral envelops. The spectrum of a clarinet sound has attenuated even harmonics, of which the spectral envelope is not smooth, either. The spectra of musical instrument sounds also evolve with time in a way that partials decay at different rates. According to previous studies on the modeling of the spectral envelopes of a musical instrument sound, there exists no universal model that generalizes different registers and various playing techniques (Jensen, 1999; Loureiro *et al.*, 2004; Burred *et al.*, 2006).

2. **Inharmonic partials**: Inharmonic partials are often observed in the string instrument sounds. The displaced partials deviate from their expected frequency positions of a harmonic model. If a harmonic model allows certain inharmonicity, the model harmonics may match the partials of different sources. If it does not allow inharmonicity, more sources may be needed to explain the stretched partials.

3. **Spurious components**: For some instruments, there are some dominant components excited along with the partials. **Phantom partials** are observed in the string instrument sounds (Harold A. Conklin, 1999), which seems to be related to the tension variation in the strings. The phantom partials appear close to the frequencies of the normal partials. For the bowed string instruments, when the three resonance modes (air mode, top mode and body mode) fall between the partials, spurious components can be boosted by the resonance. These spurious components are often observed in plucked string sounds and are sometimes rather dominant compared to the partials.

Transient

The term **transient** does not have a precise definition and it is often defined by the analysis approaches (Daudet, 2006). The transients can be simply stated as the zones of short duration with fast variation of the sound signals (Rodet and Jaillet, 2001). The transients of music signals could appear at note onsets as fast attacks, or at note offsets as fast releases. The fundamental

[1] When the envelope is observed in the power spectrum under the logarithmic scale.

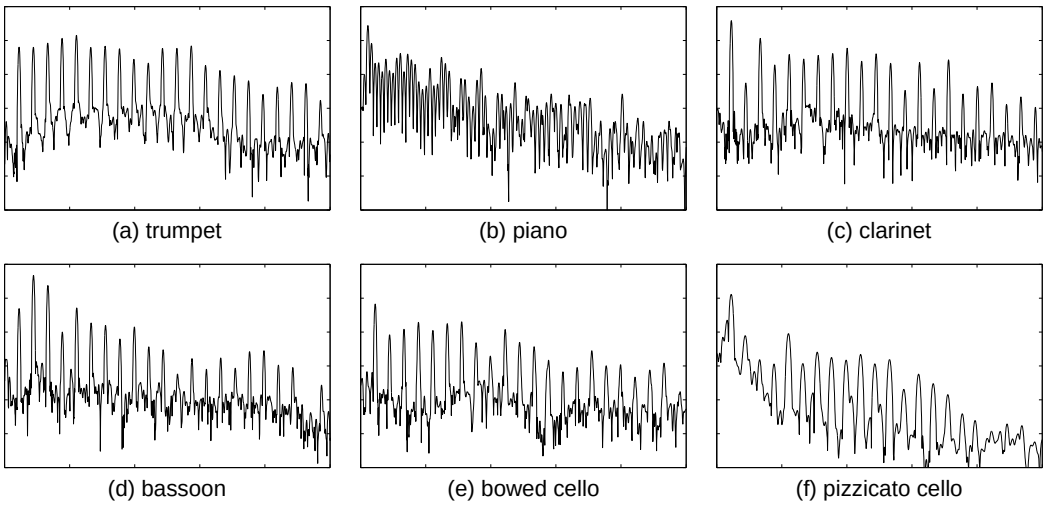


Figure 1.5: The spectra of six musical instrument sounds: (a) trumpet A3 note; (b) piano A1 note; (c) clarinet A3 note; (d) bassoon A3 note; (e) bowed cello A3 note; and (f) pizzicato cello A3 note.

frequency within the transient state poses an ill-defined problem due to its highly non-stationary nature. For bowed instruments or woodwind instruments, for example, the attack transient state might excite subharmonics (McIntyre *et al.*, 1983).

The transient of a source often appear to be impulsive and accompanied with high energy, which introduces many spurious components that may interfere with other sound sources. Recent research tends to treat the transient as a specific signal component. The transient is detected by either a non-parametric approach (Rodet and Jaillet, 2001; Röbel, 2003b; Bello *et al.*, 2005), or a parametric approach (Daudet, 2004; Molla and Torrésani, 2004).

Reverberation

Reverberation plays an important role in a music recording. A music recording in an auditorium usually requires a balance of the instrument characteristics and the room acoustics. A pair of main microphones is usually placed at the "sweet spot" to capture the whole picture of the sound scene. The recorded signal is thus a mixture of direct sounds, reflected sounds and reverberated sounds. Reverberation prolongs preceding sounds such that they overlap with the following sounds. When the recording of a monodic instrument is carried out in a reverberant environment, the recorded signal can be polyphonic (Beauchamp *et al.*, 1993; Baskind and de Cheveigné, 2003; Yeh *et al.*, 2006). The reverberated parts are quite non-stationary, which increases the complexity of the analysis of the signal.

1.3.2 Discussions

The problem of multiple-F0 estimation is far more complicated than the problem of single-F0 estimation. There are three fundamental model assumptions involved in the problem of multiple-F0 estimation: the **noise model** ($z(t)$), the **source model** ($\tilde{x}_m(t)$) and the **source interaction**

model (the effect of $\sum_{m=1}^{M} \tilde{x}_m(t)$). When the maximal number of sources is limited to one $0 \leq M \leq 1$, the porblem becomes single-F0 estimation. There is no source interaction involved in the problem of single-F0 estimation, and the inference of M becomes a voiced/unvoiced determination problem (Hess, 1983).

It is generally admitted that single-F0 estimation algorithms are not appropriate to solve the problem of multiple-F0 estimation. A naive test is to apply single-F0 estimation algorithms to a polyphonic signal and then to verify if the periodicity saliences around the correct F0s are dominant and distinct. A polyphonic signal containing four notes is tested by three time domain methods (see Figure 1.6) and five frequency domain methods (see Figure 1.7). As shown in Figure 1.6(a), the repetitive pattern in the waveform is not as clear as that of the monophonic signal shown in Figure 1.2(a). In consequence, the autocorrelation and amplitude difference functions do not show distinct peaks (or valleys) around the correct F0s. For the frequency domain methods, dominant periodicity saliences are found at the correct F0s, their subharmonics and their super-harmonics. When the energy of a source is relatively strong, the salience of its subharmonic or its super-harmonics can compete with that of a source of weaker energy. Although single-F0 estimation algorithms have limitations in analyzing polyphonic signals, they can be useful to extract F0 candidates in multiple-F0 estimation.

Another difficult problem of multiple-F0 estimation is the estimation of the number of sources. The complexity of polyphonic signals causes not only the **octave ambiguity** but also the ambiguity in the estimation of the number of sources. **Common subharmonics** have the support from the partials of concurrent sources and compete with the correct F0s. When the common subharmonic of some of the correct F0s is estimated instead, the number of sources is *underestimated*; when a source is explained by a combination of several hypothetical sources, the number of sources is *overestimated*. Moreover, spurious components and reverberation together disturb the periodic part of the sound signal, making it more difficult to achieve a robust estimation of the number of sources.

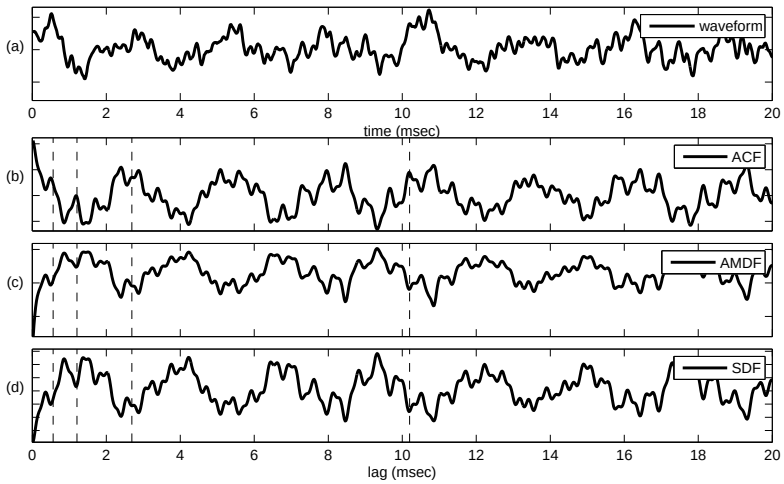


Figure 1.6: Three time-domain salience functions for a polyphonic signal containing four harmonic sources. The correct periods are marked by vertical dash lines.

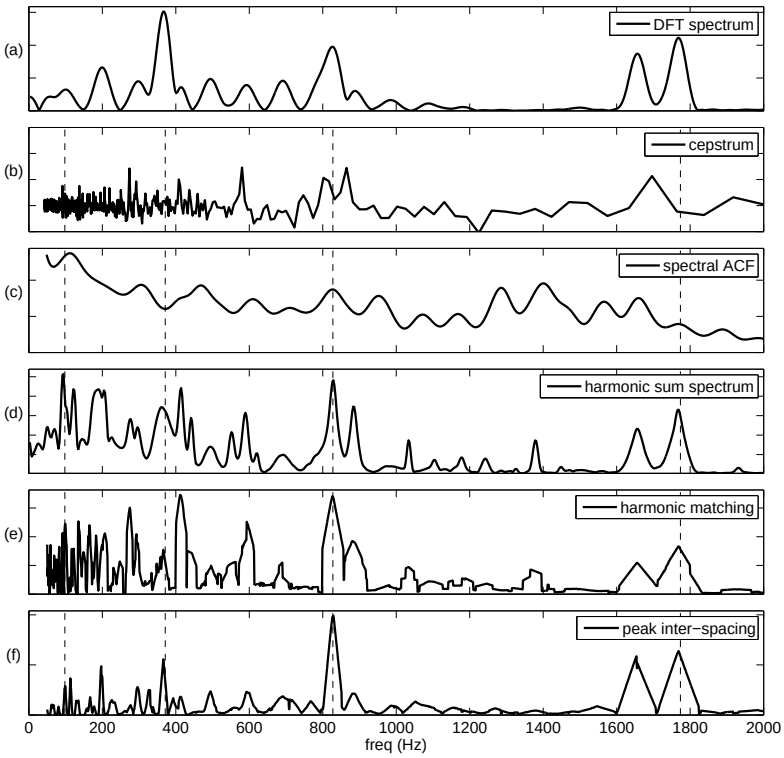


Figure 1.7: Five frequency-domain salience functions for a polyphonic signal containing four harmonic sources. The correct fundamental frequencies are marked by vertical dash lines.

2

STATE OF THE ART

In this chapter, previous studies of multiple-F0 estimation are reviewed. The related studies of, for instance, automatic music transcription and source separation are included in the review because F0s are extracted along these processes. The research of multiple-F0 estimation was initiated by the studies on separating co-channel speech signals, especially for two-speaker signals (Shields, 1970). Since then the research of multiple-F0 estimation has been extended to automatic music transcription for polyphonic music signals. Moorer (1977) started by analyzing duets and later researchers have continued to develop multiple-F0 estimation algorithms for higher polyphony. The existing methods for multiple-F0 estimation can be categorized into two groups: the iterative estimation approach and the joint estimation approach.

This categorization is different from the time/frequency domain categorization that is generally used for single-F0 estimation algorithms. The reason is that the main concern of multiple-F0 estimation is the complexity of the problem, and there exists a compromise between the efficiency and the robustness of a proposed algorithm. Theoretically, joint estimation should handle the source interaction better than iterative estimation. However, the downside is the computational cost. On the other hand, iterative estimation has the advantage of higher efficiency but is less optimal in the handling of the source interaction. Therefore, it is believed that the iterative/joint estimation categorization is more appropriate to characterize the existing methods for multiple-F0 estimation.

2.1 Iterative Estimation

The iterative approach iterates predominant-F0 estimation and the cancellation/suppression of the related sources until the termination requirement is met. Iterative estimation assumes that at each iteration there always exists a dominant source with distinct harmonic energy such that the extraction of one single F0 is reliable even when the remaining partials are fragmentary.

2.1.1 Direct cancellation

Direct cancellation applies a single-F0 estimation algorithm to extract the predominant-F0 and then eliminates *all* harmonics of the extracted source from the observed signal. This approach assumes that a complete removal of the dominant source does not influence the subsequent estimation. "Direct" cancellation here means that the source interaction such as overlapping partials is not taken care of. Parsons (1976) used Schroeder's histogram to extract the predominant F0s in a two-speaker separation problem. Once the first F0 is estimated, the spectral peaks corresponding to its harmonics are excluded before the calculation of the next histogram. The method of Lea (1992) iteratively extracts the predominant peak in the SACF as an F0 and cancels the estimate in the ACF array. de Cheveigné (1993) proposed a time-domain cancellation model and both joint cancellation and iterative cancellation are studied. The iterative cancellation algorithm estimates the predominant F0 by AMDF and cancels it by comb filtering. Direct cancellation is also applied in the spectral domain. Ortiz-Berenguer *et al.* (2005) uses spectral patterns trained from piano sounds to perform harmonic matching. Predominant sources are cancelled iteratively by means of binary masks around the matched harmonics in the observed spectrum.

2.1.2 Cancellation by spectral models

Klapuri (2003) presented an iterative estimation algorithm based on two guiding principles: harmonicity and spectral smoothness. The input signal is preprocessed by a RASTA-like technique (Hermansky and Morgan, 1994) on a logarithmic frequency scale such that the spectral magnitudes are compressed and the additive noise is removed. Predominant-F0 estimation is based on summing the spectral ACF of the preprocessed spectrum across subbands. It is pointed out that the signal may become too corrupted after several iterations of direct cancellation. The predominant source is thus smoothed before being subtracted from the spectrum. In this way, the overlapping partials still retain energy for the remaining sources. The method, called the **bandwise smooth model**, uses the average amplitude within one octave band to smooth out the envelope of an extracted source.

A perceptually motivated multiple-F0 estimation method was later presented by Klapuri (2005). Subband signals are first compressed and half-wave rectified. Harmonic matching is then performed on the summary magnitude spectrum to extract the predominant F0. A **1/k smooth model** [1] is used to attenuate the predominant source such that the energy of higher partials is

[1] Partial amplitudes are inversely proportional to the partial index.

retained for the next iteration. Klapuri (2006) also proposed a spectral model which attempts to generalize a variety of musical instrument sounds. This model is found to be similar to the 1/k smooth model.

Bach and Jordan (2005) formulated the multiple-F0 estimation problem under a graphical model framework (Jordan, 2004). The spectral model is trained from speech database as a **spline smoothing model** and the predominant F0 is obtained by maximizing the likelihood. Pitch tracking is modeled as a factorial HMM. The algorithm iterates predominant-F0 tracking and subtraction till the designated number of F0s is achieved.

2.2 Joint Estimation

Contrary to the iterative estimation approach, joint estimation evaluates possible combinations of multiple F0 hypotheses without any cancellation involved. Although the observed signal is not corrupted as that in an iterative estimation-cancellation process, the handling of overlapping partials remains a challenge.

2.2.1 Joint cancellation

A joint cancellation method was proposed by de Cheveigné (1993). This method uses the **double difference function** (DDF) that jointly cancels multiple-F0 hypotheses. The hypothetical combination producing the smallest residual is considered the final estimate. The continuous studies show that joint cancellation performs better than iterative cancellation because a single-F0 estimation failure will lead to successive errors in an iterative manner (de Cheveigne and Kawahara, 1999). However, joint cancellation is computationally more demanding than iterative cancellation. Maher and Beauchamp (1994) proposed a *two-way mismatch* method to estimate two F0s jointly. The algorithm searches for the pair of F0s that minimize the frequency discrepancies between the harmonic models and the observed peaks, i. e., the mismatch from *the predicted to the measured* and the mismatch from *the measured to the predicted*. Each match is weighted by the amplitudes of the observed peaks. In this way, the algorithm minimizes the residual by the best match.

2.2.2 Polyphonic salience function

Polyphonic salience functions aim at enhancing the salience of the underlying F0s to facilitate a later peak-picking or tracking. Many salience functions follow the pitch perception model of Licklider (1951), which suggests an autocorrelation process after cochlear filtering. This auditory model leads to the channel-lag representation of ACF in the auditory channels (Lyon, 1984). This representation is called **correlogram** (Slaney and Lyon, 1990). Weintraub (1986) applied dynamic programming algorithms to correlogram and iteratively tracked the F0s of two speakers. Wu *et al.* (2003) followed the similar approach and applied channel selection along with channel peak selection before summing the normalized ACF across channels. Multiple F0s are then tracked for two speakers under a hidden Markov model scheme. Karjalainen and Tolonen (2000)

proposed to process a two-channel SACF with special techniques such that peaks corresponding to harmonics and subharmonics are suppressed. The resulting function is called **enhanced summary autocorrelation function** (ESACF).

The combination of several single-F0 estimation functions also yields a polyphonic salience function. Min *et al.* (1988) combined ACF with AMDF as the salience function, followed by a simple tracking technique. Peeters (2006) demonstrated that the combination of spectral ACF with cepstrum provides a useful polyphonic salience function for multiple-F0 estimation. Zhou (2006) presented a method to extract the power spectrum above the noise floor, called **resonator time-frequency image** (RTFI), from which **relative pitch energy spectrum** is derived for the selection of F0 candidates.

2.2.3 Spectral matching by non-parametric models

Static models are based on the assumption that the spectral pattern of a *fixed* harmonic structure is representative of one source even for its variants that evolve with time.

Non-negative matrix factorization

Considering the decomposition of the observed power spectra $\mathbf{Y}$ with the spectral templates $\mathbf{H}$:

$$\mathbf{Y} = \mathbf{WH} \tag{2.1}$$

where $\mathbf{W}$ is the weighting matrix. Smaragdis and Brown (2003) used **Non-negative Matrix Factorization** (NMF) to decompose the spectrogram into spectral models (basis functions in $\mathbf{H}$) of each note with its intensity change along time (weightings in $\mathbf{W}$). Since the components of $\mathbf{Y}$ are non-negative by nature, NMF approximates it as a product of two non-negative matrices $\mathbf{H}$ and $\mathbf{W}$. The cost function is designed to favor the minimization of the residual with specific constraints like **sparseness** (Cont, 2006) or harmonicity (Raczynski *et al.*, 2007). Although fast algorithms have been proposed for multiple-F0 estimation (Sha and Saul, 2005; Cont, 2006), the challenge remains in the modeling of the time-varying spectra of sound sources (Virtanen, 2003b; Abdallah and Plumbley, 2004).

Specmurt

Sagayama *et al.* (2004) understands the spectrum as a convolution of a common harmonic structure with pulses at multiple fundamental frequencies. The observed signal is first analyzed by a **constant-Q** like transform to fit the nature of energy distribution on the log-frequency scale. Spectral representation on log-frequency scale facilitates the spectral deconvolution because the common harmonic pattern can be *linearly shifted* and summed to match the observed spectrum. The log-frequency spectrum and the common harmonic pattern are both transformed (by inverse Fourier transform) into the **specmurt** domain [1] in which deconvolution of the spectrum can be simply achieved by division.

[1] Specmurt is defined as the inverse Fourier transform of linear spectrum with logarithmic frequency.

2.2.4 Statistical modelling using parametric models

The statistical approach formulates the problem within a Bayesian framework. Bayesian statistical methods provide a complete paradigm for both statistical inference and decision making under uncertainty. Waveform models adaptively match the observed compound waveform in the time domain. Walmsley *et al.* (1999) employs specific prior distributions for the existence of each sources, the fundamental frequencies, the number of partials, the partial amplitudes and the residual variance. These parameters are estimated jointly across a number of adjacent frames by means of the **Markov chain Monte Carlo** (MCMC) method. Davy and Godsill (2003) extended this method by introducing a prior distribution on the inharmonicity factor. In the generative music signal model proposed by Cemgil *et al.* (2006), a higher-level parameter related to tempo was further introduced with several modifications.

Spectral models adaptively match the observed signal in the frequency domain and the phase information is often disregarded. Goto (2000) regards the observed spectrum as a weighted sum of harmonic-structure tone models. The signal parameters are estimated through the EM algorithm. Following the same concept, Kameoka *et al.* (2005a) formulates the multiple-F0 estimation problem as a time-space clustering of harmonic sounds, which is named **harmonic temporal structured clustering** (HTC) method. HTC method models the fundamental frequency, the relative partial amplitudes, the intensity, the onset and the duration, etc., of each underlying source. All the parameters are optimized by the EM algorithm (Kameoka *et al.*, 2005b) such that the superimposed HTC models approximate the observed spectrogram the best. Partials of a harmonic source are modeled as Gaussian distributions with initial spectral envelopes. The evolution of partial amplitudes are modeled by Gaussian mixtures across frames such that the synchronous evolution is constrained and the duration is adaptively modeled. Vincent (2004) models the spectra of musical instrument sounds by means of the means and the variances of partial amplitudes, partial frequencies and residuals. The observed spectrum can thus be interpreted as a sum of the spectral models with the related weighting optimized by Newton's method. A factorial model is applied to constraining the temporal continuity and to adapting the duration.

2.2.5 Blackboard system

A **blackboard system** integrates various forms of knowledge or information for solving complicated problems. In general, a blackboard system for **Auditory Scene Analysis** consists of a three-level process: low-level signal processing, mid-level grouping, and high-level stream forming. Low-level processing extracts signal components such as spectral peaks, transients (onsets/offsets), amplitude modulation and frequency modulation. Mid-level grouping interprets the signal components as features for fusion or segregation. Signal components with harmonic relation, common onsets or amplitude modulations, etc., can be clustered in one group. These grouping cues are based on the psychological findings of human perception of auditory scenes (Bregman, 1990). This approach has been widely accepted for automatic music transcription of *note pitches* (Chafe and Jaffe, 1986; Mellinger, 1991; Kashino and Tanaka, 1993; Martin, 1996;

Fernandez-Cid and Casajus-Quiros, 1998; Kashino *et al.*, 1998; Sterian, 1999; Dixon, 2000; Baumann, 2001; Chien and Jeng, 2002) in which F0s are low-level representation of *note pitches*. Higher level stream forming can incorporate prior knowledge about instrument models and music styles (the key, the scale, the tempo, etc.) to segregate objects in the same group or to eliminate less plausible groups.

2.3 On Estimating the Number of Sources

The estimation of the number of harmonic sources, called **polyphony inference**, is one of the most difficult problem for multiple-F0 estimation. Some existing methods assume that the number of sources is known. There exist few studies of polyphony inference, most of which usually rely on the threshold on a specific criterion. These criteria are summarized in the following.

1. Minimal residual
   On the assumption that the noise part is relatively low in energy compared to the periodic part, it is reasonable to minimize the residual by maximizing the energy explained by the signal model (Klapuri, 2005). The main concern is the spectral diversity of musical instrument sounds which often differ from the harmonic model. According to the description in Section 1.3.1, inharmonic partials and spurious components can not be perfectly modeled. When the energy of these components is larger than the allowed threshold on the residual, they are extracted as additional sources.

2. Minimal source energy
   To avoid the over-estimation of polyphony, a constraint on the energy of the extracted sources could be set to limit spurious detections. For a blackboard system, it is often practical to use a threshold in the segregated audio streams to prune spurious sources. However, the minimal source energy varies for signals under different SNR conditions and an adaptive threshold is needed (Kameoka *et al.*, 2005b).

3. Information criterion
   For the statistical modeling approach, there exist several information criteria such as AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion) to estimate the *model order*. In the context of multiple-F0 estimation, the model order is related to the number of sources. Information criteria usually combine the modeling error term with a penalty term (Hayes, 1996). The minimum of information criterion should represent a good balance of low model errors and the low model order. Several model orders might have to be assessed before the minimum of the information criterion can be located (Kameoka *et al.*, 2004).

4. Salience improvement
   Similar to the concept of information criterion, one can observe the salience improvements while the polyphony hypothesis is increased (Klapuri, 2006; de Cheveigné and Baskind, 2003). Adding a correct source should improve the salience more significantly than adding

an incorrect one. Compared with the information criterion, this approach, although heuristic in its nature, is more flexible for the integration of various information to validate the hypotheses.

5. Temporal modeling
   Temporal modeling integrates the observed F0 hypotheses into a statistical framework to build up continuous F0 trajectories. It is helpful for a frame-based F0 estimator to correct the errors when the estimated F0 changes abruptly. Wu *et al.* (2003) proposed to track multiple F0s within a hidden Markov model (HMM) in which the number of sources is modeled as discrete states. Transition probability between states are learned from training databases and the optimal sequence of successive states is searched by Viterbi algorithm. Ryynänen and Klapuri (2005) presented a more sophisticated algorithm which applies a note event HMM, a silence model and a musicological model to the tracking of multiple F0s.

2.4 Discussions

2.4.1 Signal representation

Most multiple-F0 estimation algorithms involve the analyses in the spectral domain. The advantage of the spectral representation is that it provides an intuitive representation of the harmonic structure and the spectral envelope of a sound source, which facilitates the modeling of sound sources. Spectral representation can be either **multi-resolution** or **fixed-resolution**.

Multi-resolution transform represents the signal with different frequency resolutions for different frequency bands. It is widely used for polyphonic music transcription (Chafe and Jaffe, 1986; Keren *et al.*, 1998; Fernandez-Cid and Casajus-Quiros, 1998; Chien and Jeng, 2002; Kobzantsev *et al.*, 2005). The constant-Q transform (Brown, 1991) is a multi-resolution transform dedicated to music signals, especially for the music based on the equal tempered scale. **Q ratio** is defined as the center frequency divided by the required bandwidth, for example, $Q = 1/(2^{24} - 1) \approx 34$ for a quarter tone resolution [1]. The kernel function of constant-Q transform can be sinusoids or other wavelets such as Gabor (Kameoka *et al.*, 2007). Similar to the multi-resolution representation, the **auditory model representation** emulates the humane hearing system of which the cochlea simulation converts the acoustic waves into a multi-channel representation of basilar membrane motion. The auditory model representation thus involves multiple subband filtering, called **cochlear filtering**, which is often approximated by the gammatone filter (Patterson and Holdsworth, 1990; Wu *et al.*, 2003; Marolt, 2004; Klapuri, 2005). The center frequencies are often selected to have a uniform distribution on a **critical band** scale. In this way, the frequency resolution is adapted to the human perception. Multiple-F0 estimation algorithms based on such a representation are called **multi-pitch estimation** algorithms.

The argument usually made for multi-resolution representation is based on its similarity to the equal tempered scale or the human auditory system. However, there exist no physical reasons

[1] 1/24-octave filter bank

that a multi-resolution transform is better in representing the structure of a harmonic sound. The advantage of multi-resolution also seems to be its disadvantage: the frequency resolution at high frequencies is sacrificed and individual partials of concurrent sources may not be distinguishable. Since the problem of overlapping partials is to be addressed, it is important to maintain the access to individual partials as much as possible. In this thesis, the STFT (Short-Time Fourier Transform) is chosen for signal representation. Although being criticized for its fixed-resolution, the STFT lays out the partials of a harmonic sound equally on the frequency axis, which provides an intuitive analysis of harmonic sounds.

2.4.2 Iterative estimation or joint estimation

Polyphonic signals differ from monophonic signals in that the components of concurrent sources may overlap. In general, iterative estimation algorithms attenuate the predominant source at each iteration. Since the subsequently extracted F0s are unknown upon the attenuation/cancellation process, it is almost impossible to estimate where the partials overlap and to prevent over-attenuation of partials. On the other hand, the joint estimation approach has the advantage that the overlapping partials can be inferred from a set of hypothetical sources. A more optimal treatment of overlapping partials can thus be expected. However, the downside of the joint estimation approach is its computational cost because the number of hypothetical combinations grows exponentially with the polyphony hypothesis. A possible solution is to combine both approaches to reach a compromise between the efficiency and the robustness of a multiple-F0 estimation system.

2.4.3 HRF0s and NHRF0s

Although the methods differ in various aspects, it is suggested to focus on the point of view of the effectiveness in extracting two groups of F0s: **harmonically related F0s** (HRF0s) and **non-harmonically related F0s** (NHRF0s). Consider two F0s $f0_H > f0_L$, both of them are considered NHRF0s if they have the following relation

$$f0_H = \frac{m}{n} f0_L, m \in \mathbb{N}, n \in \mathbb{N}, n > 1 \tag{2.2}$$

and the fraction $\frac{m}{n}$ is coprime. On the other hand, $f0_H$ is considered HRF0s if

$$f0_H = m \cdot f0_L, m \in \mathbb{N} \tag{2.3}$$

Theoretically, the harmonics of $f0_H$ overlap completely with $f0_L$. The extraction of NHRF0s and HRF0s is closely related to the fundamental assumptions of the noise model, the source model and the source interaction model (see Section 1.3.2).

The noise model is closely related to the extraction of NHRF0s. The combination of NHRF0s should explain as much as possible the periodic energy of the observed signal. If no distinction is made between the periodic part and the noise, a fixed threshold on minimal residual could either ignore weak sources or add spurious sources. Therefore, noise estimation is important for

a multiple-F0 estimation system to robustly extract NHRF0s. Different from NHRF0s, HRF0s are not much related to the noise model, but are related to the source model and the source interaction model, instead. Since the harmonics of a HRF0 may overlap completely with another source of a lower F0, the harmonics can be observed only when the energy is relatively high such that they stand out of the envelope of the overlapped source at regular harmonic intervals. Therefore, spectral smoothness principle and overlapping partial treatment are crucial to robust extraction of HRF0s.

3

PROPOSED METHOD

The objective of this thesis is to develop a frame-based multiple-F0 estimation system for monaural music signals produced by harmonic musical instruments. The observed signal is represented by the STFT, and it is assumed that the number of sources is fixed in the short-time analysis frame. The problems involving the three fundamental models, the noise model, the source model and the source interaction model, are to be dealt with in the respective parts of the thesis. The proposed method handles the three fundamental models in a way different from the existing methods for multiple-F0 estimation. Contrary to many state-of-the-art methods that do not use an explicit model for the noise part of the signal, this thesis proposes a probabilistic description of the noise level based on which the hypothetical sources are extracted. Noise modeling is meant to distinguish the components that are not necessary to be explained by a set of the hypothetical sources. The source model of the proposed method is a quasi-harmonic model without specific amplitudes. The partial frequencies and amplitudes of hypothetical sources are estimated by harmonic matching and overlap treatment. In order to correctly handle the overlapping partials, which is related to the source interaction model, the joint estimation approach is selected to evaluate a combination of hypothetical sources.

In this chapter, the signal model for the proposed multiple-F0 estimation algorithm is presented. Then, three principles: harmonicity, the smoothness of spectral envelopes and the synchronous amplitude evolution of partials, which guide the development of the system are described. They are the general assumptions of the physical properties of harmonic instrument sounds. Finally, an overview of the proposed system is given.

3.1 Generative Signal Model

The polyphonic signals comprising of multiple harmonic sound sources can be expressed as eq.(1.16). In order to model quasi-periodic and quasi-stationary sources as general as possible, several parameters are included in the generative signal model. The observed signal $y[n]$ in a short-time analysis frame is expressed as:

$$\begin{aligned} y[n] &= \sum_{m=1}^{M} y_m[n] + z[n] \\ &= \sum_{m=1}^{M} \sum_{h=1}^{H_m} a_{m,h}[n] \cos\left(\theta_{m,h}[n] + \phi_{m,h}\right) + z[n], \quad \theta_{m,h}[n] = (1+\delta_{m,h})h\omega_m n \end{aligned} \tag{3.1}$$

where n is the discrete time index, M is the number of concurrent sources and $y_m[n]$ is the **quasi-periodic part** of the mth source. H_m is the number of partials for the mth source. The hth partial of the mth source can be modeled as a sinusoid with amplitude $a_{m,h}[n]$, initial phase $\phi_{m,h}$ and frequency $(1+\delta_{m,h})h\omega_m$. ω_m represents the fundamental frequency which is the mean frequency of the time-varying phase $\theta_{m,h}[n]$. $\delta_{m,h}$ models the deviation of a partial frequency from its corresponding harmonic position. The **noise part** $z[n]$ is explained by a **generative noise model** in which the noise is understood as generated from white noise signal filtered by a frequency-dependent spectral envelope with a limited cepstral order.

The signal model expressed in eq.(3.1) specifies the signal components in the time domain. The Fourier transform of the windowed $y[n]$ can be expressed as

$$Y(\omega) = \sum_{m=1}^{M} \sum_{h=1}^{H_m} A_{m,h} e^{j\phi_0} W(j(\omega - (1+\delta_{m,h})h\omega_m)) + Z(\omega), \quad \text{for } \omega > 0 \tag{3.2}$$

where W is the Fourier transform of the window function and $A_{m,h} = \frac{a_{m,h}}{2}$ assuming $a_{m,h}[n]$ constant. The observed spectrum is understood as generated by sinusoids and noise (Doval and Rodet, 1991). All necessary information for F0 estimation is to be extracted from the properties of spectral peaks. The observed spectrum is thus modeled as a cluster of successive spectral peaks. A spectral peak is defined as the spectral regions between two neighboring spectral valleys in which all the frequency bins are considered to belong to this peak. The amplitude or the frequency of a spectral peak is defined by that of the bin with the maximal amplitude. The other extended properties of a spectral peak such as *normalized mean bandwidth*, *duration*, and *frequency coherence* are also extracted (see Appendix B). Each peak is considered either sinusoid [1] or noise. That is, non-sinusoidal components like side-lobe peaks are considered noise. Based on this model and given the observed spectrum, the most plausible F0 hypotheses are to be inferred.

[1] A sinusoid peak could be single sinusoid or a sum of sinusoids at close frequencies

3.2 Guiding Principles

Based on the generative signal model, the algorithms developed in this thesis follow three guiding principles: harmonicity, the smoothness of spectral envelopes, and the synchronous evolution of partial amplitudes within a source. These principles are the general assumptions of the physical properties of harmonic instrument sounds.

Harmonicity

The harmonicity/periodicity principle inherits from the problem definition of F0 estimation. F0 estimation algorithms, either using explicitly a quasi-periodic signal model or not, are based on the harmonicity principle that there is a repeated pattern in the observed signal. As long as the noise part $z[n]$ is effectively identified, a set of hypothetical sources should collectively explain as good as possible the components that are identified as sinusoids.

Smoothness of spectral envelopes

Few single-F0 estimation algorithms solely based on the harmonicity principle can achieve robust performance. Several techniques are developed to prevent subharmonic errors as well as super-harmonic errors. To prevent subharmonic errors, one technique is to examine the spectral envelopes of the F0 hypotheses with competitive harmonicity, and smoother envelopes are preferred. This so-called *spectral smoothness* principle (Klapuri, 2003) is based on the assumption that the spectral envelope of a harmonic instrument sound generally form a smooth contour. To estimate the F0s of concurrent harmonic sounds, this principle is particularly useful for resolving the ambiguities in the overlapping partials and for extracting HRF0s. The function of the spectral smoothness principle can be interpreted with respect to two aspects: (1) When an abrupt change is found in the spectral envelope of the hypothetical source, it can be the cue for the partial collisions, according to which the partial segregation can be carried out. (2) Subharmonic errors can be prevented because their spectral envelopes are less smooth than those related to the correct F0s.

Synchronous amplitude evolution of partials

Partials of a harmonic source should evolve in time in a similar manner because they are generated by a common control such as the bowing of the strings. Although the partials of a harmonic instrument sound may decay at different rates, the assumption of their synchronous evolution is generally useful for the segregation of sound sources. For the case in which the analysis is carried out on the time-frequency plane, this principle applies to the grouping of the partial tracks evolving synchronously as a harmonic source. For the case that the analysis is carried out exclusively at a single frame, it is possible to estimate the direction of evolution for each spectral component (Röbel, 2003a). This principle discriminates sinusoids from noise in that noise has random amplitude evolution.

The three guiding principles concerning the physical properties of harmonic instrument sounds are closely related to source segregation in auditory scene analysis (Bregman, 1990). The respective rules are: (1) A set of harmonics is not heard as a number of individual components but as one single sound source. (2) The smoothness of the spectral envelopes seems to be important for us to perceive them as one source. The partials that are raised in intensity will segregate more readily from the others. (3) Common fate cues promote the grouping of partials into one source. The cues can be correlated frequency changes, correlated amplitude changes and synchronous onsets.

3.3 System Overview

To develop a multiple-F0 estimation system, three major problems are identified as the main tasks:

1. **Noise modeling**: Derive a description of the noise level as the spectral envelope of noise components, which facilitates the identification of the sinusoids and the noise.
2. **Joint evaluation of F0 hypotheses**: Develop a joint estimation algorithm for the case when the number of concurrent sources is given.
3. **Polyphony inference**: Estimate the number of concurrent sources along with the related F0s.

Based on the generative signal model, the guiding principles, and the model assumptions (see Table 3.1), the algorithms for addressing the three problems are developed from Chapter 4 to Chapter 7. An overview of the proposed system is given in the following (see Figure 3.1).

At each analysis frame, FFT (Fast Fourier Transform) is applied to the observed signal to obtain the instantaneous spectrum. The observed spectrum is characterized by the time-frequency properties of a collection of spectral peaks. The generative noise model represents the noise envelope as a frequency-dependent spectral envelope with a limited cepstral order. The noise is considered having a nearly constant expected magnitude within a narrow band and the noise level is defined by successive Rayleigh distributions. The frequency-dependant noise level is estimated through iterative peak classification and distribution fit. Based on the assumption that the spectral peaks are of two classes: sinusoid and noise, the spectral peaks are classified by a probabilistic threshold relative to the noise level (see Chapter 4). A set of hypothetical sources

observed signal	represented as a cluster of successive spectral peaks classified as sinusoid/noise
noise model	frequency-dependent envelope of limited cepstral order applied to white noise
source model	quasi-harmonic model without specific amplitudes
interaction model	the amplitude of overlapping partials determined by the strongest source

Table 3.1: Assumptions of the proposed method.

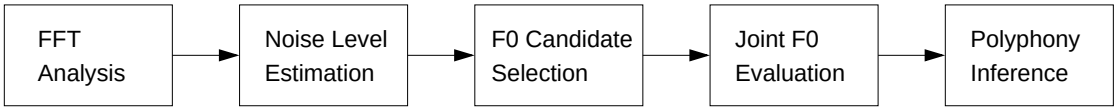


Figure 3.1: Overview of the proposed multiple-F0 estimation system.

should explain as many sinusoidal peaks as possible. For a combination of hypothetical sources, the related hypothetical partial sequences are constructed by a partial selection technique and a treatment of overlapping partials. To evaluate the plausibility of a set of hypothetical sources, a score function is proposed, which formulates the guiding principles into four criteria (see Chapter 5). To improve the efficiency of the joint estimation process, three methods for the selection of F0 candidates are studied, aiming at the reduction of the number of combinations (see Chapter6). Finally, the F0 hypotheses are consolidated to yield the most plausible set of F0s (see Section 7.1).

4

ADAPTIVE NOISE LEVEL ESTIMATION

Given a polyphonic signal composed of harmonic instrument sounds, the generative signal model explains the quasi-periodic part and the noise part by a sum of quasi-harmonic sources and by a generative noise model, respectively. The approach of the thesis is to distinguish the noise components from the sinusoidal components beforehand. Based on the preliminary estimation of noise, the match between the identified sinusoidal components and a set of hypothetical sources can be carried out afterwards. The generative noise model is described by a frequency-dependent spectral curve approximating the noise level. The noise components are assumed to be generated by filtering a white noise signal with a frequency-dependant smooth function. The estimation of the noise level is important for the estimation of NHRF0s (non-harmonically related F0s). If the energy of a hypothetical source of NHRF0s is significant compared with that of noise, it can be considered a reasonable estimate. Otherwise, it probably corresponds to a spurious detection.

In this chapter, a novel algorithm is presented for the estimation of the colored noise level for general audio signals. The modeling of the noise level is based on the assumptions that the spectral envelope of noise varies slowly in frequency and that the amplitudes of the noise peaks obey a Rayleigh distribution. By means of an iterative evaluation and adaptation of the noise level, the sinusoid/noise classification is gradually refined until the identified noise peaks are coherently explained by the estimated noise level. The evaluation of the proposed algorithm is demonstrated at the end of the chapter.

4.1 Generative Noise Model

A signal is called **white noise** if the knowledge of the past samples does not tell anything about the subsequent samples to come. The power density spectrum of white noise is constant. By means of filtering a white noise signal, correlations between the samples are introduced. Since in most cases the power density spectrum will no longer be constant, a filtered white noise signal is generally called a **colored noise** signal. With respect to statistical time invariance property, the noise may be stationary or non-stationary.

The noise is understood as generated from white noise filtered by a frequency-dependent spectral envelope which is called the **colored noise level**, or simply the **noise level**. The noise level estimation must be adaptive in time and in frequency such that non-stationary noise and colored noise can be dealt with. In this thesis, the noise level is defined as the *expected magnitude level of the observed noise peaks*. A noise peak is defined as a peak that can not be explained as a stationary or weakly modulated sinusoid. The noise level can be represented as a smooth frequency-dependent curve approximating the noise spectrum (de Krom, 1993) (see Figure 4.1). The envelope of the noise spectrum, covering most of the noise peaks, can then be related to the noise level by a constant raise in magnitude.

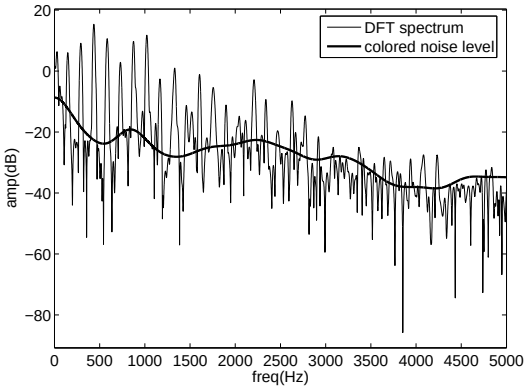


Figure 4.1: The colored noise level estimated for a real-world spectrum.

4.2 Existing Approaches to Noise Level Estimation

Noise level estimation, or noise power spectral density estimation, is usually done by explicit detection of the time segments that contain only noise, or by explicit estimation of harmonically related spectral components (for quasi-periodic signals). Since the noise components often come along with the sound source signal, the estimation of noise carried out for the noise-only segments would introduce a systematic bias. The other approach is to estimate the F0s beforehand. The harmonic parts can be extracted and the noise part can be estimated by subtracting harmonic parts from the signal. However, this is difficult for multiple-F0 estimation and not appropriate for non-harmonic sources.

A widely appreciated approach is the analysis of noise statistics. This is carried out by the statistical analysis of power spectra across consecutive frames. This approach often assumes that the analysis segment contains low energy portions and the noise present is more stationary than the embedded signal (Ris and Dupont, 2001). **Minimum Statistics** method proposed by Martin (1994) tracks the minimal values of a smoothed periodogram. He introduced later the time-varying smoothing factor for periodograms and bias compensation (Martin, 2001). Cohen (2003) followed this approach and proposed time-varying and frequency-dependent smoothing

for periodograms.

de Krom (1993) proposed to estimate the **noise floor** of a spectrum by cepstral liftering and spectral baseline shifting. Cepstral liftering is used to remove harmonic components in the spectrum, and baseline shifting is used to position the noise floor at a reasonable reference level. The noise part is estimated by the inverse transformation of the cepstrum, followed by a heuristic baseline shift. Qi and Hillman (1997) further proposed to lifter out high-quefrency parts of the cepstrum such that a smooth curve is obtained after transforming it back to the spectral domain. However, this approach relies on identifying harmonic peaks of the underlying source.

The other classical approach is to remove the sinusoids and estimate the noise afterwards. This involves sinusoidal component identification, either in a single frame (Peeters and Rodet, 1998; Hainsworth *et al.*, 2001; Röbel and Zivanovic, 2004) or by tracking sinusoidal components across several frames (David *et al.*, 2003; Lagrange *et al.*, 2005). Parametric methods usually model the observed signal as several sinusoidal components embedded in additive white Gaussian noise. The challenge faced by the parametric methods is the estimation of the model order, i.e., the number of sinusoids. Badeau *et al.* (2004) pointed out that classical methods such as maximum likelihood estimation rely on additive white noise hypothesis and tend to overestimate the model order in the presence of correlated noise. Non-parametric methods include the periodogram-based techniques and the classification of sinusoidal/noise peaks. Since the non-parametric methods is not constrained by the white noise assumption, they are of a greater freedom in dealing with modulated sinusoids embedded in the colored noise. No matter which methods, the common difficulty is in identifying sinusoids in polyphonic signals because the sinusoidal components may collide and its sinusoidality is ambiguous.

The proposed algorithm follows the non-parametric methods of Qi and Hillman (1997) and Röbel and Zivanovic (2004) to classify the spectral peaks. The advantage is to have relaxed assumptions compared to the maximum likelihood method and those reviewed by Ris and Dupont (2001). To develop a frame-based multiple-F0 estimation system, it is proposed to classify the spectral peaks in each short-time analysis frame independently. Accordingly, the non-stationary noise is assumed to be handled based on the frame-by-frame analysis. The costly tracking of sinusoidal components across the frames can then be avoided. Moreover, the spectral peak classification method proposed by Röbel and Zivanovic (2004) allows the control of the classification results such that a bias towards sinusoids or noise can be easily altered. However, this method is not suitable for polyphonic signals in which the sinusoids might collide with one another. A part of the spectral information is then ambiguous. In order to improve the spectral peak classification in polyphonic cases, an adaptive noise level estimation algorithm is developed in the following.

4.3 Modelling Narrow Band Noise

Since the noise level is frequency dependant, it is not appropriate to assume that noise is white. A more reasonable assumption is that noise has a nearly constant expected magnitude within a narrow frequency band. The assumption is thus relaxed. Accordingly, the Rayleigh distribution

may fit the distribution of the magnitudes of the noise components in each narrow band (see Appendix A). Consider a Gaussian white noise process with variance σ, the amplitude distribution of its Fourier spectrum follows Rayleigh distribution(see Figure 4.2).

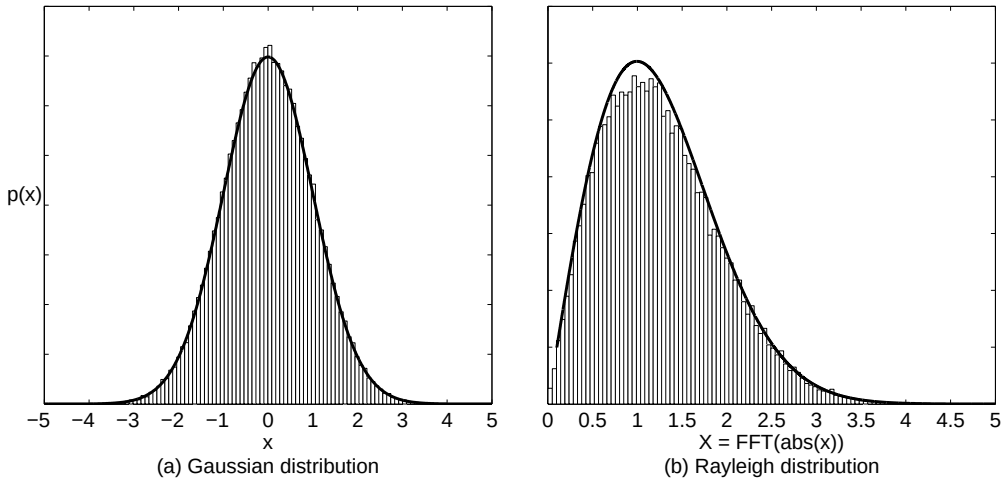


Figure 4.2: (a) Amplitude distribution of Gaussian white noise with $\sigma = 1$; (b) Spectral magnitude distribution of Gaussian white noise obeys Rayleigh distribution.

The Rayleigh distribution was originally derived by Lord Rayleigh in connection with a problem in the field of acoustics. A Rayleigh random variable X has probability density function (Johnson *et al.*, 1994):

$$p(x) = \frac{x}{\sigma^2} e^{-x^2/(2\sigma^2)} \quad \text{with } 0 \leqslant x < \infty, \sigma > 0 \tag{4.1}$$

cumulative distribution function

$$F(x) = 1 - e^{-x^2/(2\sigma^2)} \tag{4.2}$$

and the pth percentile is

$$x_p = F^{-1}(p) = \sigma\sqrt{-2\log(1-p)}, \quad 0 < p < 1 \tag{4.3}$$

In Figure 4.3, the probability density function is plotted for different values of σ ($\sigma = 0.5, 1, 1.5, 2, 2.5$ and 3). σ corresponds to the **mode** of the Rayleigh distribution, which is the most frequently observed value in X. Hence, $p(\sigma)$ corresponds to the maximum of the probability density function. Consider the Rayleigh random variable X as the observed magnitudes of spectral peaks in a narrow band, then σ represents the most frequent magnitude values of noise peaks (see Figure 4.4). Accordingly, the mode of the Rayleigh distribution can be used to derive the probability of an observed peak belonging to the background noise process. The peaks with the magnitudes smaller than σ are considered more likely to be noise. For the peaks with the magnitudes larger than σ, the larger their magnitudes, the less probable they are noise.

For a given narrow band, e.g. each frequency bin k, the noise distribution can be modeled by

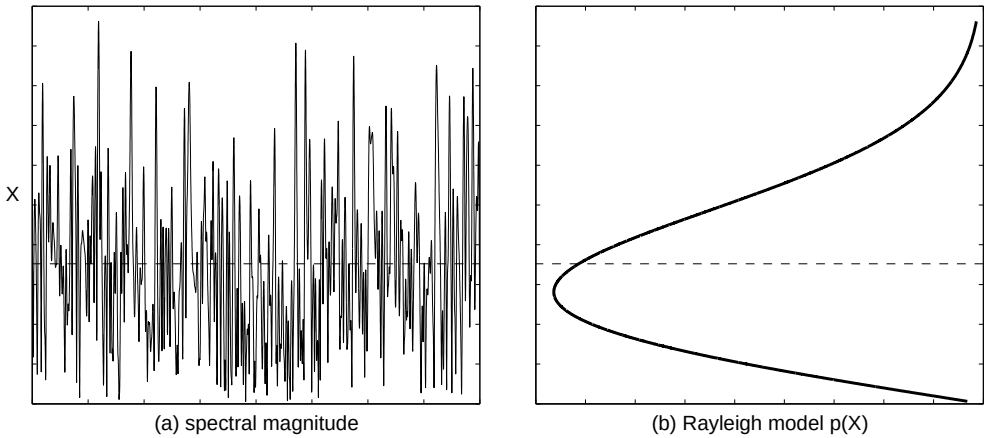


Figure 4.4: The probabilistic Rayleigh model for the spectral magnitude distribution of Gaussian white noise. The dash line represents the mean magnitude.

means of a Rayleigh distribution with mode $\sigma(k)$. Once $\sigma(k)$ is estimated for each k, the curve passing through these σ-value magnitudes represents the *estimated Rayleigh mode*, denoted by $\mathcal{L}_\sigma(k)$. With the estimated Rayleigh mode, the noise threshold can be adjusted according to a desired percentage of misclassified noise peaks (see eq.(4.3)). The related **noise envelope** $\mathcal{L}_n$ can then be calculated by multiplying $\mathcal{L}_\sigma$ with $\sqrt{-2\log(1-p)}$. Therefore, the problem is the estimation of the frequency dependent $\sigma(k)$.

It is known that the mean of a Rayleigh random variable X is

$$E[X] = \sigma\sqrt{\pi/2} \tag{4.4}$$

or equivalently

$$\sigma = \frac{E[X]}{\sqrt{\pi/2}} \tag{4.5}$$

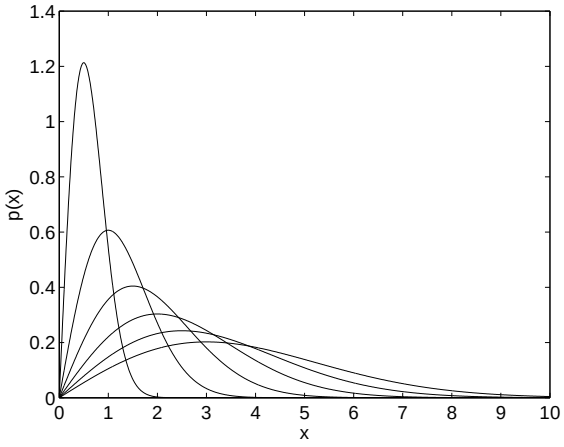


Figure 4.3: Rayleigh distributions with different σ values.

That is, the frequency dependent $\sigma(k)$ can be calculated if the mean noise magnitude $E[X]$, which is also frequency dependent, can be estimated. An intuitive way to estimate $E[X]$ is to collect, for each bin, sufficient observations of the noise magnitudes and to estimate the mean value. However, this approach is not viable for a frame-based analysis algorithm. The chosen approach is to estimate the noise level as the cepstrally liftered curve and then to relate it to the **mean noise level** $\mathcal{L}_m$. The advantage of this approach is that the statistics of individual spectral bins are not evaluated separately. As long as there are sufficient noise peaks in the observed spectrum, the cepstrally liftered curve can be estimated as a smooth envelope across all the spectral bins (Qi and Hillman, 1997). The mean noise magnitude of each bin can thus be estimated. This approach is based on

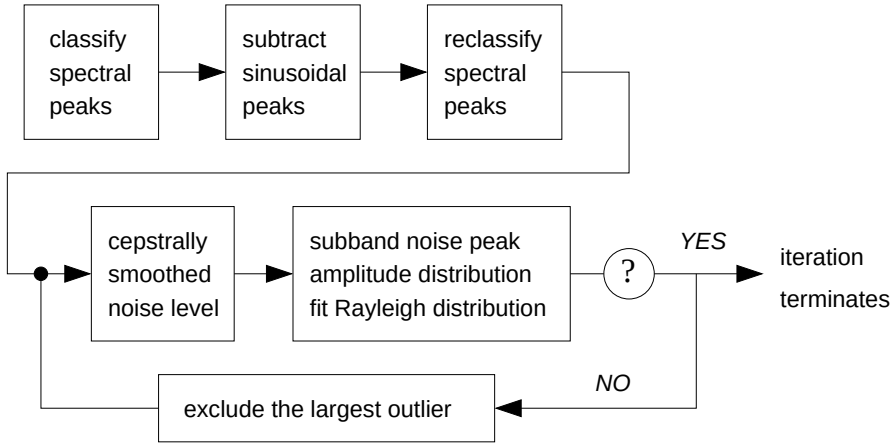


Figure 4.5: Overview of the adaptive noise level estimation algorithm.

the assumption that the noise level is changing only slowly with the bin index k to the extent that the cepstrally liftered curve describes its variation with frequency. Based on the noise model and the assumptions, the adaptive noise estimation algorithm is presented in the next section.

4.4 Iterative Approximation of the Noise Level

The algorithm starts with the estimation of the residual spectrum (see Figure 4.5). There are two processes involved: the classification of spectral peaks and the subtraction of sinusoids. For the classification of spectral peaks, three spectral peak descriptors proposed by Röbel and Zivanovic (2004) are used because they are designed to properly deal with non-stationary sinusoids (see Appendix B). For the estimation of sinusoidal parameters, the method proposed by Abe and Smith (2005) is chosen because it provides an efficient and accurate estimation for weakly modulated sinusoids (see Appendix C). After the initial classification of spectral peaks, the sinusoids are subtracted from the observed signal. Because both methods are not meant to deal with closely spaced sinusoids, some sinusoids may not be correctly classified or effectively removed. The estimation error due to the model inconsistency (single sinusoid instead of multiple sinusoids) is expected to be compensated in the later stages. The main function of subtracting sinusoids is to provide sufficient residual peaks for a proper statistical measure of the magnitude distribution, even if the frequency resolution is limited and sinusoidal peaks are very dense.

Once the residual spectrum S_R is obtained, the classification of spectral peak is carried out again, followed by the process of the iterative approximation of the noise level (see Figure 4.5). The reason for the reclassification of spectral peaks is that the spectrum has changed after the subtraction of sinusoidal peaks. The residual spectrum is divided into subbands of an equal bandwidth [1]. When the noise distributions in all subbands fit the related Rayleigh distributions, the iteration process terminates. The evaluation of the distribution fit is achieved by using a statistical measure because: (1) the amount of the observed samples is usually not large enough to draw the underlying distribution and (2) statistical measures are representative

[1] For an analysis frequency up to $8k$Hz, the spectrum is equally divided into 25 subbands.

of a distribution and are more efficient for verifying the underlying distribution. To this end, **skewness** is selected for the distribution fit. Skewness is a measure of the degree of asymmetry of a distribution (Stuart and Ord, 1998). If the right tail (tail at the large end of the distribution) extends more than the left tail does, the skewness of the observed samples is positive. If the reverse is true, the skewness is negative. If the two tails extend symmetrically, the skewness becomes zero, e.g. Gaussian distribution. The skewness of a distribution is defined as

$$Skw(X) = \frac{\mu_3}{\mu_2^{3/2}} \tag{4.6}$$

where μ_i is the ith central moment. The skewness of Rayleigh distribution is independent of $\sigma(k)$:

$$Skw_{rayl} = \frac{2(\pi-3)\sqrt{\pi}}{\sqrt{(4-\pi)^3}} \approxeq 0.6311 \tag{4.7}$$

If the distribution of the noise magnitudes in a subband is assumed Rayleigh, the remaining sinusoids can be tested by means of the skewness of the magnitude distribution. Since the number of the observed samples is limited to a small number, the biased estimation needs to be corrected. Besides, the small-number observation may result in positive or negative skewness. If there are more sinusoidal peaks (than noise peaks) with significant amplitudes within a subband, the skewness will be negative. Therefore, the following condition: $0 < Skw(\{X_n\}_b) \cdot C_{skw} < Skw_{rayl}$ is proposed for a better convergence, where $\{X_n\}_b$ are the set of noise peak magnitudes in the bth subband, and $C_{skw} = \frac{\sqrt{N_s(N_s-1)}}{N_s-2}$ is the correction factor given the number of samples N_s from a population. Another issue is that, however, the assumption that the narrow band noise has a nearly constant expected magnitude may not hold true for the selected bandwidth of a subband. If $\sigma(k)$ in the subband is not constant, which is the case for the colored noise, the distribution of noise magnitudes in the subband will not be Rayleigh. To improve the consistency of the skewness test, the noise magnitudes are rescaled by means of normalizing X_n with the current estimated Rayleigh mode $\mathcal{L}_\sigma$. Accordingly, the distribution fit condition is

$$0 < Skw(\{\frac{X_n(k)}{\sigma(k)}\}_b) \cdot C_{skw} < Skw_{rayl} \tag{4.8}$$

The noise level approximation can be realized by iterating the following processes:

1. Calculate the cepstrum of the *noise spectrum*[1]. The cepstrum is the inverse Fourier transform of the log-magnitude spectrum. The dth cepstral coefficient is formulated as

$$c_d = \frac{1}{2}\int_{-\pi}^{\pi} \log(X_n(\omega))e^{i\omega d}d\omega \tag{4.9}$$

   By truncating the cepstrum and using the first D cepstral coefficients, we reconstruct a smooth curve representing the mean noise level $\mathcal{L}_{mlg}$ (on the logarithmic scale) as a sum

[1] It is constructed from interpolating the magnitudes of noise peaks.

$\mathcal{L}_{mlg}$	the mean noise level on the log amplitude scale	eq.(4.10)
$\mathcal{L}_{\sigma}$	the estimated Rayleigh mode	eq.(4.13)
$\mathcal{L}_{m}$	the mean noise level on the linear amplitude scale	eq.(4.14)
$\mathcal{L}_{n}$	the noise envelope/threshold	eq.(4.15)

Table 4.1: Summary of the noise levels.

of the slowly-varying components.

$$\mathcal{L}_{mlg}(\omega) = \exp(c_0 + 2\sum_{d=1}^{D-1} c_d \cos(\omega d)) \tag{4.10}$$

The cepstral order D is determined by (Röbel and Rodet, 2005): $D = F_s/max(\Delta f_{max}, BW) \cdot C$, where F_s is half the sampling frequency, Δf_{max} is the maximal frequency gap between the consecutive noise peaks, BW is the subband bandwidth, and C is a parameter to set.

2. The mean noise level $\mathcal{L}_{mlg}$ is the expected value of *log amplitudes* $\log(X_n)$. It is necessary to correct it as the expected value of *linear amplitudes* X_n. Assuming $Y = \log X \triangleq \Phi(X)$, we can calculate the expected value of Y (Rivet *et al.*, 2007) :

$$\begin{aligned}
E[Y] &= \int_{-\infty}^{\infty} yp(y)dy = \int_{-\infty}^{\infty} yp_x(\Phi^{-1}(y))|\frac{d\Phi^{-1}(y)}{dy}|dy \\
&= \int_{-\infty}^{\infty} y\frac{e^{2y}}{\sigma^2}e^{-\frac{e^{2y}}{2\sigma^2}}dy \\
&= \int_{0}^{\infty} \log(x)\frac{x}{\sigma^2}e^{-\frac{x^2}{2\sigma^2}}dx \\
&= \int_{0}^{\infty} (\log(\sigma) + \log(y))ye^{-y^2/2}dy, \text{ where } \ y = x/\sigma \\
&= \log\sigma + \int_{0}^{\infty} \log(y)ye^{-y^2/2}dy \\
&= \log\sigma + \int_{0}^{\infty} \frac{1}{2}\log(y^2/2)e^{-y^2/2}ydy + \int_{0}^{\infty} \frac{\log(2)}{2}e^{-y^2/2}ydy \\
&= \log(\sigma) - \frac{\gamma}{2} + \frac{\log(2)}{2} \approx \log(\sigma) + 0.058
\end{aligned} \tag{4.11}$$

where

$$\gamma = -\int_{0}^{\infty} \log(z)e^{-z}dz = 0.577215... \tag{4.12}$$

is the Euler constant. The derived result relates $\log(\mathcal{L}_{mlg})$, or $E[Y]$, to the estimated Rayleigh mode $\mathcal{L}_{\sigma}$:

$$\mathcal{L}_{\sigma} = \frac{\mathcal{L}_{mlg}}{e^{0.058}} = 0.9437\mathcal{L}_{mlg} \tag{4.13}$$

This cepstrally smoothed curve *interpolates* the σ values across the analysis frequency range. Notice that if the estimated Rayleigh mode is calculated, without the correction of

eq.(4.13), by using $\mathcal{L}_{mlg}$ of eq.(4.10), there will be a systematic error of underestimation ($\mathcal{L}'_\sigma = \sqrt{2/\pi}\mathcal{L}_{mlg} \approx 0.7979\mathcal{L}_{mlg}$). The corrected **mean noise level** (on the linear scale) shall be

$$\mathcal{L}_m = \sqrt{\pi/2}\mathcal{L}_\sigma \tag{4.14}$$

3. For each subband, check if the condition for the distribution fit is achieved (see eq.(4.8)). If the condition is not achieved, the largest outlier is excluded. That is, the largest outlier is re-classified as sinusoid. Otherwise, the iterative process is terminated.

When all the subbands meet the requirement of the distribution fit, the estimated Rayleigh mode $\mathcal{L}_\sigma$ can be used to derive a probabilistic classification of all the spectral peaks into noise and sinusoidal peaks by means of the pth percentile of Rayleigh distribution

$$\mathcal{L}_n = \mathcal{L}_\sigma\sqrt{-2\log(1-p)} \tag{4.15}$$

with a user selected value for p. In Table 4.1, the four noise levels involved in the derivation of the algorithm are summarized. Notice that if the underlying noise level varies a lot with frequency so that the proposed model cannot capture the noise level evolution, the procedure may not converge, or may not converge to a reasonable estimate.

4.5 Testing and Evaluation

Demonstration

To demonstrate the proposed algorithm, two examples are tested: a white noise signal and a polyphonic signal. In both cases, the sampling frequency is $16k$Hz, the cepstral order coefficient is set $C = 1$, and the percentage of noise to be included is $p = 0.8$ (see eq.(4.15)), that is, 20% of the noise peaks are allowed to be misclassified according to the Rayleigh distribution.

In Figure 4.6, a white noise spectrum is shown with the estimated noise level. The estimated mean noise level $\mathcal{L}_m$ does approximate the constant white noise mean. The estimated noise envelope $\mathcal{L}_n$ is noted as the **noise threshold** to notify that this is a user-adjustable level. The noise peaks are identified as the spectral peaks with magnitudes below this threshold. To further demonstrate how the proposed algorithm adapts to the noise level in frequency, a polyphonic signal is tested (see Figure 4.7). The estimated noise envelope seems to follow well the variation of the noise floor.

Evaluation of noise level estimation

To evaluate the proposed algorithm, three kinds of signals are tested for the estimation of mean noise level : (1) white noise, (2) single sinusoid embedded in white noise, and (3) twenty sinusoids embedded in white noise. The analysis window is the Blackman window of length $L = 93$ms.

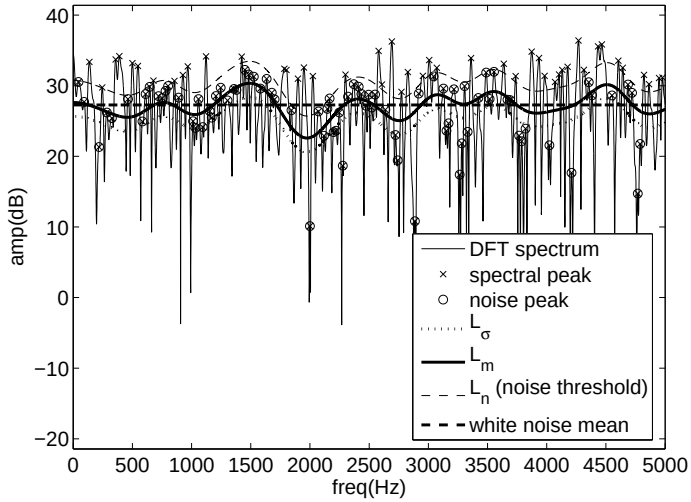


Figure 4.6: Noise level estimation for a white noise signal.

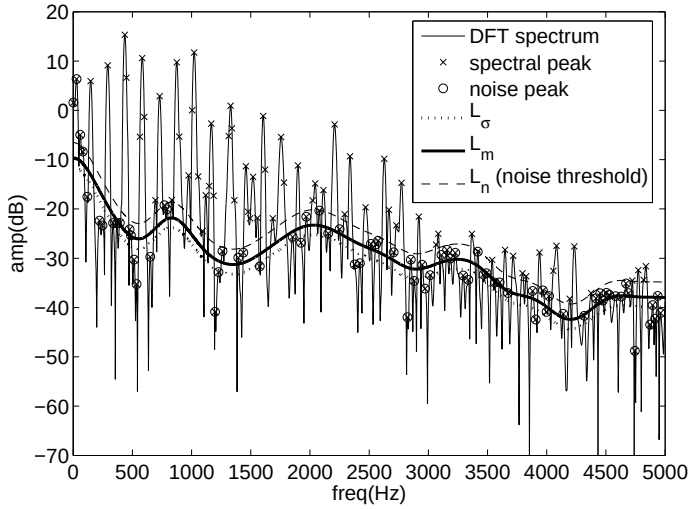


Figure 4.7: Noise level estimation for a polyphonic signal.

The sinusoids are created using exponentially damped sinusoids as follows

$$s[n] = Ae^{-\alpha n}e^{j(\beta n^2+\omega n+\phi)} \tag{4.16}$$

where A is instantaneous amplitude at the reference time index (usually in the center of the window), α is the AM (amplitude modulation) rate, β is half the FM (frequency modulation) rate (frequency slope), ω is instantaneous frequency at the reference time index, and ϕ is the initial phase. The parameters for synthesizing sinusoids are randomly selected from uniformly distributed variables of which the parameter ranges are specified in Table 4.2. Notice that the sinusoids are allowed to overlap in the last test such that the synthesized signals resemble the real-world polyphonic signals. The evaluation metrics uses the bias/variance analysis (Keijzer

and Babovic, 2000):

$$\begin{aligned}\mathrm{MSE}(\hat{L}) &= \frac{1}{N_f N_b}\sum_{i=1}^{N_f}\sum_{j=1}^{N_b}(\hat{L}_{ij} - L_i^w)^2 \\ &= \frac{1}{N_f}\sum_{i=1}^{N_f}(\bar{L}_i - L_i^w)^2 + \frac{1}{N_f N_b}\sum_{i=1}^{N_f}\sum_{j=1}^{N_b}(\hat{L}_{ij} - \bar{L}_i)^2 \\ &= \mathrm{Bias}^2 + \mathrm{Variance}\end{aligned} \tag{4.17}$$

where $\hat{L}$ is the estimated mean noise level, L^w is the white noise level, N_b is the number of frequency bins, N_f is the number of analysis frames, and $\bar{L}_i$ is the mean of $\hat{L}_{ij}$ within an analysis frame.

parameter	A	α	ϕ	$\omega/2\pi$	$\beta/2\pi$
range	$[0.1 \quad 0.5]$	$[0 \quad 0.3]/L$	$[-\pi \quad \pi]$	$[0.01 \quad 0.3]$	$[-0.5/L^2 \quad 0.5/L^2]$

Table 4.2: The parameter distribution range for randomly synthesizing sinusoids.

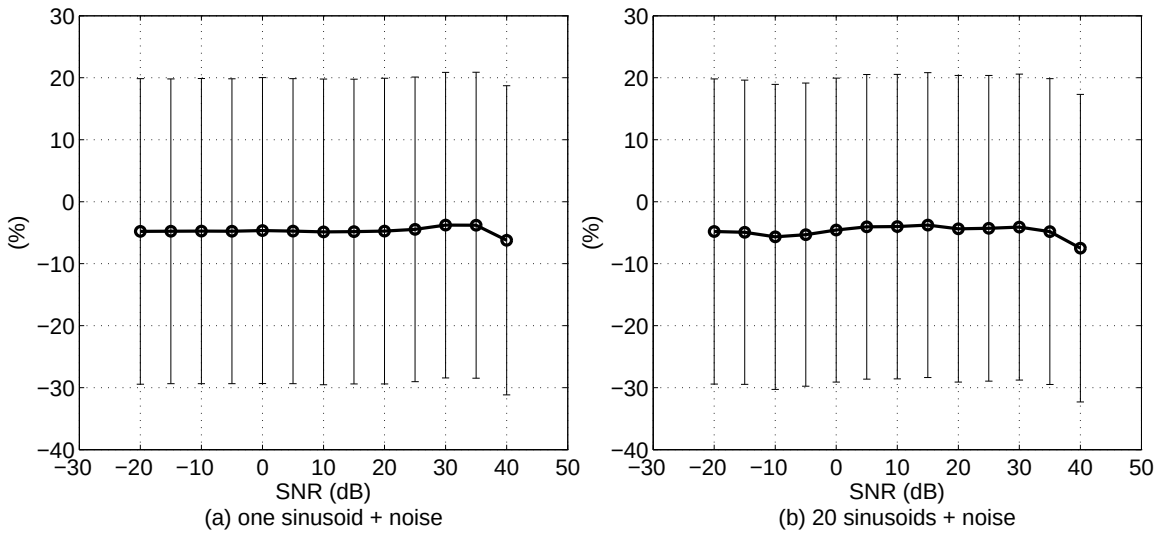


Figure 4.8: The bias and variance of the estimated noise level of synthetic signals.

For the white noise test, the bias is −4.88% and the standard deviation is 24.58%. The systematic bias is possibly due to the subtraction of sinusoids which removes a certain amount of noise energy from the observed signal. The large estimation variance could be due to the cepstral order which is too high for the white noise spectrum that is rather flat. The testing results using sinusoids embedded in white noise are shown in Figure 4.8 for the SNR (Signal-to-Noise Ratio) from −20dB to 40dB. For the case in which only one sinusoid is embedded in white noise, the proposed mean noise level estimator gives consistent performance compared to the white noise test. When there are more sinusoids embedded, the bias seems to vary with different SNRs but it is confined within 1%.

On selecting the noise threshold

The noise threshold is controlled by the user-adjustable parameter p (see eq.(4.15)). It can be formulated as follows, in terms of the difference in dB of the two noise level curves $\mathcal{L}_n$ and $\mathcal{L}_m$.

$$\Delta_{\mathrm{dB}} = 20\log_{10}(\mathcal{L}_n) - 20\log_{10}(\mathcal{L}_m) = 20\log_{10}(\sqrt{-4\log(1-p)}) \tag{4.18}$$

from which

$$p = 1 - e^{(-\pi/4)\cdot 10^{(\Delta_{dB}/10)}} \tag{4.19}$$

In this way, the setting of p can be interpreted as the raise of the noise level from $\mathcal{L}_m$ by Δ_{dB}. Notice that the proposed algorithm simplifies the magnitude distribution of spectral bins by that of spectral peaks. In consequence, it is more intuitive to consider the inclusion of the percentage of noise peaks, denoted by p_n, while selecting a proper p. To investigate the relations between p_n and the percentage of noise bins, the white noise signal is again used, assuming all the spectral bins and all the spectral peaks are noise. The related parameters are summarized in Table 4.3 and their relations are shown in Figure 4.9.

p	percentile of Rayleigh distribution
p_n	percentage of noise peaks
p_b	percentage of noise bins

Table 4.3: Summary of the measures of the percentages of noise components.

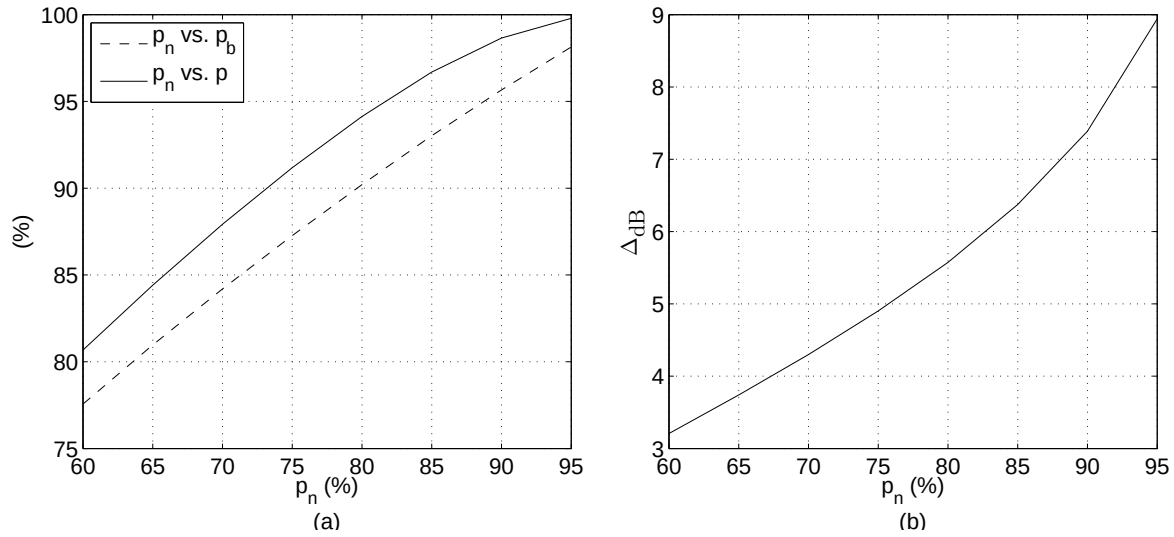


Figure 4.9: The noise thresholds related to the percentages of noise components: (a) the percentage of noise peaks vs. the percentage of noise bins and p; (b) the percentage of noise peaks vs. Δ_{dB}

The dash line in Figure 4.9 (a) shows the relation between the percentage of the noise bins ,

denoted by p_b, and the designated percentage of noise peaks. It is found that p_b is larger than p_n in average. The corresponding p is plotted as the solid line in Figure 4.9 (a), which is larger than p_b by about 3.3% in average. The equivalent raise in dB of the noise level is shown in Figure 4.9 (b). Figure 4.9 provides a look-up table just as the reference which gives a general idea for the selection of p. For example, when one wants to include 90% of the noise peaks, the noise level is raised above more than 95% of the noise bins, which corresponds to about 7.4dB.

5

JOINT EVALUATION OF MULTIPLE F0 HYPOTHESES

Noise level estimation provides a probabilistic classification of the spectral peaks into sinusoids and noise. According to the generative signal model, sinusoidal peaks are considered the partials of the quasi-periodic sources. To estimate the F0s of the quasi-periodic sources, it is proposed to jointly evaluate a set of F0 hypotheses. Each F0 hypothesis is related to its HPS (**Hypothetical Partial Sequence**), and the correct combination of HPS shall have their partials match as many sinusoidal peaks as possible in the observed spectrum. The source model and the source interaction model are handled in the processes of constructing HPS: harmonic matching and overlap treatment. The specially designed treatment of overlapping partials intends to remove the ambiguity based on a hypothetical combination. It is proposed to evaluate the plausibility of hypothetical sources based on three guiding principles: (1) *spectral match with low inharmonicity*; (2) *spectral smoothness*; and (3) *synchronous amplitude evolution within a single source*.

In this chapter, the development of the joint estimation algorithm is focused on the case in which the number of sources is known. The objective is to develop a score function based on the three guiding principles to evaluate a set of F0 hypotheses. The joint estimation algorithm plays an important role in the scoring of all possible combinations based on which the polyphony can be inferred in a later stage. This chapter begins with the description of the construction of the HPS for a combination of F0 hypotheses. Then, the score function is presented to evaluate the plausibility of a combination of hypothetical sources. Finally, the proposed algorithm is evaluated by mixtures of harmonic instrument sounds.

5.1 Generating Hypothetical Sources

The source model of an F0 hypothesis is a set of harmonic grid without specific amplitudes. Given a set of F0 hypotheses, the frequencies and the amplitudes of their HPS are estimated by two processes: (1) *partial selection* and (2) *overlapping partial treatment*. The partials of a hypothetical source can be determined by matching the harmonics of the related quasi-harmonic model with the observed peaks. To remove the ambiguity in the overlapping partials of HPS, it is proposed to re-estimate, for each hypothetical source that overlaps, the partial amplitudes based on the interpolation of non-overlapping partials.

5.1.1 Harmonic matching for partial selection

For each F0 hypothesis, the degree of match in frequency is evaluated between the model harmonics and the observed peaks. A tolerance interval is designated in the neighborhood of each model harmonic, which allows to handle the inharmonic partials. The spectral peaks situated in the tolerance interval are considered *matched* peaks, otherwise *unmatched* ones. The **degree of deviation** of the ith observed peak from the hth harmonic is expressed as

$$d_m(i) = \begin{cases} \frac{|f_i - f_{m,h}|}{\alpha f_{m,h}} & \text{if } |f_i - f_{m,h}| < \alpha f_{m,h}, \\ 1 & \text{otherwise.} \end{cases} \tag{5.1}$$

where f_i is the frequency of the ith observed peak and $f_{m,h}$ is the frequency of the hth harmonic of the source model. The use of index m is to refer to the mth hypothetical source in a combination. α determines the tolerance interval $\alpha \cdot f_{m,h}$. When an observed peak is situated outside the related tolerance interval, it is considered unmatched and $d_m(i)$ is set to 1. Therefore, $0 \leq d_m(i) \leq 1$. In order to adaptively search the peaks matching to the model harmonics, the harmonic frequency of the model is updated by $f_{m,h+1} = f_i + f_m$ (when a matched peak is found for the hth harmonic) where f_m denotes the F0 value. If the hth harmonic does not match any observed peaks, the next harmonic frequency is updated by $f_{m,h+1} = f_{m,h} + f_m$. The first harmonic is initiated by a hypothetical F0, which implicitly constrains the harmonic locations around the multiples of the hypothetical F0. In this way, the measure of harmonic matching is based on both the **spectral location** principle and the **spectral interval** principle (Klapuri, 2004).

In the case of single-F0 estimation, the tolerance interval can simply be set as large as $0.5 \cdot f_m$ to allow inharmonic partials, whereas disallowing the overlaps of the tolerance intervals of adjacent harmonics. In the case of multiple-F0 estimation, however, α shall be determined in a more precise way because partials of concurrent sources may fall into the same tolerance interval. A constraint can be posed on the frequency difference of two adjacent partials (see Figure 5.1). The maximum and the minimum of the frequency difference between two adjacent partials are

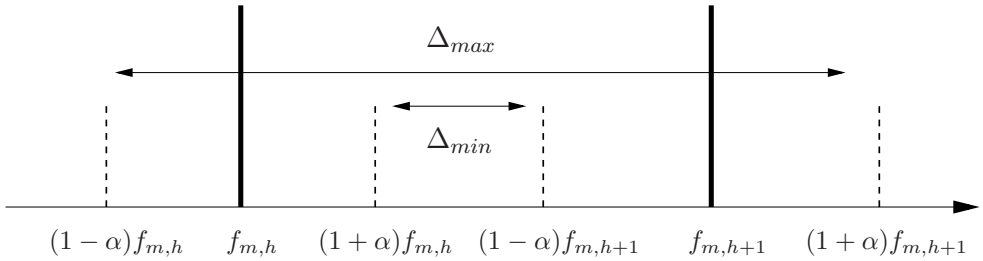


Figure 5.1: The maximum and the minimum of the frequency difference of two adjacent partials. The tolerance interval of each partial is the region between two dash lines around the partial.

$$\Delta_{max} = (1+\alpha)f_{m,h+1} - (1-\alpha)f_{m,h} \approx f_m + (2h+1)\alpha f_m \tag{5.2}$$

$$\Delta_{min} = (1-\alpha)f_{m,h+1} - (1+\alpha)f_{m,h} \approx f_m - (2h+1)\alpha f_m \tag{5.3}$$

in which the approximations $f_{m,h+1} - f_{m,h} \approx f_m$ and $f_{m,h+1} + f_{m,h} \approx (2h+1)f_m$ are used. The allowed frequency range for a peak to match a harmonic is thus $(2h+1)\alpha f_m$, which is set to be limited by βf_m where β is defined as a factor of the fundamental frequency f_m. Then, α can be selected according to

$$\alpha \leq \frac{\beta}{2h+1} \tag{5.4}$$

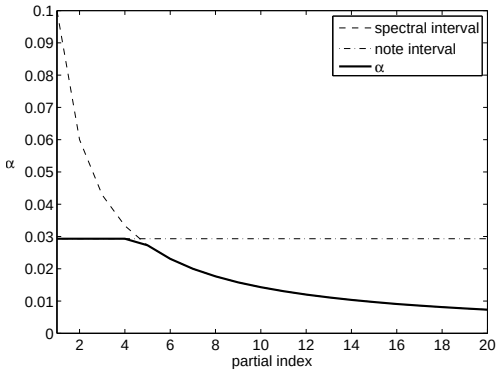


Figure 5.2: α as a function of partial index.

and β is a set to 0.3 (see the dash line in Figure 5.2). For the tolerance intervals of lower partials, a constraint is further set, empirically, according to a quarter-tone frequency resolution. That is, $\alpha = 2^{1/24} - 1 = 0.029$. With the two constraints on the tolerance interval, α can thus be determined (see the solid line in Figure 5.2).

For polyphonic signals, there may be more than one peak falling into the same tolerance interval. Some proposed to select the *nearest* peak around the related harmonic (Parsons, 1976; Duifhuis and Willems, 1982). Others proposed to select the *maximal* peak (Klapuri, 2006). All these techniques take into account only one peak in the tolerance interval and neglect the others. However, both techniques are not robust when the matched peaks include spurious peaks or the partials of concurrent sources. Selecting the nearest peak may select a spurious peak; whereas selecting the maximal peak may select the partial of other sources. The proposed partial selection technique begins with assigning the first partial to the nearest peak. For the consecutive partials, two *peak candidates* are considered: (1) the nearest one and (2) the one of which the mainlobe cover the related model harmonic. By means of comparing the average amplitude of the previously selected three partials with the amplitudes of the two peak candidates, the peak candidate forming a smoother envelope is allocated to the hypothetical source.

5.1.2 Overlapping partial treatment

Given a combination of hypothetical sources, the overlapping partial positions can be easily inferred. This information can then be used to estimate the partial amplitudes of each hypothetical source. The goal is to make the best of the unambiguous information derived from the non-overlapping partials to remove the ambiguity in overlapping partials. It is assumed that an overlapping partial still carries important information about at least the HPS that locally has the strongest energy. Therefore, the overlapping partial treatment aims to allocate the overlapping partial amplitude to this HPS. Based on the spectral smoothness principle, the strategies to estimate the amplitudes in the overlap positions are listed below:

- Partials having potential collisions are determined by the peaks that match to more than one hypothetical sources. The overlap treatment is carried out in order of the partial frequency.
- In each overlap position, the *local energy* of each HPS is estimated by the interpolation of the amplitudes of the neighboring partials that do not overlap (Maher, 1990). The amplitude of the overlapping partial is exclusively assigned to the HPS with the largest local energy. The overlapping partial of that HPS is labeled as *credible* and is used like a non-overlapping partial for the consecutive interpolation. For the rest of the colliding sources, their amplitudes in the overlap position are replaced by the interpolated amplitudes respectively. The use of the interpolated amplitudes is meant to maintain the local smoothness of the envelope for the partial amplitudes that can not be easily inferred.
- When one of the neighboring partials is overlapped, the amplitude of the non-overlapping partial determines the local energy. If both the neighboring partials are overlapped, the partial of the related source is considered *not credible*. In this case, the amplitude of the overlapping partial is replaced by the interpolated amplitude if the observed amplitude is larger than the interpolation.
- When the amplitude of the overlapping partial is smaller than all the interpolated amplitudes of the colliding sources, it is difficult to infer which hypothetical source contributes the most. In this case, the colliding sources *share* the overlapping partial. The overlapping partial in all HPS is labeled as credible for the consecutive interpolation. The idea is to keep as many as possible the credible partials for the treatment of consecutive overlapping partials.

An example of the overlapping partial treatment is demonstrated in Figure 5.3. The envelopes of the HPS after this treatment are smoother. Following the bandwise smooth model (see Section 2.1.2), the proposed treatment uses the interpolated amplitude as an estimator of overlapping partials. To examine the estimator based on the bandwise smooth model, its accuracy is evaluated by harmonic instrument sound samples. The estimation error [1] is calculated between an observed partial amplitude and the interpolated amplitudes of the neighboring partials. This test consider

[1] HPS are compressed exponentially by a factor of 0.5 beforehand.

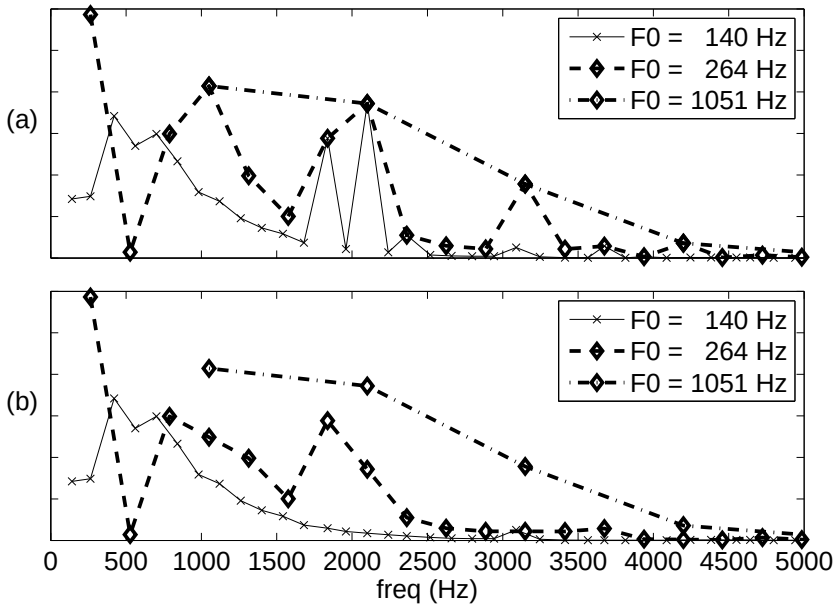


Figure 5.3: An example of the overlapping partial treatment: (a) HPS constructed by partial selection; (b) HPS after the treatment of the overlapping partials.

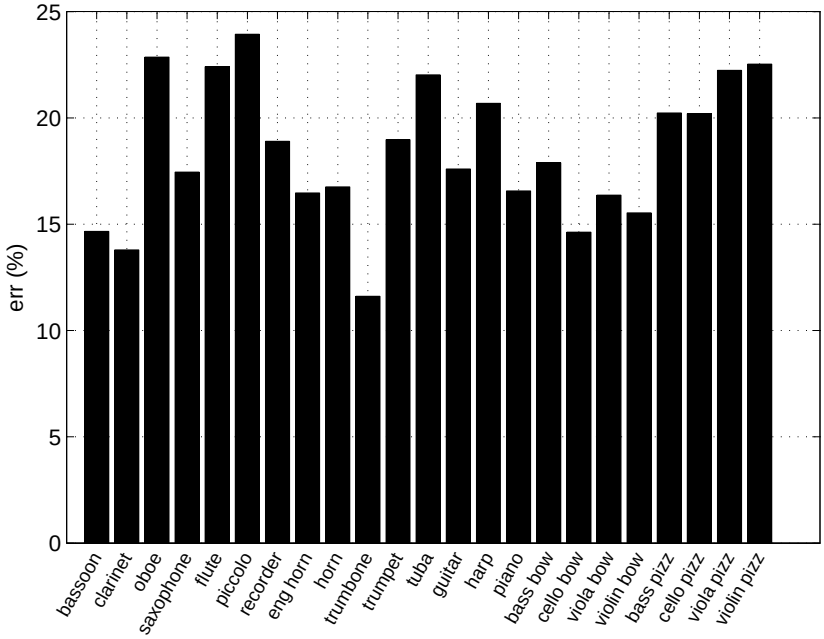


Figure 5.4: The estimation errors of partial amplitudes using the interpolated amplitudes.

only the partials that are larger than the interpolated amplitudes. Twenty-three instruments are selected for this test (see Figure 5.4). The partials of the flute sounds and the piccolo sounds decay rapidly along the frequency. Therefore, the interpolation causes dominant errors when a relatively strong partial is used with a relatively weak partial, resulting in an over-estimation of the partial amplitude. Similarly, the strong resonances of, for instance, the oboe and the plucked string instruments, which boost certain partials immensely can cause a significant estimation error, too.

5.1.3 Spectral flattening

To further attenuate the dynamics of the spectral envelopes, HPS are compressed exponentially such that an "equalized" spectral envelope is attained. This technique, called **spectral flattening**, makes HPS more appropriate to be evaluated with regard to the spectral smooth principle. The usual way to spectral flattening is to take the logarithm of the spectrum or compress the spectrum exponentially (Sreenivas and Rao, 1981). It is observed that taking the logarithm of the spectrum eliminates the characteristics of the spectral envelopes and boosts the noise at the same time. However, the "unsmoothness" of spectral envelopes is useful for the extraction of HRF0s (see Section 6.2). Hence, the more flexible way is to apply an exponential compression such that the compromise between the attenuation of the dynamics and the preservation of the characteristics can be controlled. Karjalainen and Tolonen (2000) investigated the exponential compression of the spectrum for SACF, and a factor of 0.67 was found to be a good choice. This compression factor, in fact, corresponds to the *power law* relation between loudness level LL and sound intensity I (Stevens, 1970): $LL = C_1 \sqrt[3]{I} = C_2 \sqrt[3]{(\Delta p)^2}$ where Δp is the sound pressure variation. C_1 and C_2 are frequency dependent constants. Empirically, it is observed that the exponential compression of 0.5, i. e., the square root, has a similar compression effect to that of 0.67. In addition, the square root is of better computational efficiency. Therefore, the exponential compression of 0.5 is generally applied to the HPS for evaluating the smoothness of the spectral envelopes.

5.2 Score Function

After the construction of the most reasonable HPS for a set of F0 hypotheses, a score function is designed to evaluate the plausibility of the combination of the hypothetical sources. Based on the three guiding principles, four score criteria are proposed: harmonicity (HAR), mean bandwidth (MBW), spectral centroid (SPC), and the standard deviation of mean time (SYNC). The score function is formulated as the linear combination of the four criteria.

5.2.1 Harmonicity

The score criterion HAR evaluates the harmonic matching between the combination of the hypothetical sources and the observed spectral peaks. It is based on the harmonicity principle which indicates the harmonicity and the explained energy of a hypothetical source. To derive

the combinatorial property of harmonicity for M hypothetical sources, their individual deviation vectors are first combined as follows:

$$D_M(i) = \min\left(\{d_m(i)\}_{m=1}^{M}\right) , \quad \forall i \in I \tag{5.5}$$

That is, each observed peak is matched with the closest partial in the hypothetical sources such that the resulting combination explains the observed spectrum with the lowest inharmonicity. HAR is thus defined as the weighted sum of $D_M(i)$ for all peaks

$$\text{HAR} = \frac{\sum_{i=1}^{I} Spec(i)^a \cdot D_M(i)}{\sum_{i=1}^{I} Spec(i)^a} \tag{5.6}$$

where the **peak salience** $Spec(i)$ is the *sum of linear amplitudes* for all the bins within the ith spectral peak. The reason of not using the peak energy (the sum of squared amplitudes) is to not emphasize the dynamics of partial amplitudes. In order to equalize the significance of the peaks, an exponential compression $a = 0.5$ is applied to the peak salience. Because the subharmonics may have competitive harmonic matching compared to that of the correct F0s, three score criteria are designed to evaluate the plausibility of each hypothetical source.

5.2.2 Mean bandwidth

To score the spectral smoothness of a hypothetical source, the frequency content of the envelope of a HPS is evaluated using the mean bandwidth as a criterion. Each HPS is first assembled with its flipped sequence to construct a symmetrical sequence g_m. This process is meant to avoid the discontinuity at the first partial, and to obtain a smooth representation of HPS (see Figure 5.5(a)). Applying K-point [1] FFT to g_m, the spectrum G_m of HPS is acquired (see Figure 5.5(b)). The score criteria mean bandwidth is defined as follows

$$\text{MBW}_m = \frac{1}{K/2}\sqrt{2 \cdot \frac{\sum_{k=1}^{K/2} k|G_m(k)|^2}{\sum_{k=1}^{K/2} |G_m(k)|^2}} \tag{5.7}$$

which indicates the degree of energy concentration in the low-frequency region of G_m. In this way, the envelope of g_m with smaller variations results in a smaller value of MBW_m. The function of MBW is to discriminate a correct F0 from its subharmonics. For a subharmonic F0/n, the envelope is "disturbed" at every n partial, which reflects the periodic peak in the spectrum of its HPS. That is, the high frequency components are dominant at multiples of $1/n$ measured on the normalized frequency scale. To further illustrate the function of MBW, a clarinet sound signal is used to demonstrate the difference in the envelope smoothness for the HPS of F0 and the HPS of F0/2 (see Figure 5.5). Although the resonance structure of the clarinet sound does not result in a smooth spectral envelope, the envelope of F0/2 is less smooth than that of F0. Due to the missing even harmonics, the envelopes are disturbed every 4 partials in HPS, which results in the dominant frequency components at multiples of 0.25 (see Figure 5.5(b)). Compared with the

[1] Two times the next power of 2 of the length of g_m.

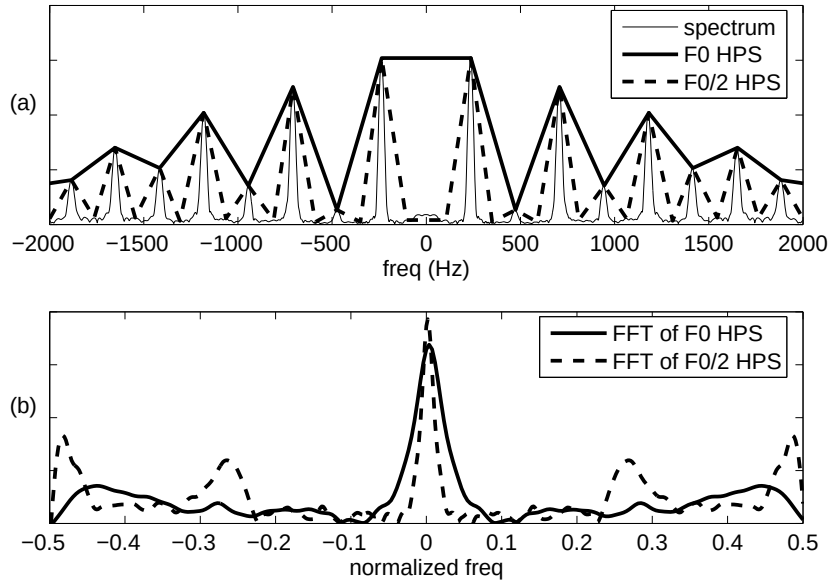


Figure 5.5: Spectral smoothness comparison of a clarinet sound playing the note Bb3.

HPS of F0, the HPS of a subharmonic like F0/2 generally has more high frequency energy. That is, the energy spreads more widely in frequency and MBW is larger.

5.2.3 Spectral centroid

For harmonic instrument sounds, the spectral centroids tend to lie around the lower partials because the higher partials often decay rapidly. According to this general property related to the spectral smoothness principle, the centroid can evaluate the energy spread of a HPS:

$$\mathrm{SPC}_m = \frac{1}{B/2}\sqrt{2 \cdot \frac{\sum_{n=1}^{N_m} n[\mathrm{HPS}_m(n)]^2}{\sum_{n=1}^{N_m}[\mathrm{HPS}_m(n)]^2}} \tag{5.8}$$

where N_m is the length of HPS_m. B is a normalization factor determined by $\lfloor F_{90}/F0_{min} \rfloor$. F_{90}, called the **spectral roll-off** (Peeters, 2003), which stands for the frequency limit containing 90% of spectral energy in the analysis frequency range, and $F0_{\min}$ is the minimal F0 hypothesis in search. Since the spectral envelopes of harmonic instrument sounds are not always smooth, SPC works as a further test, in addition to MBW, of the related physical property. This criterion works as a penalty function for the subharmonics of which the HPS matches the partials of concurrent sources.

5.2.4 Synchronicity

To evaluate the synchronicity of the temporal evolution of the partials in a HPS, **mean time** is estimated for individual spectral peaks. Mean time is an indication of the center of gravity of the signal energy (Cohen, 1995). It can be defined in the frequency domain as the weighted sum of group delays (see eq.(B.6)). The *mean time of a spectral peak* can be estimated by considering

only the frequency bins within a spectral peak, which can characterize the amplitude evolution of the related source (Röbel, 2003a). For a coherent HPS, the synchronous evolution of partials is expected. This results in a small variance of mean time w.r.t. the matched peaks. The *mean time of a hypothetical source*, denoted as T_m, is calculated as the power spectrum weighted sum of the mean time of the hypothetical partials. The standard deviation of the mean time of the partials is then formulated as

$$\mathrm{SYNC}_m = \frac{1}{L/2}\sqrt{\sum_{i\in\mathrm{HPS}_m}\{(\bar{t}_i - T_m)^2 \cdot w_m(i)\}} \tag{5.9}$$

where L is the window size, $\bar{t}_i$ denotes the mean time of the ith observed peak. The weighting vector w_m, normalized to sum to one, is constructed from HPS by disregarding (set to zero) the overlapping partials of which the spectral phases are possibly disturbed. Since this criterion makes use of the noise, an exponential compression factor of 0.23 is applied to w_m in order to raise the significance of the noise components (see the **specific loudness** descriptor in (Peeters, 2003)). w_m avoids the use of the disturbed phases of overlapping partials and makes use of the spurious peaks to penalize a HPS containing more noise peaks.

Notice that the three criteria MBW_m, SPC_m and SYNC_m are evaluated *individually* for each hypothetical source. To combine the individual criteria into combinatorial ones, they are weighted by the **effective salience** of the respective hypothetical sources. The effective salience is the sum of the peak salience of the "credible" partials within a HPS. The term "effective" is used because the ambiguous partials have been treated, and what remain in the HPS are representative of the hypothetical source. The weighted sum of individual criteria is normalized by the sum of effective salience, giving rise to the resulting score criteria with values between 0 and 1. The score function is formulated as a linear combination of the four criteria:

$$\mathcal{S} = p_1 \cdot \mathrm{HAR} + p_2 \cdot \mathrm{MBW} + p_3 \cdot \mathrm{SPC} + p_4 \cdot \mathrm{SYNC} \tag{5.10}$$

where $\{p_j\}_{j=1}^4$ are the weighting parameters. The four criteria are designed in a way that a smaller weighted sum stands for a better score. Notice that HAR favors lower hypothetical F0s whereas MBW, SPC and SYNC favor higher ones. Therefore, the criteria perform in a complementary way. The weighting parameters shall be optimized to balance the relative contribution of each criterion such that the score function generally ranks the correct F0s the best. To refine the F0 values, the linear regression is applied to the effective hypothetical partials for a given combination of F0 hypotheses.

5.3 Score Criteria for Musical Instrument Sounds

To demonstrate the functions of the score criteria, the score criteria for single-note sound signals are illustrated (see Figure 5.6). The note samples are collected from four databases: RWC Musical Instrument Sound database, McGill University Master Samples, IRCAM Studio On Line

database and Iowa Musical Instrument Samples database. There are 23 categories of musical instruments. The four criteria are signal descriptors when we refer to monophonic signals. The characteristics of musical instrument sounds are reflected in the score criteria. For example, the resonances at odd harmonics of the clarinet sound give rise to a rather unsmooth spectral envelope and thus the MBW is larger. The resonance maximum of the bassoon sound is often higher than the first partial, which results in a wider energy spread and thus the SPC is larger. Pizzicato sounds are comparatively noisy and their HPS tend to match more noise peaks, which results in larger SYNC.

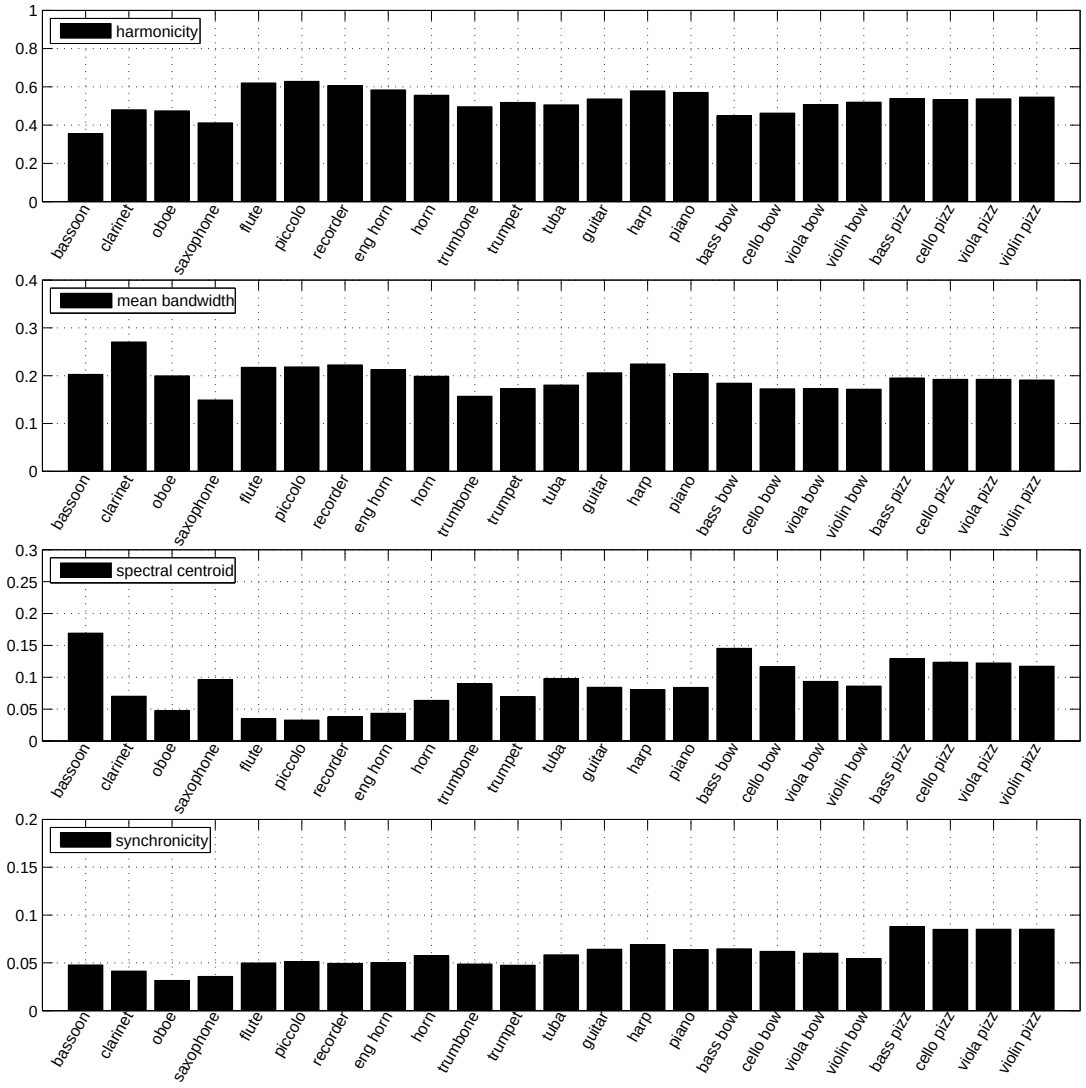


Figure 5.6: Score criteria for a variety of musical instruments

The comparisons of the score criteria among F0, F0/2 (subharmonic F0) and 2F0 (superharmonic F0) are also demonstrated (from Figure 5.7 to Figure 5.10). To facilitate the comparison, the score criterion values related to F0/2 and 2F0 are shown at the same note index in the respec-

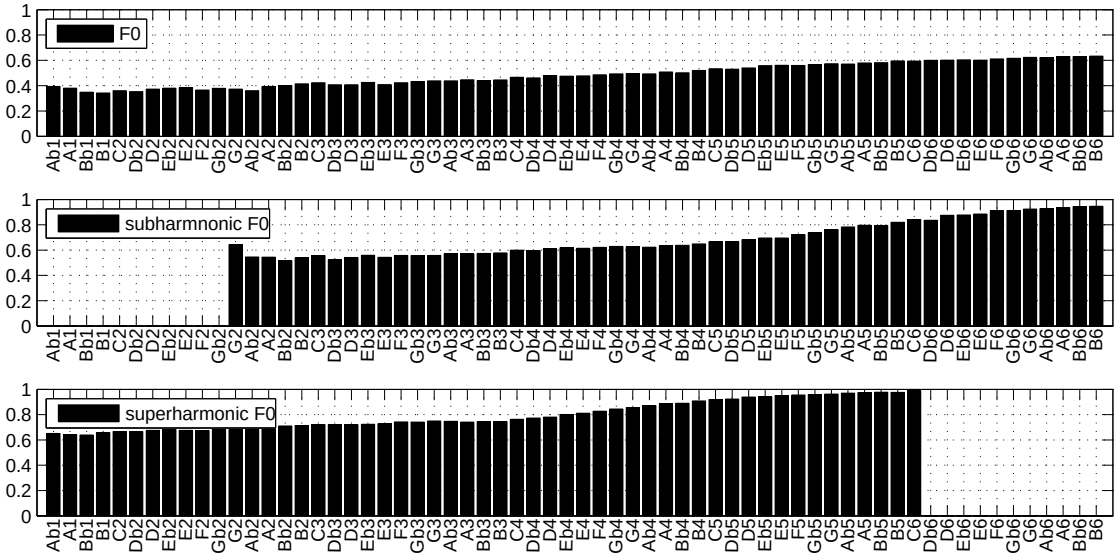


Figure 5.7: HAR comparison

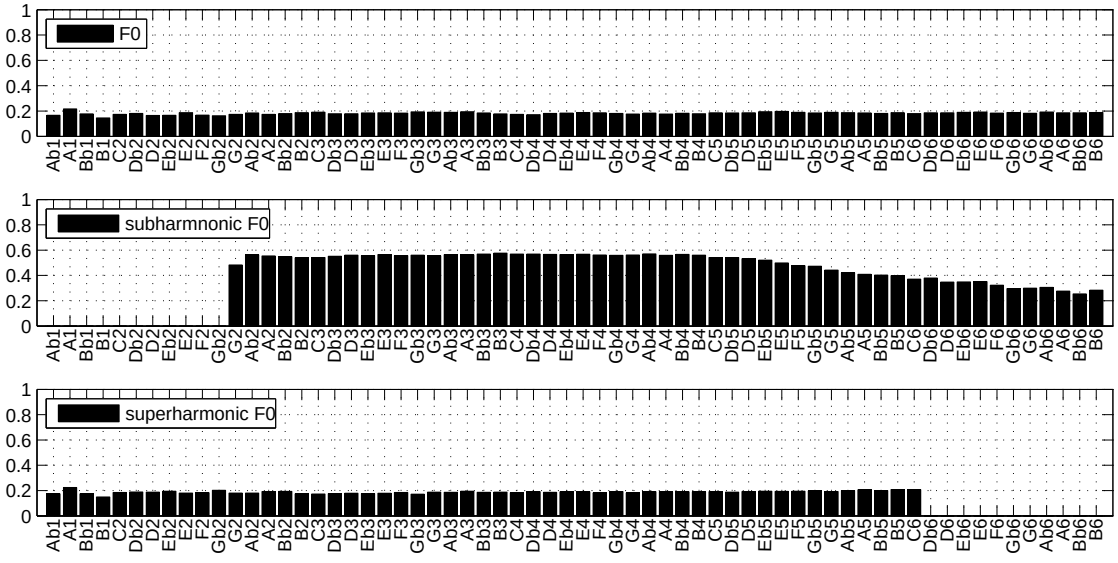


Figure 5.8: MBW comparison

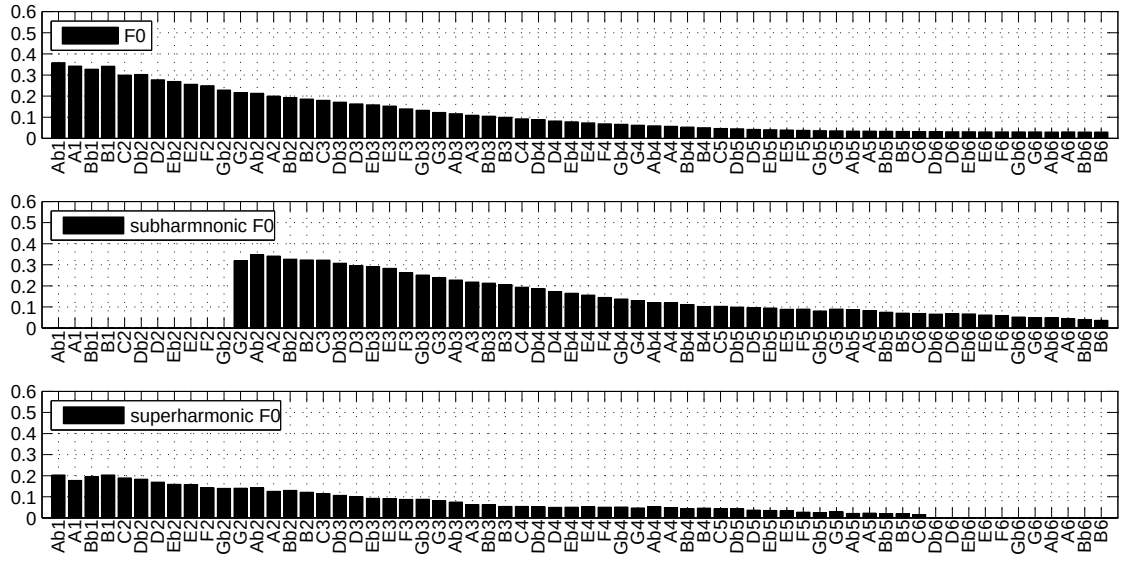


Figure 5.9: SPC comparison

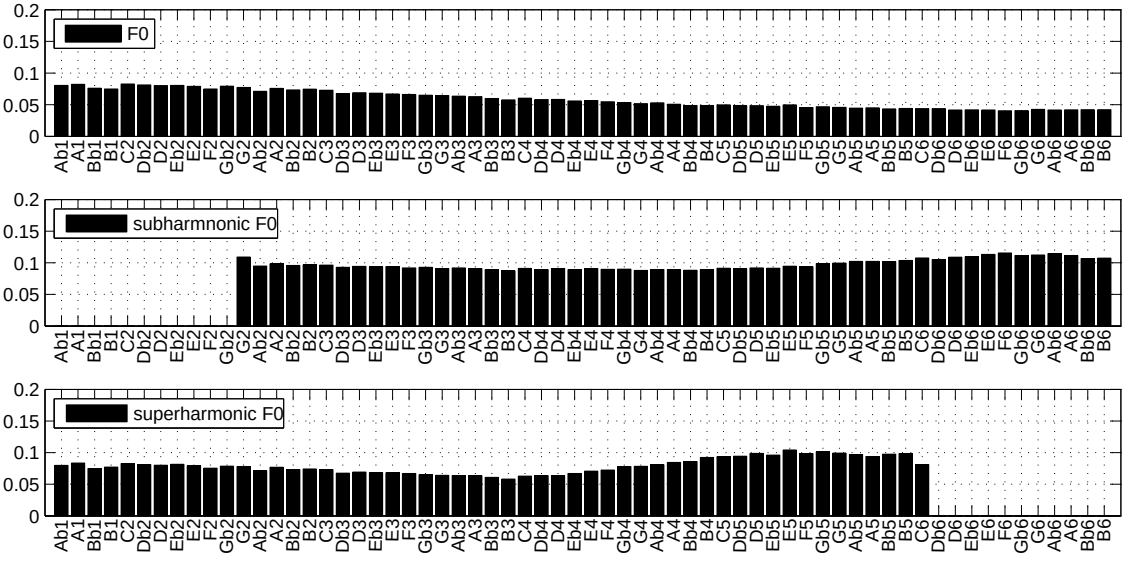


Figure 5.10: SYNC comparison

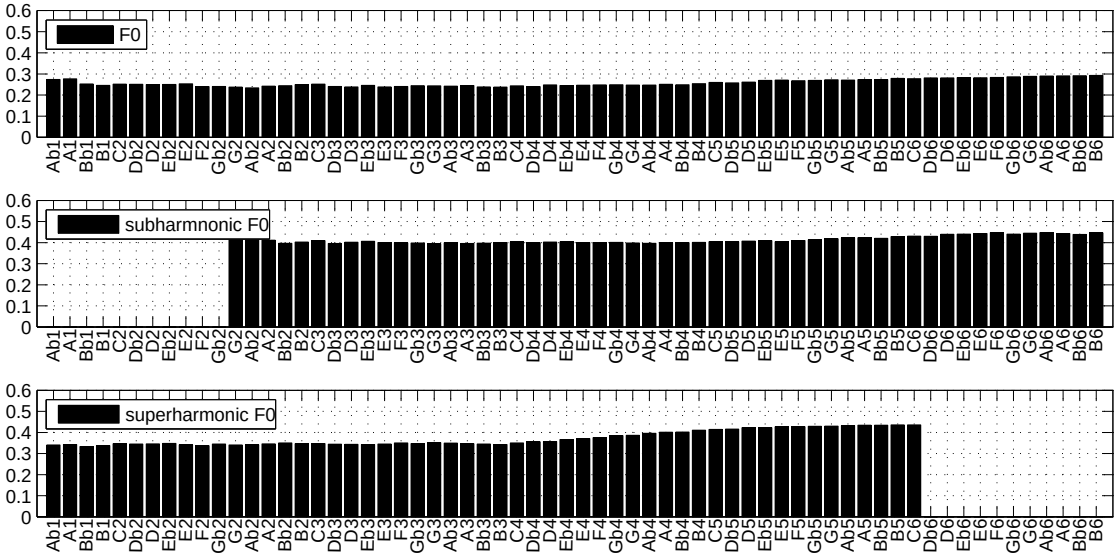


Figure 5.11: SCORE comparison

tive sub-figures. As expected, F0 is better than 2F0 for HAR because there are more harmonics in the F0 model than those in the 2F0 model (Figure 5.7). Surprisingly, F0 is still better than F0/2. This is probably because the partial selection prefers smooth envelopes such that more noise peaks are selected for the F0/2 model. Since the noise peaks have been discarded for HAR, F0/2 is worse than F0 in general. In addition, for inharmonic sounds F0/2 might have a worse harmonic matching compared to F0. F0 and 2F0 do not differ much for MBW, which conforms to the spectral smoothness principle (see Figure 5.8). On the other hand, F0/2 is largely disfavored by about 0.4 compared to the MBW of F0 and 2F0. Because SPC is also designed to disfavor lower F0s, the results are similar to that of MBW (see Figure 5.9). SYNC makes use of noise peaks to disfavor subharmonics and F0/2 is disfavored by about 0.05 in general (see Figure 5.10). For the higher notes, the increase in SYNC of 2F0 may be due to the less-stationary nature of high frequency partials. Finally, the resulting scores are also compared (see Figure 5.11). The weighting parameters are trained by a database containing not only monophonic samples but also polyphonic samples (see Section 5.4). The resulting score has rather equal quality for the range of notes considered (compare the plot for F0 from Figure 5.7 to Figure 5.11).

5.4 Evaluation

Following the evaluation method of Klapuri (2003), musical instrument sound samples are semi-randomly mixed to create the database (see Section 7.2.2) for the evaluation of the presented joint estimation algorithm. Non-transient parts of monophonic samples are pre-selected and then mixed with equal mean-squared energy. Multiple-F0 estimation of a polyphonic sample is carried out within a single frame. The number of sources, or the polyphony, is given for the score

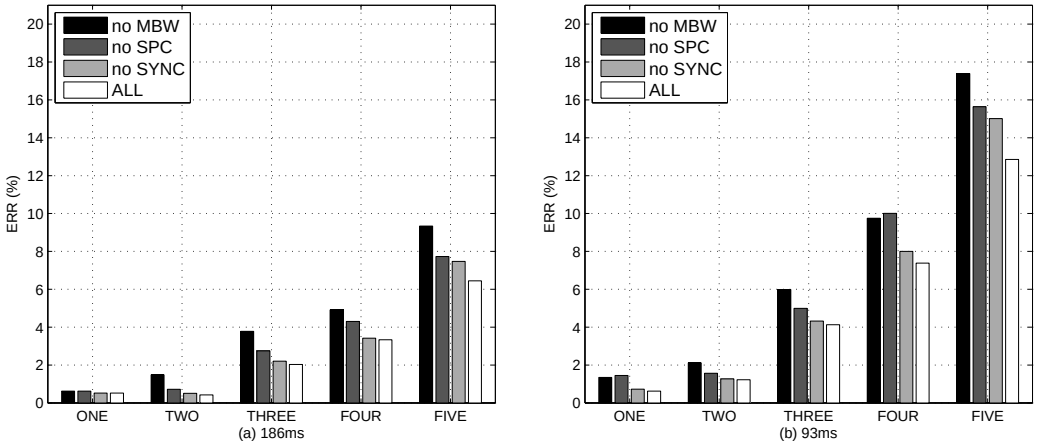


Figure 5.12: Comparison of the functioning of the score criteria: deactivation of MBW (no MBW), deactivation of SPC (no SPC), deactivation of SYNC (no SYNC) and activation of all the criteria (ALL).

function to find the most probable set of F0 hypotheses. The F0 search range is set between 50Hz and 2000Hz. The maximal analysis frequency is set at 8000Hz. Blackman window is chosen as the analysis window. A correct estimate should not deviate from the reference value by more than 3%. The error rates are computed as the number of wrong estimates divided by the total number of target F0s. To train the weighting parameters $\{p_j\}_{j=1}^{4}$, 100 polyphonic samples for polyphony from one to five are created as the training database. The parameters are trained by the evolutionary algorithm (Schwefel, 1995) and the set resulting in the best performance [1] is selected for the score function. For the evaluation, two analysis window sizes, 186ms and 93ms, are tested for the polyphony from one to five (see Table 5.1). This experiment follows the setups of Klapuri (2003) but different in the semi-random way of mixing monophonic sound samples. The results demonstrate the competitive performance of the proposed algorithm compared to several algorithms proposed by Klapuri (2006). To study the significance of MBW, SPC and SYNC, a further test is carried out in which one of the three criteria is deactivated. This test shall give an idea about how effectively each criterion disfavors wrong hypothetical sources according to its guiding principle. HAR is activated for all the tests because it is essential for F0 estimation. The comparison with the original result is shown in Figure 5.12. It is observed that the deactivation of any of the three criteria degrades the overall performance.

window size	ONE	TWO	THREE	FOUR	FIVE
186ms	0.05%	0.42%	2.03%	3.33%	6.45%
93ms	0.06%	1.23%	4.13%	7.38%	12.86%

Table 5.1: Evaluation of the joint estimation algorithm in the case that the number of F0s is given.

[1] $\{p_j\}_{j=1}^{4} = \{0.3774, 0.2075, 0.2075, 0.2075\}$

6

ITERATIVE EXTRACTION OF F0 CANDIDATES

Given the number of sources, the joint estimation algorithm jointly evaluates all the possible combinations among F0 candidates. The joint estimation approach has the advantage over the iterative approach in the handling of overlapping partials. However, the number of combinations grows exponentially with the number of F0 candidates as well as the polyphony. N candidates amount to $\binom{N}{M}$ combinations for the polyphony M. If the F0 candidates are, for instance, sampled on a 1Hz grid between 50Hz and 2000Hz, there will be more than one billion combinations to evaluate for a polyphony of three. Therefore, the selection of F0 candidates plays an important role in a joint estimation approach. A proper candidate selection scheme helps to reduce unnecessary calculations while keeping the robustness of an F0 estimation algorithm.

In this chapter, two approaches to F0 candidate selection are studied. One is based on the use of a threshold for a polyphonic salience function. Two salience functions are proposed: one is based on the harmonicity principle and the other is based on the beating of adjacent partials. A candidate selection method based on the iterative estimation approach is also presented. This method first iteratively estimates NHRF0s (non-harmonically related F0s). HRF0s (harmonically related F0s) are then extracted from the probable harmonic sequences within the HPS of NHRF0s. The three methods are evaluated with respect to the accuracy of the selection of candidates and the accuracy of the estimation of multiple F0s, as well as the number of candidates.

6.1 Polyphonic Salience Functions

The usual approach is to select F0 candidates according to the salience measure of a single-F0 estimation algorithm. The salience measure indicates the degree of periodicity/harmonicity, or other kinds of information. Because the single-F0 estimation algorithm is applied to polyphonic signals, the resulting salience function is called the polyphonic salience function (see 2.2.2). This approach usually depends on a threshold for the salience to select the locally salient F0 candidates. In most cases, the subharmonics and the super-harmonics of the correct F0s have competitive salience to that of the correct F0s. The difficulty in setting the threshold is to reach the compromise between the selection of the correct F0s and the reduction of their subharmonics and super-harmonics (see Figure 1.7). To avoid an excessive selection of the subharmonics, spectral pattern matching (see Section 2.2.3) can be a solution provided that the source models are representative of the spectral envelopes involved. The blackboard system usually performs partial tracking beforehand and selects the continuous trajectories for the related F0 candidates (see Section 2.2.5). In the statistical signal modeling approach, a good selection of F0 candidates as initial conditions can help the parameter adaptation converge to optimum with fewer iterations (Dubois and Davy, 2007). Goto (2000) used the *fixed-point* estimation (Abe *et al.*, 1995) to select dominant and stable partials as F0 candidates. Kameoka *et al.* (2007) picked the 60 largest peaks in the observed spectrogram within 400 consecutive frames. In the following, two polyphonic salience functions are proposed, considering the easy integration of the joint estimation algorithm.

6.1.1 Harmonicity

The score criterion *harmonicity* (see eq.(5.6)) can be used as a harmonicity measure of an F0 hypothesis. This criterion indicates the inharmonicity of an F0 hypothesis; therefore, the local minima represent potential candidates (see Figure 6.1(a)). The harmonicity criterion in general favors low-F0 hypotheses over high-F0 hypotheses because the number of partials of a hypothetical source decreases with the increase of its related F0. The threshold is suggested to be as high as 0.95 to ensure the inclusion of all correct F0s. However, when the number of sources increases, the number of spurious candidates increases immensely. The spurious candidates are mostly the subharmonics of the correct F0s and the subharmonics of the harmonics of the correct F0s.

6.1.2 Partial beating

A signal with more than one sinusoidal component exhibits periodic fluctuations, called **beating**, in the temporal envelope. The rate of beating depends on the frequency difference between every two sinusoidal components (Klapuri, 2004). Given two sinusoids of different frequencies, the magnitude of the beating is determined by the sinusoid of the *smaller* magnitude. This property is useful for distinguishing the HPS of a correct F0 from that of its subharmonics. Because the HPS of the subharmonics generally match more noise peaks than the HPS of the related F0 does (see Figure 5.5), the resulting beating should be smaller. Accordingly, the partial beating

measure is defined as follows

$$\mathrm{PB} = \frac{1}{A_{max}} \sum_{h=1}^{H_b} \min(a(h), a(h+1)) \tag{6.1}$$

where $a(h)$ is the partial amplitude *relative to the noise envelope* of the hth hypothetical partial. A_{max} is the maximal amplitude of the observed peaks, which is used for normalizing PB between 0 and 1. H_b is the number of partials, which is empirically set to be maximally 10 partials. The use of a fixed number of partials is for the purpose of the normalization of PB among all F0s in search. Two adjacent partials are considered a pair to produce the beating. For each pair of the beating, the smaller one determines the effect of the beating for their frequency difference, that is, the F0. The reason to use the partial amplitude relative to the noise envelope is to disregard the beatings related to the noise components. In this way, only when two sinusoidal partials beat each other can the effect of beating be dominant. The partial beating measure is related to the spectral smoothness principle. When the smoothness of a spectral envelope is disturbed by dominant overlapping partials, the beating of a large overlapping partial with its adjacent partials is determined by the smaller partial. That is, the overlapping partials are taken care of implicitly. Although the partial beating measure avoids the selection of a significant number of subharmonic candidates, super-harmonic candidates often have competitively dominant beating effects.

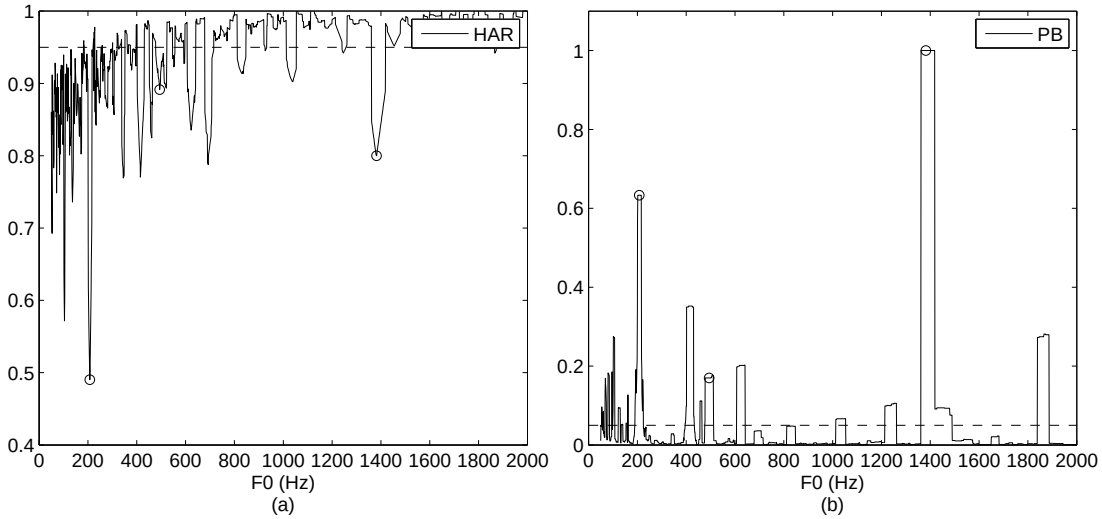


Figure 6.1: Two polyphonic salience function for F0 candidate selection: (a) HAR (harmonicity); and (b) PB (partial beating). The circles indicate the correct F0s at around 208Hz, 494Hz and 1382Hz. The empirical thresholds for both criteria are shown as the respective dash lines at 0.95 and 0.05.

A salience function can be derived from combining both harmonicity and partial beating principles (See Appendix E). This is the **inter-peak beating** (IPB) function demonstrated in Section 1.2.2. Since IPB gives similar results of PB with respect to the selection of candidates, it will not be evaluated at the end of this chapter.

6.2 Extraction of Non-Harmonically Related F0s (NHRF0s)

The method based on HAR is prone to select subharmonic candidates, whereas the method based on PB is prone to select super-harmonic candidates. No matter which methods, setting a threshold of a polyphonic salience function has the drawback that the compromise made between the selection of weaker sources and the reduction of the number of F0 candidates is unavoidable. This is because the polyphonic salience functions are often normalized with respect to the total energy of the signal. When a strong source is present in the signal, the other weaker sources may not have competitive salience compared with that of the F0 hypotheses related to the harmonics of the strong source. One way to deal with this normalization issue is to iteratively extract F0 candidates and suppress their harmonics to update the normalization factor of the salience function. Suppressing an extracted source has the advantage that the salience function can be normalized with respect to the residual of the signal. In this way, the common subharmonics of the correct F0s can be gradually suppressed and the weak sources can thus be iteratively extracted (see Figure 6.2). However, the iterative suppression of all partials of an extracted source has the danger of suppressing the salience of its HRF0s as well. In consequence, the salience of HRF0s will not be significant enough to be extracted afterwards. To address this issue, an iterative algorithm is proposed to extract the candidates of NHRF0s as well as those of HRF0s.

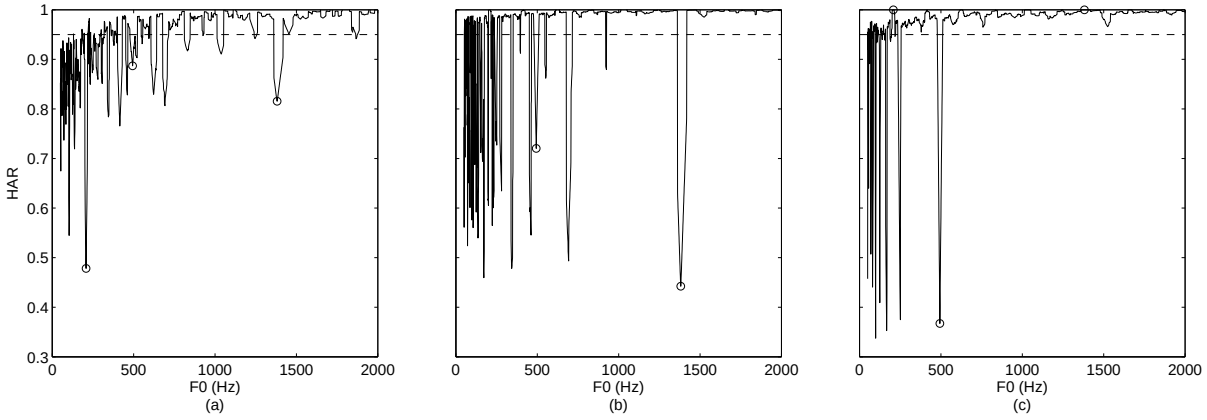


Figure 6.2: Iterative suppression of the extracted sources and the resulting HAR. The three circles mark the salience of the three notes: Ab3 ($\approx$ 208Hz), B4 ($\approx$ 494Hz) and F6 ($\approx$ 1382Hz) mixed in the signal. (a) original HAR; (b) HAR after suppressing the note Ab3; (c) HAR after suppressing the notes Ab3 and F6.

The extraction of NHRF0s involves three parts: predominant-F0 estimation, the verification of an extracted F0 candidate and a criterion to stop the iteration (see Figure 6.3). For predominant-F0 estimation, the score function is used to extract the most probable F0. To suppress an extracted source, the *peak weights* (see eq.(5.6)) of the related peaks matched to the partials are set to zero. The harmonic matching of the HAR criterion is thus normalized with respect to the remaining spectral peaks. The other score criteria are fixed on the assumption that even without the treatment of the overlapping partials, MBW, SPC and SYNC are still able

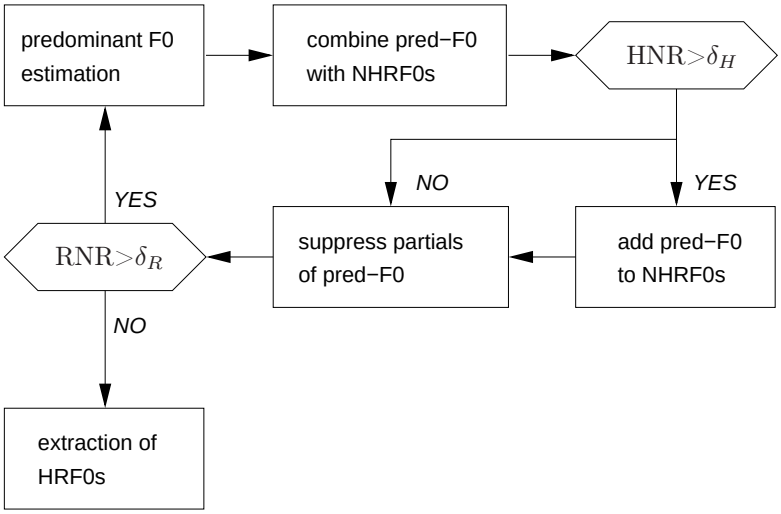


Figure 6.3: Iterative extraction of F0 candidates.

to collectively disfavor the candidates of subharmonic F0s and those of super-harmonic F0s. To avoid the extraction of spurious candidates, the HPS of the predominant F0, denoted by "pred-F0", is first constructed, considering its combination with the HPS of the extracted NHRF0s (see Section 5.1), and then its **harmonic-to-noise ratio** (HNR) (Qi and Hillman, 1997) is evaluated. The HNR is defined as the average ratio between the *partial* peak amplitudes and the related noise level, considering only the first 10 partials. If the HNR is smaller than the threshold δ_H, the extracted F0 is considered spurious and is then discarded. In order to calculate a more reliable HNR, the overlapping partials are inferred from the combination of the HPS of the extracted NHF0s and then reallocated (see Section 5.1.2). The verification of a pred-F0 by its HNR is meant to compensate the predominant-F0 estimation part in which the overlapping partials are not taken care of. The iterative extraction of NHRF0s continues until the **residual-to-noise ratio** (RNR) is below the threshold δ_R. The RNR is defined as the average ratio between the *residual* peak amplitudes and the related noise level, considering only the residual peaks larger than the related noise level.

In order to study the thresholds δ_R and δ_H for RNR and HNR, respectively, the performance of the score function is investigated using the tails of harmonic instrument samples where the partials become weaker. On the assumption that there is only one harmonic source in each note of instrument sound samples, their RNR can serve as the reference values for δ_R to prevent the extraction of spurious sources. On the other hand, HNR characterizes the SNR limit of the quasi-periodic parts that an F0 estimator is able to extract. It is expected that the single-F0 estimator fails when the harmonicity/periodicity is low and noise is dominant. For each single-note sample, the *lowest* RNR and the *lowest* HNR in all the analysis frames in which the F0s are correctly estimated are used to learn the thresholds δ_R and δ_H, respectively. Each threshold is learned as a function of the MIDI note number (see Figure 6.4). To learn the threshold δ_R for each MIDI note, it is proposed to use the *maximal* RNR of all the related samples. Because there are fewer residual peaks in polyphonic signals than those in monophonic signals, it is expected to compensate the normalization issue by taking the maximum. δ_R is thus determined by the

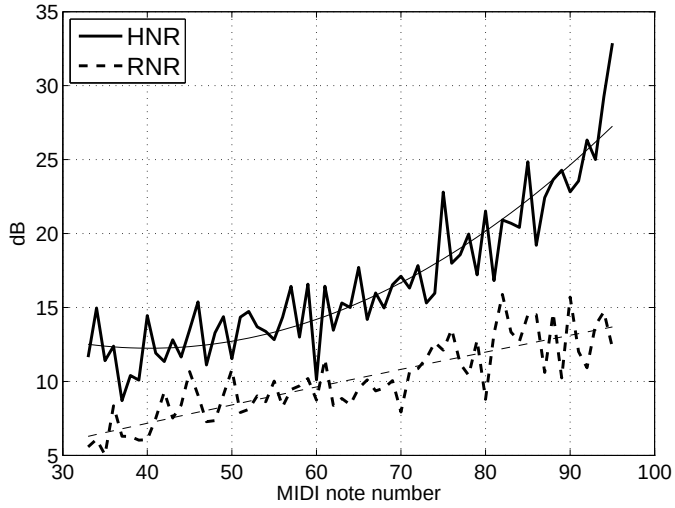


Figure 6.4: The note-dependant thresholds for HNR and RNR, learned from the tails of harmonic instrument sound samples for the notes ranging from Ab1 to B6. The thick lines represent the learned thresholds, and the thin lines represent their second-order polynomial approximation curves.

second-order polynomial approximation to $\mathrm{RNR}_{\max}$ (the dash lines in Figure 6.4). During the process of iterative extracting NHRF0s, δ_R is updated with respect to the lowest F0 extracted. To learn the threshold δ_H for each MIDI note, it is proposed to use the *average* HNR of all the related samples (the solid lines in Figure 6.4). However, the learned HNR increases rapidly as the F0 becomes higher, which implies that the score function tends to fail for high F0s. Since the HNR learned in this way does not provide reasonable results for some instances, the threshold is set according to δ_R and the average difference between the two approximation curves: $\delta_H = \delta_R + 6.67dB$.

6.3 Detection of Harmonically Related F0s (HRF0s)

Each NHRF0 represents a *harmonic group* in which HRF0s are to be extracted. This process, called **harmonic split**, extracts the hypothetical sources of HRF0s from the HPS of NHRF0s. The idea of harmonic split can be demonstrated by a simple example (see Figure 6.5). Consider three harmonic sources of fundamental frequencies $F0_a$, $F0_b$ and $F0_c$, in which $F0_b/F0_a = 3/2$ and $F0_c = 2 \cdot F0_a$. Assuming that the NHRF0s $F0_a$ and $F0_b$ are extracted in the previous stage, harmonic split aims at extracting $F0_c$ from the HPS of $F0_a$. The harmonic relation between $F0_a$ and $F0_c$ causes the partials of $F0_a$ to boost at regular frequency intervals (every $2 \cdot F0_a$ in this case). Therefore, it is assumed that as long as a HRF0 is dominant and disturbs the envelope smoothness of the related NHRF0, it is reasonable to consider the HRF0 to be an F0 candidate.

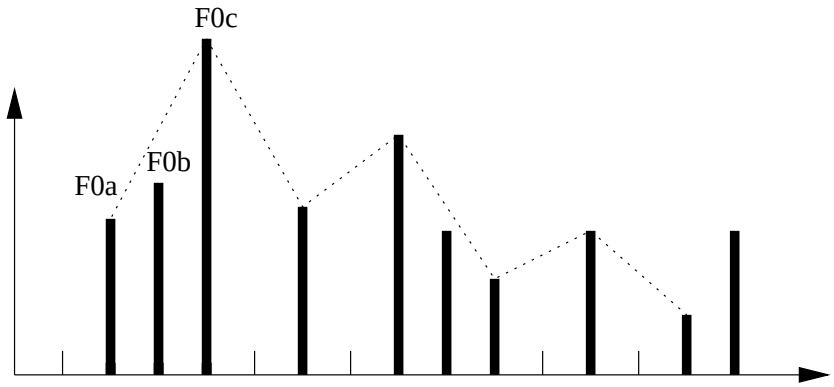


Figure 6.5: The extraction of a HRF0 $F0_c$ from the HPS of a NHRF0 $F0_a$. $F0_a$ and $F0_b$ are related by a perfect fifth in a well-tempered scale. $F0_a$ and $F0_c$ are related by one octave.

Note-dependent tone models

To measure how much the smoothness of a spectral envelope is disturbed, it is necessary to refer to certain models upon the comparison with the observed envelopes. Through the development of the score function, the bandwise smooth model is used, which makes use of the interpolation of partial amplitudes (see Section 5.1.1). However, this model serves to estimate the expected amplitudes of overlapping partials on condition that the overlap positions have been inferred from a set of F0 hypotheses. In the iterative extraction process, F0s are consecutively extracted and it is very difficult to predict the overlapping partials. In addition, harmonic instrument sounds are of various resonance characteristics, which often results in boosted partials within the resonance frequency ranges. In consequence, unsmooth envelopes shall be allowed. In order to quantify the allowable degree of disturbance, the note-dependent tone models of harmonic instrument sounds are proposed. The reason to use the note-dependent tone models is that the underlying musical instruments are usually not known *a priori* in the problem of multiple-F0 estimation.

A tone model of a note is defined by its partial amplitude sequence, which can be learned from a collection of musical instrument sound samples: McGill University Master Samples, Iowa University Musical Instrument Samples, IRCAM Studio On Line and RWC Musical Instrument Sound Database. For each sound sample, the partial sequences at each frame are extracted and then weighted by its harmonicity to favor the estimates of good periodicity. The weighted partial sequences are averaged for all the instruments playing at the same note. The note-dependant tone models are trained for the notes from Ab1 to B6. The partial amplitudes of the models are normalized with respect to the first partial, i. e., the fundamental. It is proposed to learn two types of tone models for each note: **low-fundamental model** and **high-fundamental model**. The low-fundamental model is of a weak fundamental, which represents a spectral envelope with boosted partials at resonance frequencies higher than the first partial (see Figure 6.6(a)). On the other hand, the high-fundamental model is of a strong fundamental, which represents a spectral envelope with a fast decay for higher partials (see Figure 6.6(b)). The 1/k smooth model (see Section 2.1.2) is a kind of the high-fundamental model.

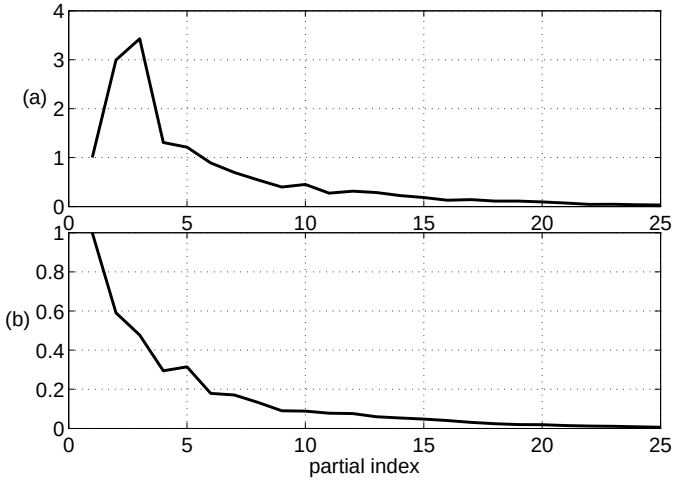


Figure 6.6: Two types of tone models for the note E3: (a) low-fundamental model; and (b) high-fundamental model.

Given an observed partial sequence $\mathbf{P}$ and the related note, the likelihood of the low-fundamental model $\mathbf{M}_L$ and that of the high-fundamental model $\mathbf{M}_H$ are compared. The more likely one is selected for the calculation of the envelope disturbance measure. The likelihood of a model is evaluated by means of the least-square error estimation. A smaller error between $\mathbf{P}$ and $\mathbf{M}$ stands for a higher likelihood. By introducing a scaling factor s_p for $\mathbf{P}$, the least-square error estimation minimizes $\parallel s_p \cdot \mathbf{P} - \mathbf{M} \parallel^2$ with

$$s_p = \frac{\mathbf{P} \cdot \mathbf{M}}{\parallel \mathbf{P} \parallel^2} \tag{6.2}$$

Since the observed $\mathbf{P}$ may contain overlapping partials, the scaling factor is estimated from the first three partials on the assumption that the overlaps at the first several partials have less impact on the partial amplitudes and s_p can be reasonably deduced.

Envelope disturbance measure

By comparing the observed partial sequence of a NHRF0 with the selected tone model, HRF0s are going to be extracted according to an envelope disturbance measure. For a HRF0 hypothesis at the kth harmonic of a NHRF0, the **envelope disturbance** (ED) is evaluated. ED is the mean amplitude difference of the first five harmonics between the model and a HRF0 hypothesis, considering only the partials that are larger than those of the tone model. The threshold of ED is trained for each partial of a note, using harmonic instrument sound samples as a reference to allow certain unsmoothness of the spectral envelopes. ED is similar to the **spectral irregularity** proposed by Zhou (2006). Instead of using the tone models learned from harmonic instrument sounds, he used the partial interpolation to estimate the spectral envelope. The issue of this method is that the overlapping partials might be used for the interpolation to estimate the "expected" envelope, which is the main reason why here the note-dependant models are used for the extraction of HRF0s.

6.4 Evaluation

Following the same experiment setup for the joint estimation algorithm (see Section 5.4), the three proposed methods for F0 candidate selection are evaluated in terms of multiple-F0 estimation error rate, predominant-F0 estimation error rate, the number of F0 candidates, and the candidate selection error rate. Two window sizes, 186ms (see Figure 6.7 and Figure 6.9) and 93ms (see Figure 6.8 and Figure 6.10), are tested. Compared to the iterative extraction method, the two methods using HAR and PB, respectively, in general have lower error rates in multiple-F0 estimation, predominant-F0 estimation, and F0 candidate selection. However, the robustness seems to be gained from the increase of the number of F0 candidates. HAR in average selects more than 30 candidates. PB has the best robustness but the number of candidates increases with the increase of the polyphony. As described in Section 6.1, this drawback may arise from the increasing number of partials, which results in more spurious candidates. The iterative method ITR in general selects 15 candidates. Although the overall performance of ITR is worse, it does reduce considerably the required computations, especially for the higher polyphony. For example, in the case of five concurrent sources, ITR reduces more than one thousand times of combinations compared to HAR. The price to pay is the increase of the error rate by only $1 \sim 2\%$ for multiple-F0 estimation, which is considered acceptable for the related gain in efficiency.

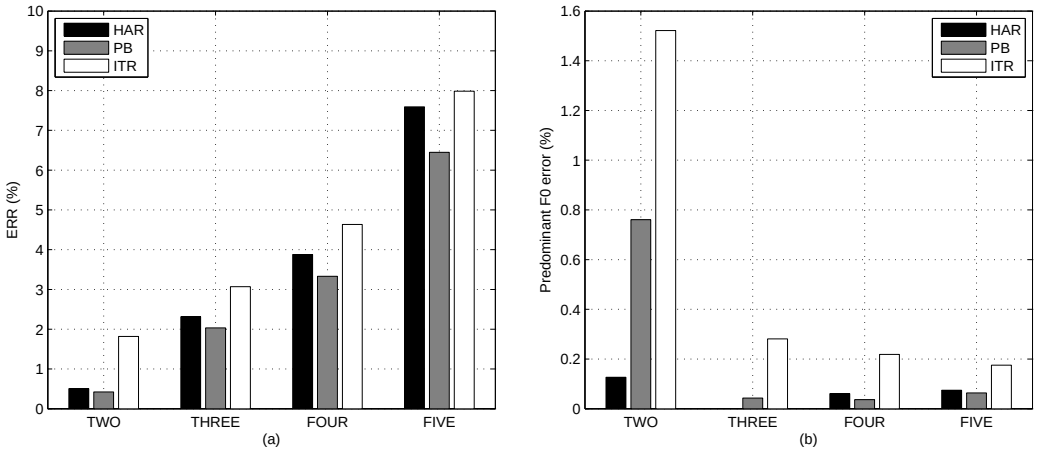


Figure 6.7: Multiple-F0 estimation error rates and predominant-F0 estimation rates for 186ms window size.

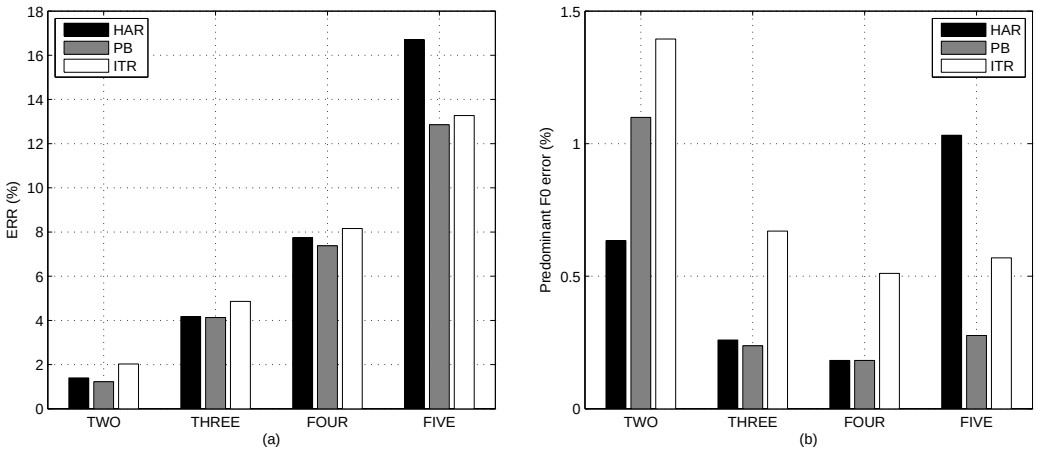


Figure 6.8: Multiple-F0 estimation error rates and predominant-F0 estimation rates for 93ms window size.

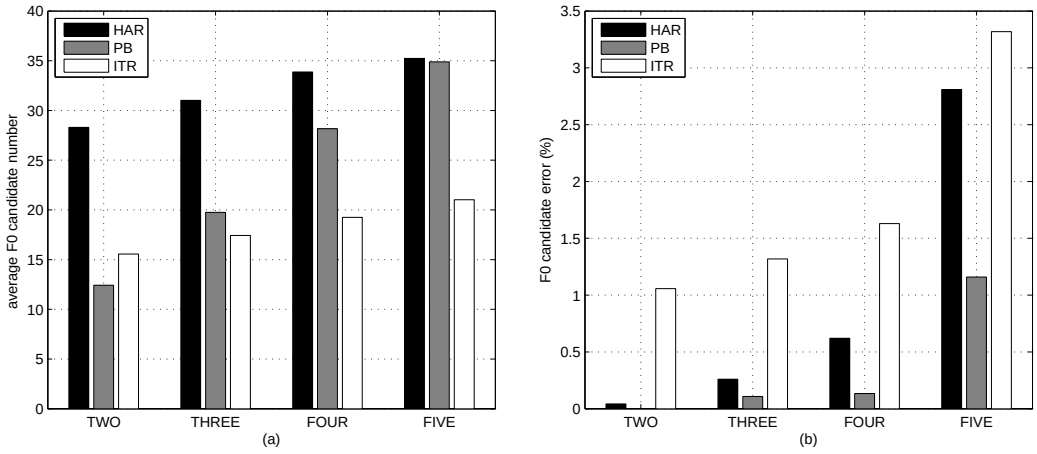


Figure 6.9: Average F0 candidate number and F0 candidate selection error rates for 186ms window size.

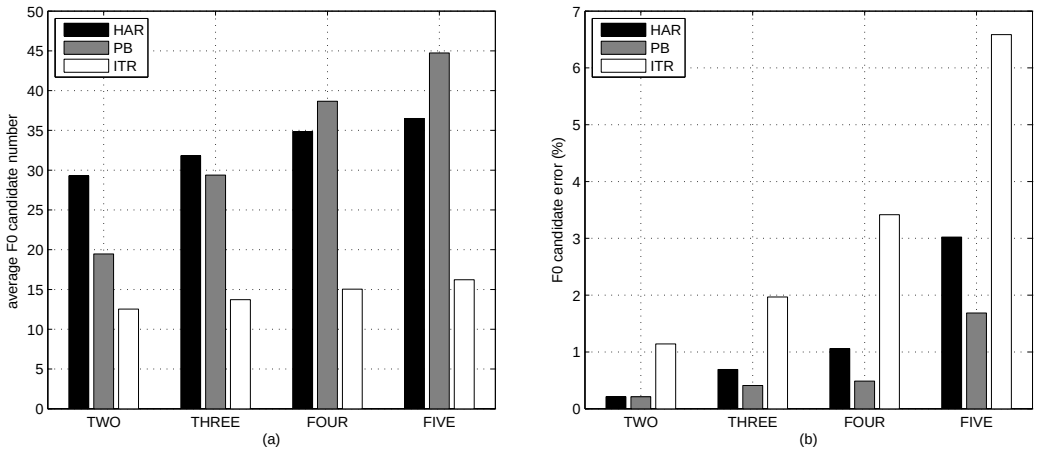


Figure 6.10: Average F0 candidate number and F0 candidate selection error rates for 93ms window size.

7

ESTIMATING THE NUMBER OF CONCURRENT SOURCES

Of the three fundamental models in the multiple-F0 estimation problem, the modeling of the noise is handled by the adaptive noise level estimation algorithm, and the modeling of a combination of quasi-periodic sources is explored in the development of the joint estimation algorithm that takes care of the overlapping partials. The last problem yet to study is the estimation of the number of sources, called **polyphony inference**. The proposed strategy is to first estimate the maximal polyphony and then to consolidate the F0 estimates according to two criteria: the explained energy and the improvement on the spectral smoothness. In the polyphony inference stage, the estimated noise level and the most probable hypothetical combinations lay a foundation for the final verification of NHRF0s as well as HRF0s.

In this chapter, the polyphony inference algorithm is presented, which completes the multiple-F0 estimation system. In order to evaluate the proposed system, a systematic method is proposed to create a polyphonic music database. Then, the proposed system is evaluated for polyphony settings ranging from one to six. The result is compared with a previous version which has been evaluated in the public competition at MIREX 2007. Finally, a multiple-F0 tracking algorithm for monodic instrument solo recordings is presented as a practical application.

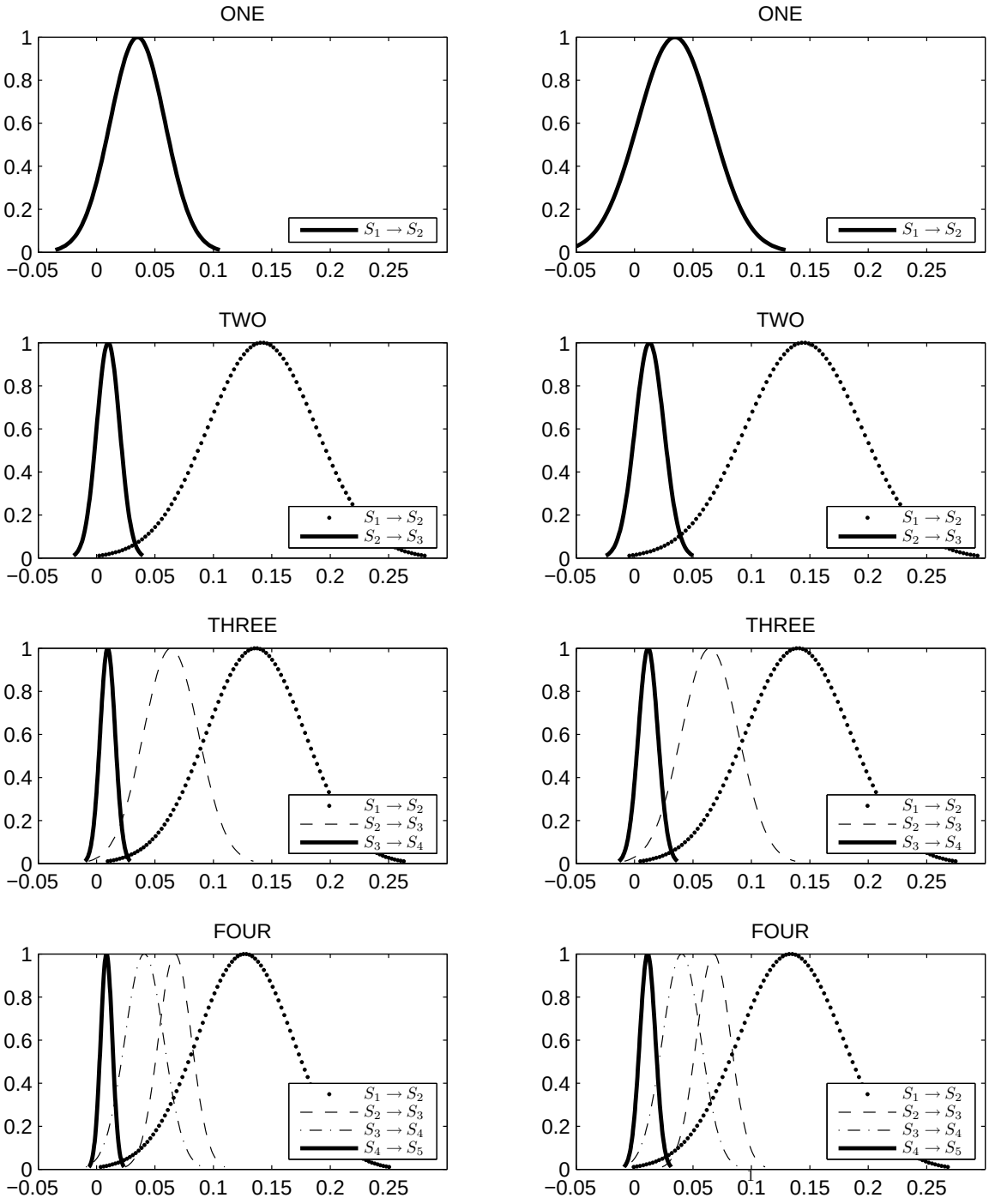
7.1 Polyphony Inference

7.1.1 Estimation of the maximal polyphony

Supposing that the inference of the polyphony starts from one hypothetical source and adds gradually one source after another, the combination of the hypothetical sources should explain more and more energy and adapt the resulting spectral envelopes to be smoother and smoother. Theoretically, it is reasonable to expect the combination of the correct number of F0s to give the highest score. Therefore, it is proposed to investigate how the score improves when hypothetical sources are added consecutively. The artificially mixed polyphonic database (see Section 7.2.2) is used for this inspection.

The correct polyphony is denoted by N; the polyphony hypothesis is denoted by m. Given the correct polyphony, the combinations of $\{m = 1, \cdots, m, \cdots, M, M+1 = N+1\}$ sources are evaluated by the score function. The best score of each hypothesis, denoted by S_m, is used to calculate the **score improvement**. It is observed that as the polyphony hypothesis is increased towards N, the resulting score in general becomes better, whereas the score improvement gradually approaches zero. To illustrate the observed score improvements, they are modeled by Gaussian distributions with respective means and variances (see Figure 7.1). The observation does not come up to the expectation that the combination with the correct polyphony always gives the best score. The combinations of $M+1$ hypothetical sources in general give rise to a better score than those of M hypothetical sources do. However, the score improvement can be a useful criterion because an additional hypothetical source does not significantly improve the score. In this way, the consecutive combination of increasing polyphony can be terminated by setting a threshold Δ_s for the score improvement when the polyphony hypothesis is increased from M to $M+1$. When $S_{M+1} - S_M < \Delta_s$, the hypothesis M is considered the most plausible number of sources. The polyphony is correctly inferred if $M = N$.

It is found that the Gaussian distributions do not have a unique model of $S_m \rightarrow S_{m+1}$ for each polyphony. In consequence, setting a universal threshold for each correct polyphony may not work properly. Another possibility could be to model the probability of a polyphony hypothesis M as $p(M) = \prod_{m=1}^{M} p(S_m \rightarrow S_{m+1})$ and choose the polyphony hypothesis that maximizes the resulting probability. However, it is found that the parameters of the Gaussian models depend on the diversity of harmonic instrument sounds and the way they are mixed. Therefore, it is suggested to use the score improvement for the estimation of the **maximal polyphony**. Beginning with the hypothesis $m = 1$, the score improvements of the polyphony hypotheses $\{m = 1, \cdots, m, \cdots, M, M+1\}$ are consecutively evaluated till the last hypothesis $M+1$ is reached on condition that $|S_{M+1} - S_M| < \Delta_s$. The hypothesis M is considered to be the inferred maximal polyphony M_{max}. In the current implementation, Δ_s is set at 0.01 to terminate the consecutive combination.



(a) window size 186ms

(b) window size 93ms

Figure 7.1: Score improvement observations modeled by Gaussian distributions. $S_m \to S_{m+1}$ represents the score improvement from the polyphony hypothesis m to $m+1$. Column (a) shows the tests using the 186ms window size; column (b) shows the tests using the 93ms window size.

7.1.2 Consolidation of multiple F0 hypotheses

Given the number of F0s, the four score criteria HAR, MBW, SPC and SYNC collectively give a score to a hypothetical combination of F0s, which determines its ranking in all possible combinations. When the polyphony is given, the combination scored the best gives the related F0 hypotheses as the final estimate. The weighting parameters of the score criteria have been optimized by the evolution algorithm such that the correct combination is ranked the best in general. However, the polyphonic sample database used for training consists of signals mixed with sources of equal energy. That is, the weighting parameters do not guarantee to perform well when the underlying sources are of different energy. In fact, most errors in the test described in Section 5.4 occur when the fixed weighting of score criteria does not guarantee to rank the correct F0s the best. However, the correct F0s are generally ranked in the first several places. Therefore, it is possible to infer the correct F0s from the combinations that are ranked in the first several places. This consolidation process is important for the case in which the concurrent sources are of different energy.

The polyphony inference algorithm is shown in the following pseudo code table. Given all the top-five combinations for all polyphony hypotheses, the individual F0s are first listed in order of their salience. The inference algorithm begins with the F0 hypotheses found in the top-five combinations with polyphony hypothesis M_{max}. Then, the best combination is going to be inferred by iteratively validating the F0 hypotheses. Beginning with the most likely F0, each F0 hypothesis is gradually added to verify their contributions to the explained energy and the spectral smoothness. If an F0 hypothesis (to be added) is higher than any of the previously selected F0 hypotheses, it shall either explain more energy or improve the envelope smoothness of the HPS that have partials overlapping with its HPS. On the other hand, if an F0 hypothesis (to be added) is lower than any of the previously selected F0s, it shall explain more energy. Otherwise, it is considered a spurious source that is composed of noise. When an F0 hypothesis meets the requirements for a valid source, it is removed from the F0 hypothesis list and added to the set of the final estimates. For each polyphony hypothesis, the algorithm searches for the matched combinations. When no matched combination is found, the validation process stops. The polyphony is thus inferred along with the set of F0 estimates.

The individual salience of an F0 hypothesis is derived from the individual score that is defined by the score function in which the combinatorial criteria are replaced by the individual criteria. For an F0 hypothesis f within the cth-ranked combination of the polyphony m, the individual score is

$$\varsigma_{f,m,c} = p_1 \cdot (1 - \mathrm{eff}_{f,m,c}) + p_2 \cdot \mathrm{mbw}_{f,m,c} + p_3 \cdot \mathrm{spc}_{f,m,c} + p_4 \cdot \mathrm{sync}_{f,m,c} \tag{7.1}$$

where $\mathrm{eff}_{f,m,c}$ is its effective salience. Reminded that the score function returns a score between 0 and 1 that is inversely proportional to the plausibility of a combination, "*1-score*" is used as the salience measure. Accordingly, the salience of an F0 hypothesis is defined as

$$\varsigma_f = \frac{1}{N_f} \sum_{m,c} (1 - \varsigma_{f,m,c}) \cdot (1 - \mathcal{S}_{m,c}) \tag{7.2}$$

Algorithm: polyphony inference

input : The list of F0 hypotheses $\mathcal{H} = \{\mathrm{F0}_1, \mathrm{F0}_2, \cdots, \mathrm{F0}_Q\}$ in order of salience along with the top-five combinations for all polyphony hypotheses $\{\mathcal{C}_m\}_{m=1}^{M_{max}}$

output: The inferred polyphony M with the set of selected F0s $\mathcal{F}$

Initialization of $\mathcal{F} = \{\mathrm{F0}_1\}$ and $M = 1$
Initialization of residue $E_R = 1$
Initialization of the F0 hypothesis set $\mathcal{T} = \{\mathrm{F0}_2, \cdots, \mathrm{F0}_Q\} = \{\mathrm{F0}(j)\}_{j=1}^{J}$ to combine with $\mathcal{F}$
while $J > 0$ **do**
  for $c = 1$ *to* J **do**
    if $\{F0_c \bigoplus \mathcal{F}\} \bigcap \mathcal{C}_{M+1}$ **then** `/* The matched combination is found */`
      if $E_{eff}(F0_c) > E_{noise}$ **then** `/* The effective salience is larger than noise salience */`
        if *F0$_c$ overlaps lower F0s in* $\mathcal{F}$ **then**
          if $\max\{\Delta MBW\}_{m=1}^{M} > \Delta MBW_{model}$ **then** `/* Smoother envelopes */`
            add $\mathrm{F0}_c$ to $\mathcal{F}$
            update M and E_R
            break the For loop
          else
            if $\Delta E_R > E_{noise}$ **then** `/* reduction in residual salience larger than noise salience */`
              add $\mathrm{F0}_c$ to $\mathcal{F}$
              update M and E_R
              break the For loop
            end
          end
        else `/* F0`$_c$ `is the lowest */`
          if $\Delta E_R > E_{noise}$ **then**
            add $\mathrm{F0}_c$ to $\mathcal{F}$
            update M and E_R
            break the For loop
          end
        end
      end
    end
  end
  if *any F0 is added to* $\mathcal{F}$ **then**
    update $\mathcal{T}$ and J by removing the selected F0s
  else
    termination of the While loop
  end
end

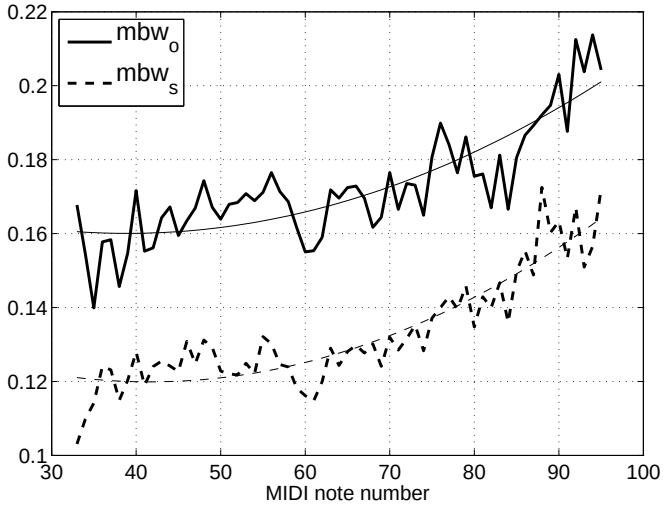


Figure 7.2: MBW comparison between those of the original spectral envelopes and those of the smoothed spectral envelopes. The two thin lines are second-order polynomial fitting the trained MBW data.

The individual salience is first weighted by the combinatorial salience $1-\mathcal{S}_{m,c}$ and then summed for all the matched combinations of all polyphony hypotheses. N_f is the normalization factor. In this way, an F0 hypothesis appearing in the combinations with higher score is considered more important. And an F0 hypothesis appearing more frequently in combination with other F0 hypotheses is considered more important as well.

Once the F0 hypotheses are ranked in order of salience, the validation process makes use of two criteria, the *explained energy* and the *improvement of spectral smoothness*, to iteratively add an F0 hypothesis to the set of final estimates. The explained energy is evaluated in terms of the reduction in the *residual salience*, denoted by ΔE_R. The residual salience is defined as the sum of the peak salience of the remaining peaks that are not yet explained. ΔE_R therefore characterizes the energy explained exclusively by an F0 hypothesis. A NHRF0 hypothesis should explain most sinusoidal peaks such that ΔE_R is larger than the noise salience E_{noise}. The noise salience is determined, at the end of the F0 candidate selection process (see Section 6.2), by summing the peak salience of the peaks classified as noise. E_{noise} serves as the threshold for the decrease in the residual salience ΔE_R. In this case, an added F0 hypothesis is considered valid if it explains the residual peaks with an amount more than the noise salience, that is, $\Delta E_R > E_{noise}$. This condition is important for the validation of a NHRF0 hypothesis because its non-overlapping partials should explain a significant amount of salient peaks. Notice that either a NHRF0 hypothesis or a HRF0 hypothesis should have its effective salience larger than the noise salience ($E_{eff} > E_{noise}$).

The improvement of spectral smoothness is an important criterion for the validation of HRF0s because adding a HRF0 usually improves the smoothness of the spectral envelopes of the extracted sources. Since an additional HRF0 tends to improve the resulting spectral smoothness as well, it is necessary to put a constraint on the improvement of spectral smoothness. To achieve this goal, it is proposed to observe the variation of the score criterion MBW. Notice

that MBW is designed in a way that smoother envelopes result in smaller values. The improvement of spectral smoothness is required to exceed what can be allowed for harmonic instrument sounds. To learn the threshold of MBW as the allowed improvement of a spectral envelope, selected instrument samples of RWC Musical Instrument Sound Database (Goto, 2003) are used. Given an observed partial sequence of a harmonic sound, the hypothetical sources of the F0s at the partial frequencies are considered the HRF0 hypotheses. For each HRF0 hypotheses, the decrease of MBW, denoting ΔMBW, are evaluated. ΔMBW is the difference of MBW before, denoted by mbw_o, and after, denoted by mbw_s, smoothing out [1] the overlapping partials of a HRF0 hypothesis. For each analysis instance, mbw_o of the correct F0 and mbw_s of the HRF0 hypothesis that results in the maximal ΔMBW are retained. For each musical note, the calculated mbw_s and mbw_o are averaged for all the analysis instances of all the instruments (see Figure 7.2). They are further modeled, as a function of the MIDI note numbers, using a second-order polynomial. The threshold for the improvement of spectral smoothness is then defined as $\Delta\text{MBW}_{model} = (mbw_o - mbw_s)/mbw_o$.

7.2 Database Construction

To study and to evaluate a multiple-F0 estimation algorithm, a polyphonic database with a representative corpus and verifiable ground truth is necessary. For speech signals, there exist quite a few monophonic databases ready for generating polyphonic speech signals (Wu *et al.*, 2003; Roux *et al.*, 2007). For music signals, there exist nowadays three types of polyphonic music signals used as evaluation corpus:

1. Mixtures of monophonic samples
   With a variety of musical instrument sound samples available, polyphonic signals can be mixed either randomly (Klapuri, 2003; Yeh *et al.*, 2005), or musically (Kitahara *et al.*, 2007; Li and Wang, 2007).

2. Synthesized music from MIDI files
   Synthesized polyphonic music can be rendered from MIDI files by sequencers with sound modules (Kashino and Tanaka, 1993; Dixon, 2000; Marolt, 2004; Sterian, 1999) or samplers (Kashino *et al.*, 1998).

3. Real recordings
   Real recordings can be recordings of multi-tracks [2] or stereo/mono mix-down tracks (Marolt, 2004; Goto *et al.*, 2002; Goto, 2003). Some use YAMAHA Disklavier for piano recordings triggered by MIDI events (Poliner and Ellis, 2006; Monti and Sandler, 2002; Bello, 2003).

Provided that the selected single-F0 estimation algorithm is robust, the ground truth F0s of mixtures of monophonic samples can be individually estimated from the respective monophonic sources. The concern, however, is that the final mixtures may not have the same statistical

[1] A smoothed out partial is replaced by the amplitude interpolation of its adjacent partials.
[2] `bass-db.gforge.inria.fr/BASS-dB/`

properties as those found in music. To increase the relevance of the test corpus for real world applications, the corpus should take the musical structure into account. From this point of view, synthesized music from MIDI files and real recordings are more suitable as corpora. Despite the wide availability of music corpora, the establishment of their ground truth remains an issue.

7.2.1 Annotating real recordings

Nowadays, more and more evaluations of multiple-F0 estimation algorithms use real recordings of mix-down tracks (Ryynänen and Klapuri, 2005; Kameoka *et al.*, 2005b). The annotation process usually starts with a reference MIDI file, followed by the alignment of the note onsets and offsets with the observed spectrogram. Assuming that the notes in the reference MIDI file correspond exactly to what have been played in the authentic performance, the annotation process is in fact the **score alignment** with the recorded signals. In order to further discuss the issues concerning the annotation of real recordings, a score alignment procedure is described in the following.

Given a reference MIDI file and a real recording of the same musical piece, the MIDI notes are first aligned with the real recording automatically (Rodet *et al.*, 2004; Kaprykowsky and Rodet, 2006). Then, the details like note offs, slow attacks, etc., are manually corrected using AudioSculpt (Bogaards *et al.*, 2004). Innovative tools in AudioSculpt have been developed to facilitate verification and modification of signal analysis and manual annotation (see Figure 7.3). Automatically aligned MIDI notes are manually corrected according to the following procedure:

1. Overlay MIDI notes on the spectrogram as a piano-roll like representation. Adjust *MIDI note grid* by tuning for the best reference frequency at note A4.

2. Generate time markers by automatic onset detection (Röbel, 2006) and adjust the probability threshold according to the observed spectrogram.

3. Verify and adjust note onsets detected around transient markers visually and auditorily. In addition to the waveform and spectrogram, the *harmonics tool*, *instantaneous spectrum* (synchronous to the navigation bar), etc., provide visual cues for the evolution of harmonic components. The *diapason* allows accurate measurement and sonic synthesis at a specific time-frequency point. *Scrub* provides instantaneous synthesis in a single FFT frame, which allows users to navigate auditorily at any speed controlled by hand. Users can also listen to arbitrarily shaped time-frequency zones.

4. Align markers automatically with the verified transient markers using *magnetic snap*.

5. If any inconsistency is found between the MIDI file and the real performance, missing notes can be added and unwanted notes eliminated.

Despite all the powerful tools for manual annotation, timing ambiguities need to be resolved based on subjective judgements. Above all, for reverberated recordings, reverberation extends the end of notes and overlaps the following notes in time and in frequency. If one aims at finding out when a musician stops playing a note, a scientific description of reverberation (Baskind,

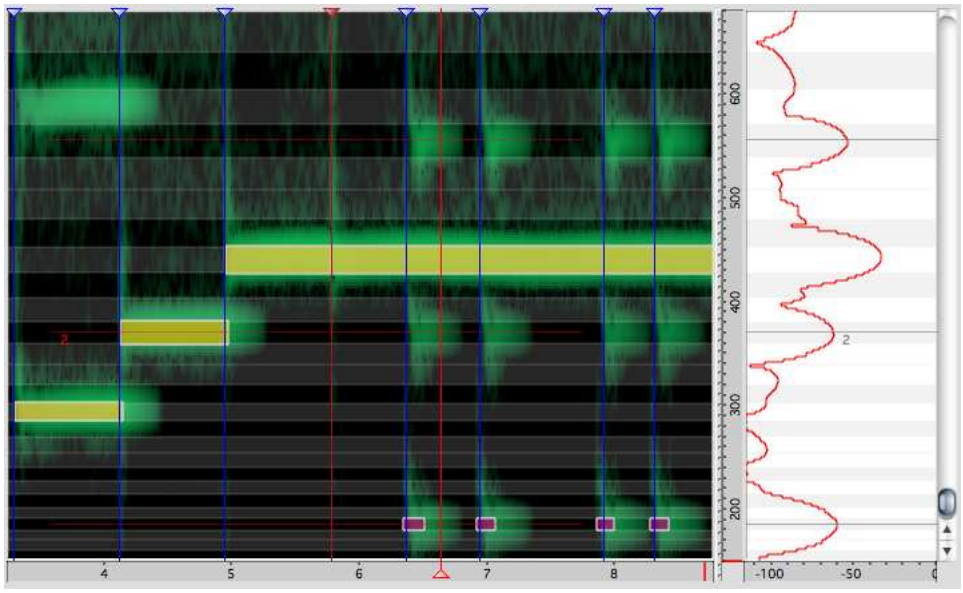


Figure 7.3: Screenshot of AudioSculpt during annotation showing MIDI notes, MIDI note grids, onset markers, the instantaneous frequency spectrum and the harmonics tool.

2003) is necessary to identify the end of the playing. Due to reverberation, real recordings of monodic instruments usually appear to be polyphonic, which requires a multiple-F0 tracking (Yeh *et al.*, 2006). However, the description of reverberation is not yet available for polyphonic recordings in a reverberant environment. Alternatively, if one defines the end of a note as the end of its reverberated part, the ambiguity occurs when (1) certain partials are boosted by the room modes and extended longer than the others, and when (2) reverberation tails are overlapped by the following notes and the end of reverberation is not observable.

If manual annotation/alignment is reliably done for non-reverberated recording, it is still disputable in what accuracy one can extract multiple F0s as ground truth. Due to all the issues discussed above, evaluation based on unverifiable reference data endangers the trustworthiness of the reported performance. Therefore, it is believe that ground truth shall be derived by means of an automatic procedure from the isolated clean notes of the polyphonic signals. However, real recordings with clean and separate notes are almost impossible to carry out. One possible way could be to record multiple tracks of solos in a non-reverberant environment and mix the tracks afterwards. However, a solo track can be polyphonic for instruments like pianos or guitars, which still raise the issues in annotation. Therefore, two methods are proposed to build polyphonic music databases for the evaluation of multiple-F0 estimation algorithms. One follows the method proposed by Klapuri (2003) which randomly mixes monophonic instrument samples to generate polyphonic signals. The other is to render isolated notes from MIDI files to generate polyphonic music. Both methods assure the access to monophonic signals of which the ground truth can be reliably established and easily verified.

7.2.2 Artificially mixed polyphonic samples

Polyphonic samples are generated by mixing monophonic samples of four databases: McGill University Master Samples [1], Iowa University Musical Instrument Samples [2], IRCAM Studio On Line [3] and RWC Musical Instrument Sound Database [4] (see Table 7.1). There are in general three playing dynamics: *ff*, *mf* and *pp*. For certain instruments in the four databases, sounds with different playing techniques are also recorded. For example, both playing straight and playing vibrato are included for wind instruments. For bowed string instruments, recordings are also made by string plucking (*pizzicato*). Although mallet percussion instrument sounds evoke distinct pitches, they are excluded because of their unique partial structures.

Instrument Family	Instruments
Reed	bassoon, clarinet, oboe, saxophone, English horn, accordion, etc.
Flute	flute, pan flute, piccolo, recorder, shakuhachi, organ, etc.
Brass	cornet, French horn, trumpet, trombone, tuba, etc.
Plucked string	archlute, guitar, harp, harpsichord, shamisen, etc.
Struck string	piano
Bowed string	violin, viola, cello, double bass, etc.

Table 7.1: Selected harmonic instruments from four musical instrument sample databases.

Four polyphonic mixture sets are generated: two-voice, three-voice, four-voice and five-voice mixtures, labeled as TWO, THREE, FOUR and FIVE respectively. To mix M-voice polyphonic samples, M out of twelve tones (C, Db, D, Eb, E, F, Gb, G, Ab, A, Bb and B) are preliminarily assigned and then the related samples ranging from 50Hz to 2000Hz (corresponding to notes from Ab1 to B6) are randomly selected to mix. Segments of good periodicity in monophonic samples are selected and then mixed with equal mean-square energy. Around 1500-2000 samples for each database are generated in a way that each combination of tones are of equal proportion. The ground truth of multiple F0s is established by estimating the F0 of each monophonic sample before mixing. The F0 search ranges are limited to one half tone around the related note. Random mixing has the advantage of generating a great variety of polyphonic signals in a small number of instances, which facilitates the training and the testing of a multiple-F0 estimation algorithm.

7.2.3 Synthesized polyphonic music

In order to perform a musically relevant evaluation, synthesized polyphonic music more appropriate than random mixtures of monophonic samples. The biggest advantage of synthesized music is that one can have access to every single note from which the ground truth can be established. The argument against synthesized music is often that it is "non-realistic", but few have doubts about the ground truth. It seems that MIDI note event data is considered the ground

[1] http://www.music.mcgill.ca/resources/mums/html/index.htm
[2] http://theremin.music.uiowa.edu/MIS.html
[3] http://forumnet.ircam.fr/402.html?&L=1
[4] http://staff.aist.go.jp/m.goto/RWC-MDB/rwc-mdb-i.html

truth, but it is not true. In fact, MIDI `note off` events are messages requesting the sound modules/samplers to start rendering the end of notes, which usually extends the notes to sound longer after `note off`. Thus, creating the reference data for the rendered audio signal from its original MIDI file is not straightforward. The extended note duration depends on the settings of sound modules or samplers, which is controllable and thus predictable. In order to retain each sound source for reliable analysis as automatic annotation, a systematic method is presented to synthesize polyphonic music from MIDI files along with verifiable ground truth.

There are several ways to synthesize a musical piece from a MIDI file: mixing monophonic samples according to MIDI `note on` events, rendering MIDI files using sequencers with either sound modules, software instruments, or samplers. The chosen approach is to render MIDI files with samplers for the following reasons: (1) Sequencers and samplers (or Sound Bank players) allow us to render MIDI files with real instrument sound samples into more realistic music. Many efforts have been made to provide large collections of musical instrument sound samples such as McGill University Master Samples , Iowa Musical Instrument Samples , IRCAM Studio On Line and RWC Musical Instrument Sound Database. These sample databases contain a variety of instruments with different playing dynamics and styles for every note in playable frequency ranges, and they are widely used for research. (2) There exists an enormous amount of MIDI files available for personal use or research and there is, therefore, a great potential for expanding the database. Currently, the RWC Musical Instrument Sound Database as well as the Standard MIDI Files (SMF) of RWC Music Database (Goto *et al.*, 2002; Goto, 2003) are selected for synthesis. There are a total of 3544 samples of 50 instruments in RWC-MDB-I-2001 and 315 high quality MIDI Files in RWC-MDB-C-2001-SMF, RWC-MDB-G-2001-SMF, RWC-MDB-J-2001-SMF, RWC-MDB-P-2001-SMF and RWC-MDB-R-2001-SMF. (3) We are free to edit a MIDI file for evaluation purposes to produce several versions from the original MIDI file. For example, limiting the maximal concurrent sources by soloing the designated tracks, changing instrument patches, mixing with or without drums and percussion tracks, etc.

Creating instrument patches

While continuous efforts are being made to manually annotate music scene descriptors for RWC musical pieces (Goto, 2006), no attention is paid to the labeling of RWC Musical Instrument Sound Database RWC-MDB-I-2001. Each sound file in RWC-MDB-I-2001 is a collection of individual notes across the playing range of the instrument and a mute gap was inserted between adjacent notes. The segmentation should not only separate individual notes but also detect onsets for rendering the precise timing of MIDI `note on` events because harmonic sounds are preceded by breathy or noisy regions for certain instrument samples. If the samples are segmented right after the silence gap, they sometimes lead to noticeable delays when triggered by MIDI events to be played by a sampler. These noisy parts in musical instrument sounds come from the sound generation process. When instruments are played with lower dynamics (*pp*), it takes much more time to establish the regimes of oscillation. In order to achieve precise onset rendering, AudioSculpt is used to segment individual notes.

When receiving MIDI event messages, a sampler can render musical instrument samples

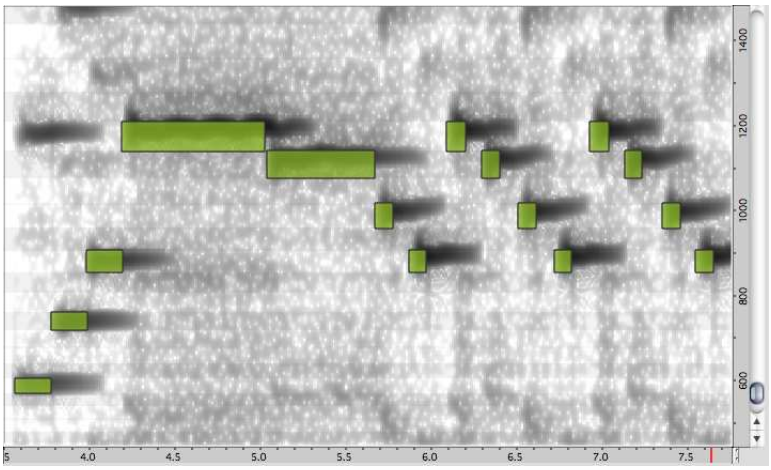


Figure 7.4: Comparison of MIDI notes with the spectrogram of the rendered audio signal

according to the **keymap**s defined in an **instrument patch**. A sample can be assigned to a group of MIDI notes called a **keyzone**. A set of keyzones is called a keymap, which defines the mapping of individual samples to the MIDI notes at specified velocities. For each MIDI note of a keymap, we assign three samples of the same MIDI note number but different dynamics (often labeled as *ff*, *mf* and *pp*). The mapping of the three dynamics to 128 velocity steps is listed in Table 7.2. In this way, an instrument patch includes all the samples of a specific playing style, which results in more dynamics in the rendered audio signals [1].

dynamics	MIDI velocity range
ff	100-127
mf	44-99
pp	0-43

Table 7.2: Mapping the playing dynamics to the MIDI velocity range

Rendering MIDI files into multiple monophonic audio tracks

Once the instrument patches are created, MIDI files can be rendered into polyphonic music by a sequencer+sampler system. Direct rendering of all the tracks into one audio file would prevent the possibility of estimating the ground truth using a single-F0 estimation algorithm. One might then suggest rendering each MIDI track separately. However, this is not a proper solution, not only for polyphonic instrument tracks (piano, guitar, etc.) but also for monodic instrument tracks.

To discuss the issues, one example is illustrated in Figure 7.4. The MIDI notes are extracted from the flute track of *Le Nozze di Figaro* in RWC-MDB-C-2001-SMF. After rendering them by a sequencer+sampler system using the flute samples of RWC-MDB-I-2001, the spectrogram of the rendered audio signal is shown along with the MIDI notes. Each rectangle represents one MIDI

[1] In this work, two playing styles are used: normal and pizzicato.

note, with time boundaries defined by `note on` and `note off`, and with frequency boundaries defined by a quarter tone from its center frequency. It is observed that even if the MIDI note events do not overlap, the rendered signals may still overlap in time as well as in frequency, depending on the **delta time** between the note events and the **release time** parameter of the instrument patch.

In order to access individual sound sources for verifiable analysis, it is necessary to prevent the overlaps of concurrent notes as well as those of consecutive notes. Therefore, each MIDI track is split into tracks of separate notes such that the rendered signals do not overlap. Given the release time setting of an instrument patch, concurrent and consecutive notes in a MIDI track can be split into several tracks under the following condition:

$$T_{\text{note on}}(n) \geq T_{\text{note off}}(n-1) + T_{\text{release}} \tag{7.3}$$

where $T_{\text{note on}}(n)$ is the `note on` time of the current note, $T_{\text{note off}}(n-1)$ is the `note off` time of the previous note, and T_{release} is the release time setting of the instrument patch. In this way, the rendered notes are guaranteed not to overlap one another and individual sound sources can always be referred to whenever necessary. When splitting a MIDI file into several, the control messages [1] are retained in the split tracks such that the rendered notes are exactly the same as the notes rendered from the original MIDI file (see Figure 7.5).

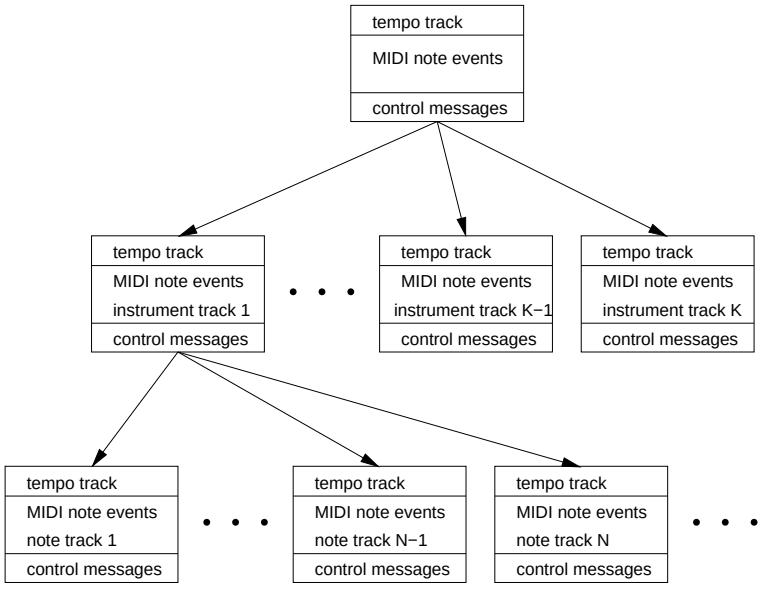


Figure 7.5: Splitting MIDI files into several containing tracks of separate notes

Once MIDI notes are rendered into non-overlapping samples, the ground truth can be established from the analysis of each rendered note sample. The ground truth F0 should be annotated for each analysis frame. Given the MIDI note number, the reference F0 can be calculated as

[1] Channel messages such as pitch bend are included.

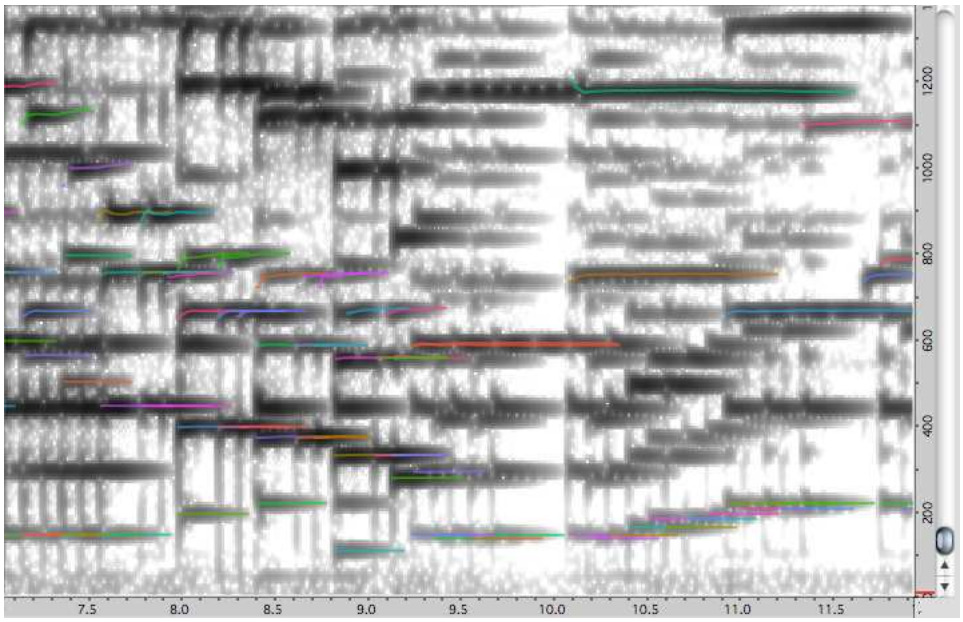


Figure 7.6: Ground truth of multiple F0 tracks in which the F0s of each note signal are estimated by the YIN algorithm.

follows:

$$F_{\text{note}} = \frac{F_{\text{A4}}}{32} \cdot 2^{(\text{MIDI note number}-9)/12} \tag{7.4}$$

It is not always correct to calculate F_{note} with a fixed F_{A4} (for example, $440Hz$) because the tuning frequency F_{A4} may differ; moreover, recorded samples may not be played in tune. As the example illustrated in Figure 7.4, the MIDI notes are placed at center frequencies related to the tuning frequency $F_{\text{A4}} = 440$. The D6 note around 1200 Hz either (1) has a higher tuning frequency, or (2) is not played in tune.

In order to obtain precise F0s as ground truth, F0 estimation is carried out twice for each sample: a coarse search followed by a fine search. The coarse search uses F_{note} with $F_{\text{A4}} = 440$ for a frequency range $F_{\text{note}} \cdot [0.6 \;\; 1.4]$. Then, the search range is limited to one semi-tone, centered at the energy-weighted average of the coarsely estimated F0s. The YIN algorithm is suggested for the estimation of the $F0$ because it has been evaluated to be robust for monophonic signals (de Cheveigné and Kawahara, 2002) and it is available for research use. A window size of $93ms$ is used for the analysis of the reference F0s. For each F0 track of a sample, only the parts of good periodicity serve as ground truth. The aperiodic parts at the transients and near the end of notes are discarded by a threshold of the aperiodicity measure in the YIN algorithm. The estimated F0s of individual note samples collectively establish the ground truth of the synthesized music signal (see Figure 7.6).

7.3 Evaluation

To evaluate the presented multiple-F0 estimation system, the evaluation metrics proposed by Poliner and Ellis (2006) is used, which takes into account the estimation of the number of sources. In order to formulate all the measures of the evaluation metrics, it is necessary to describe some terms beforehand. N_{sys} denotes the estimated polyphony reported by the estimation system; N_{ref} denotes the ground truth polyphony; N_{corr} denotes the number of correctly estimated F0s, N_{miss} denotes the number of *missing F0s*, N_{subs} the denotes the number of *substitution F0s*, and N_{inst} denotes the number of *insertion F0s*. Prior to listing the formulates of the measures, it is simple to investigate the estimation results by two cases (see Table 7.3).

$N_{sys} \geqslant N_{ref}$	$N_{sys} < N_{ref}$
$0 \leqslant N_{corr} \leqslant N_{ref}$	$0 \leqslant N_{corr} \leqslant N_{sys}$
$N_{miss} = 0$	$N_{miss} = N_{ref} - N_{sys}$
$N_{subs} = N_{ref} - N_{corr}$	$N_{subs} = N_{sys} - N_{corr}$
$N_{inst} = N_{sys} - N_{ref}$	$N_{inst} = 0$

Table 7.3: Error measures for two cases

N_{corr} is what often called **True Positives**, which is bounded by $\min(N_{sys}, N_{ref})$. N_{miss} is what often called **False Negatives**, which is non-zero only if $N_{sys} < N_{ref}$. The rest of the errors are often called **False Positives**, which include both N_{subs} and N_{inst}. N_{inst} is non-zero only if $N_{sys} > N_{ref}$. By summarizing the two cases, the measures of the evaluation metrics can be formulated, for T analysis frames, as follows:

1. Total error

$$E_{tot} = \frac{\sum_{t=1}^{T} \max(N_{ref}(t), N_{sys}(t)) - N_{corr}(t)}{\sum_{t=1}^{T} N_{ref}(t)} \tag{7.5}$$

2. Missing error

$$E_{miss} = \frac{\sum_{t=1}^{T} \max(0, N_{ref}(t) - N_{sys}(t))}{\sum_{t=1}^{T} N_{ref}(t)} \tag{7.6}$$

3. Substitution error

$$E_{subs} = \frac{\sum_{t=1}^{T} \min(N_{ref}(t), N_{sys}(t)) - N_{corr}(t)}{\sum_{t=1}^{T} N_{ref}(t)} \tag{7.7}$$

4. Insertion error

$$E_{inst} = \frac{\sum_{t=1}^{T} \max(0, N_{sys}(t) - N_{ref}(t))}{\sum_{t=1}^{T} N_{ref}(t)} \tag{7.8}$$

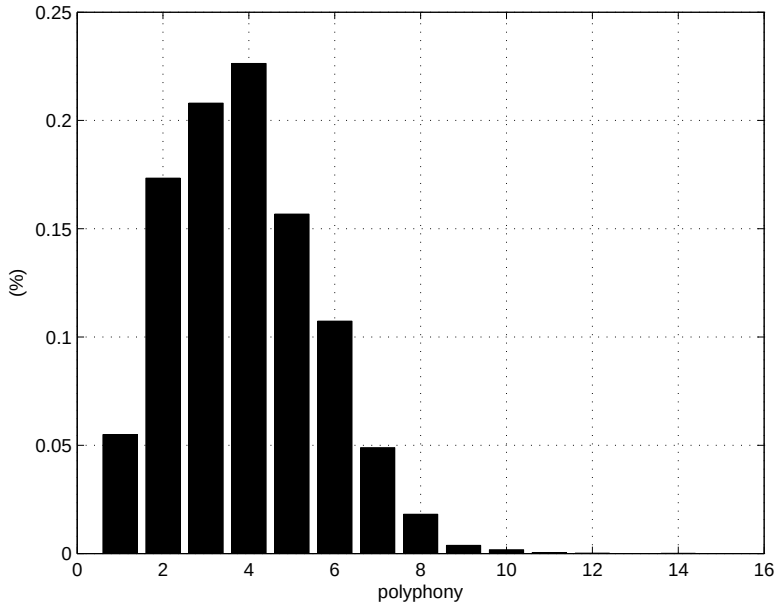


Figure 7.7: Polyphony distribution of the synthesized music database

5. Recall

$$Rcl = \frac{N_{corr}}{N_{ref}} \tag{7.9}$$

6. Precision

$$Prs = \frac{N_{corr}}{N_{sys}} \tag{7.10}$$

7. Overall Accuracy

$$Acc = \frac{N_{corr}}{N_{corr} + N_{miss} + N_{subs} + N_{inst}} \tag{7.11}$$

The proposed multiple-F0 estimation system is evaluated by two polyphonic databases containing (1) sources with equal energy and (2) sources with different energy. In the first case, the polyphonic samples mixed from monophonic sources with equal energy is used (see Section 7.2.2). The results of the estimated polyphony, the accuracy rates and the error rates are shown in Figure 7.8, 7.10, and 7.11, respectively, for the polyphony up to 5. In the second case, 26 pieces of synthesized music (see Section 7.2.3) are used. The results of the estimated polyphony, the accuracy rates and the error rates are shown in Figure 7.9, 7.12, and 7.13, respectively. Although the number of concurrent notes may increase to more than 10 (see the polyphony distribution in Figure 7.7), the results of polyphony up to 6 are the main concerns. In both cases, concurrent sources related to the same note are regarded as one single source. An example of the estimation F0s of a piece of synthesized music is demonstrated in Figure 7.15.

When the sources are of equal energy, the estimated polyphony has a tendency to overestimation for lower polyphony ($M < 5$). This is probably due to the energy normalization process while mixing monophonic samples into polyphonic ones, which boosts the spurious spectral components and causes additional F0 hypotheses to be estimated as extra sources. In addition, ΔMBW_{model} is learned as the *average* of various spectral envelopes. This threshold is probably too low for the artificially mixed polyphonic samples in which the stationary parts right after the onsets are selected and the spectral envelopes are usually less smooth due to several strong partials boosted by the resonances. When the energy of the sources varies, the estimated polyphony has a distinct peak at the correct polyphony for the low polyphony but less precise for the polyphony higher than four ($M > 4$). The tendency to overestimation in the first case is alleviated in the second case (compare Figure 7.8 and Figure 7.9). The two thresholds, E_{noise} and ΔMBW_{model}, seem to work well for the polyphony up to four. However, for the polyphony higher than four, the performance starts to decline. This could be due to the weak sources that are not detectable by either the explained energy or the improvement of spectral smoothness, which causes quite a few missing F0s. When a weak source of HRF0 is totally buried in a strong source of the related F0, it is the most difficult to extract the "buried" HRF0.

To further investigate the performance of the proposed system, we evaluate the accuracy rate with the related error rates. In the first case, the overall accuracy is 72.77%. In the second case, the overall accuracy is 64.75%. The decline in accuracy due to the varying energy is about 10%. The recall rate is also about 10% lower compared to that in the first case. There are several issues related to the performance degradation when the energy of sources varies. In addition to the issues related to E_{noise} and ΔMBW_{model}, the polyphony inference algorithm needs to be improved because no care is taken to remove an incorrectly inferred F0 that is reasonably added in the inference process. The polyphony inference algorithm shall adaptively remove a less probable F0 after several iterations if the resulting combination does support it with high probability.

The performance of the proposed multiple-F0 estimator can also be characterized by the likelihood function $p(M|\hat{M})$ where $\hat{M}$ is the estimated polyphony (see Figure 7.14). That is, given the estimated polyphony $\hat{M}$, it is possible to reason about how likely the true polyphony M can be.

MIREX 2007

The proposed system has participated in MIREX (Music Information Retrieval Evaluation eXchange) 2007 for *Multiple Fundamental Frequency Estimation* which is the first ever evaluation of many existing multiple-F0 estimation algorithms. Since the proposed system does not involve the tracking of F0s across frames, it participated in the first task: frame-by-frame evaluation. The evaluation database contains woodwind quintet (French horn, clarinet, bassoon, flute and oboe) recordings and synthesized music using RWC music instrument samples (clarinet, violin, cello, electric bass, electric guitar and saxophone) and RWC MIDI files. Four pieces of music are selected to generate the testing corpora by mixing two to five solo tracks, which sums up to 28 pieces. The ground truth of the woodwind quintet recordings is annotated by YIN for each

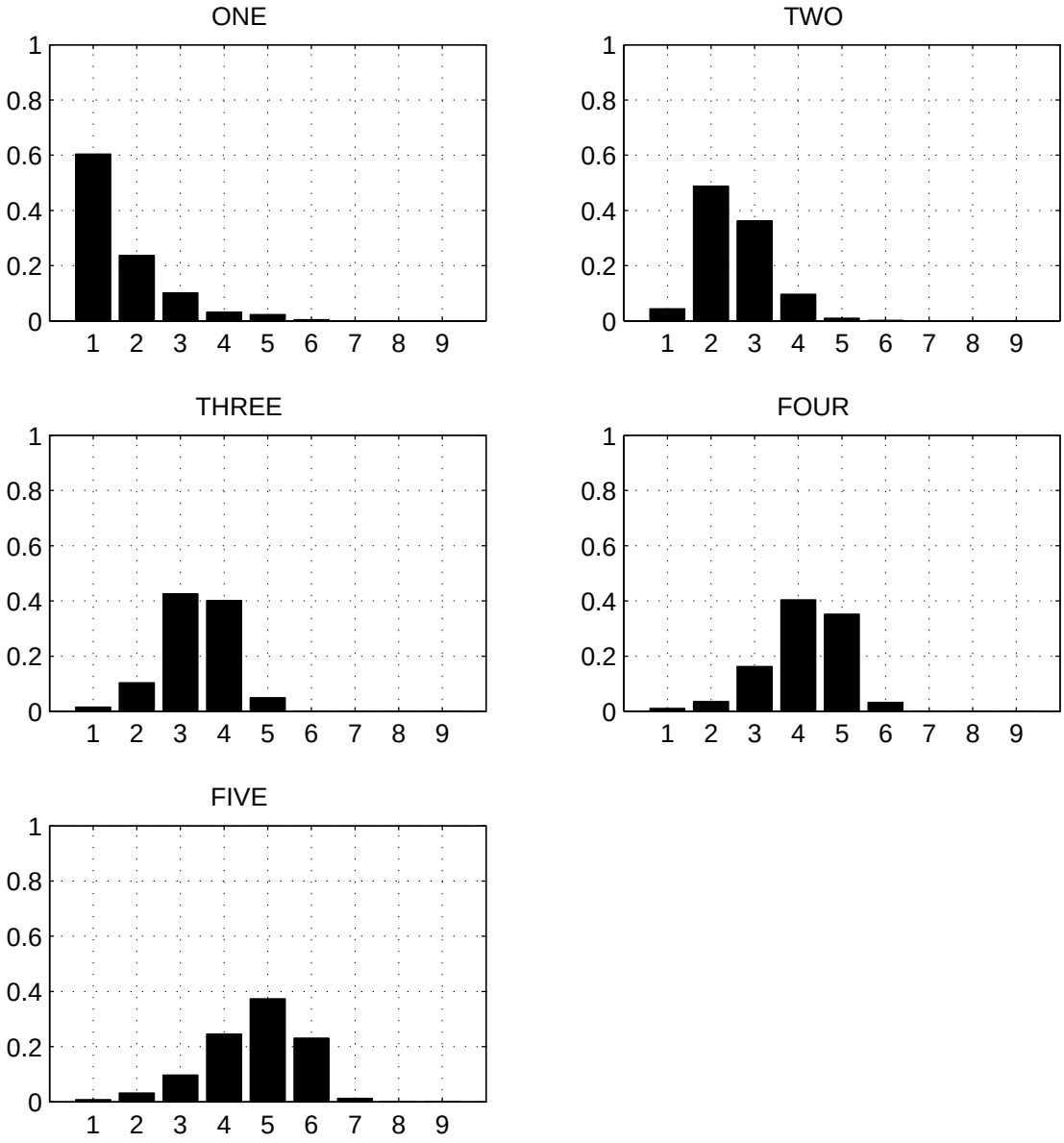


Figure 7.8: Estimated polyphony for sources with equal energy

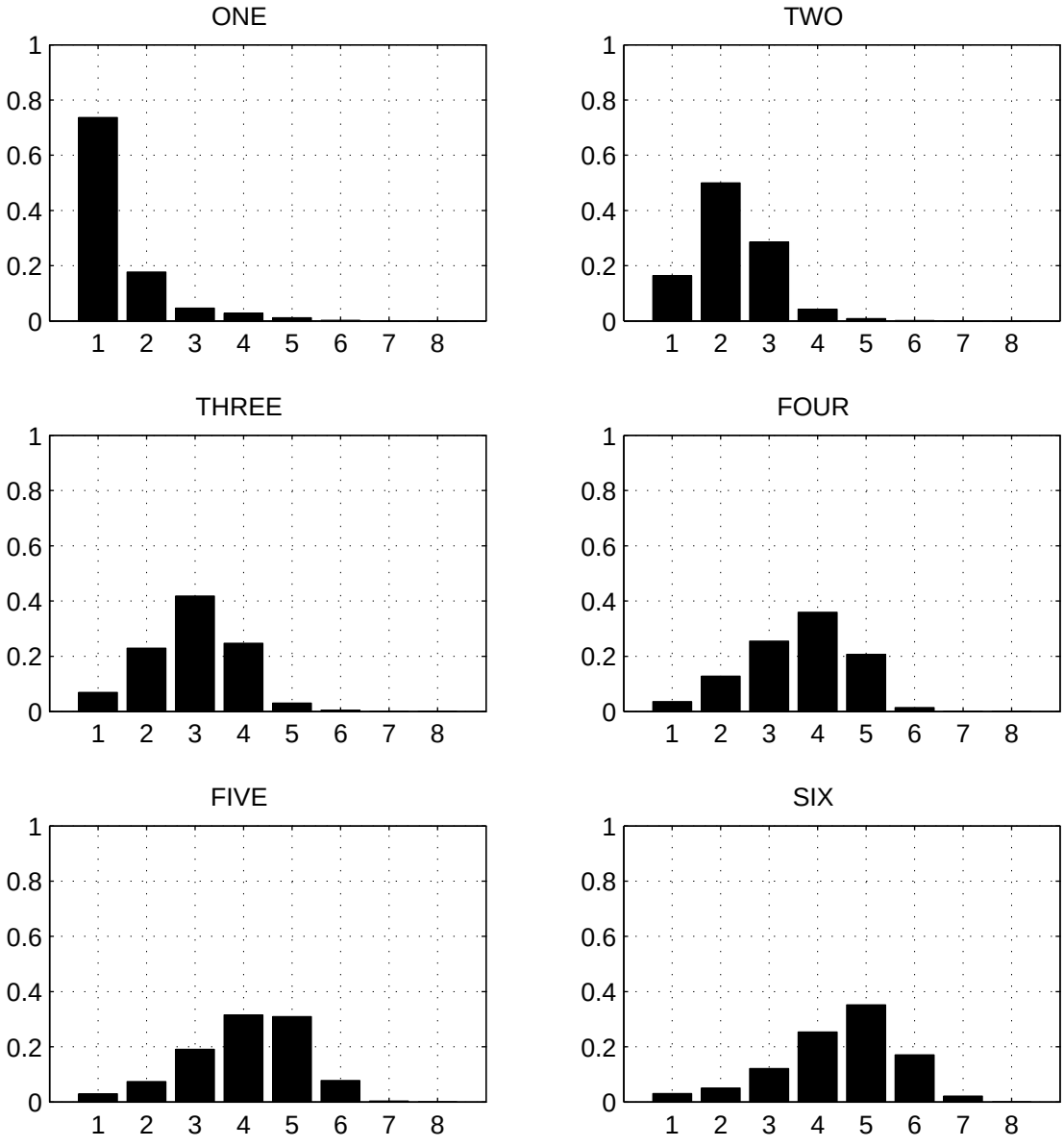


Figure 7.9: Estimated polyphony for sources with different energy.

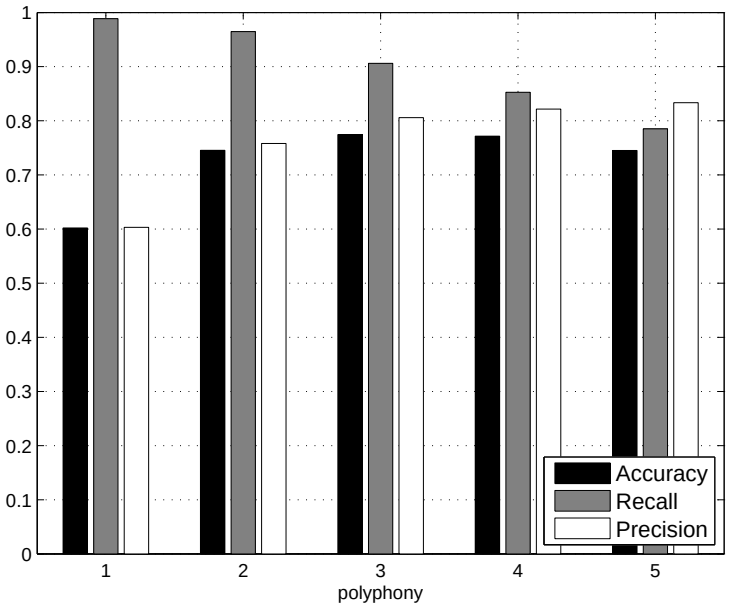


Figure 7.10: Accuracy, recall and precision results for sources with equal energy.

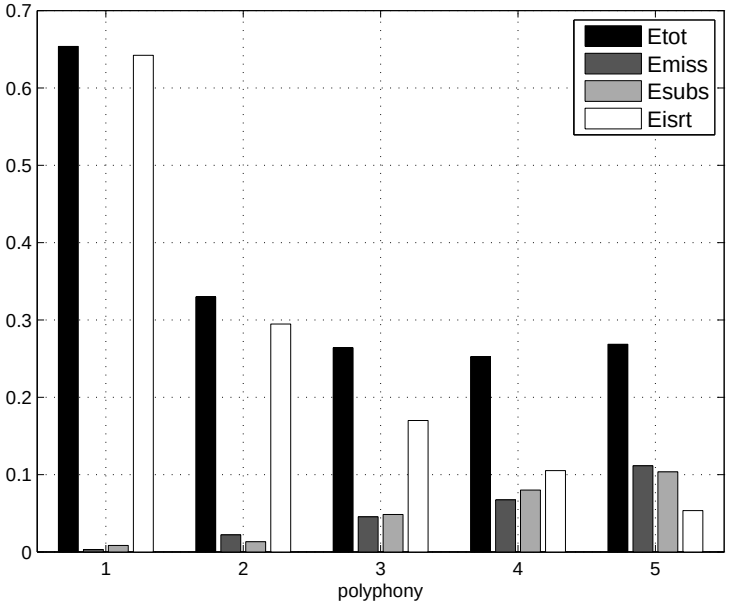


Figure 7.11: Error results for sources with equal energy.

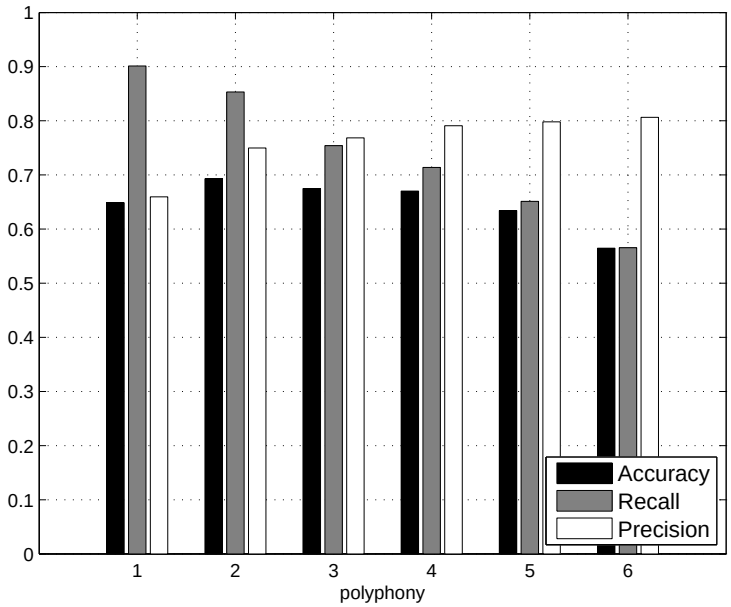


Figure 7.12: Accuracy, recall and precision results for sources with different energy.

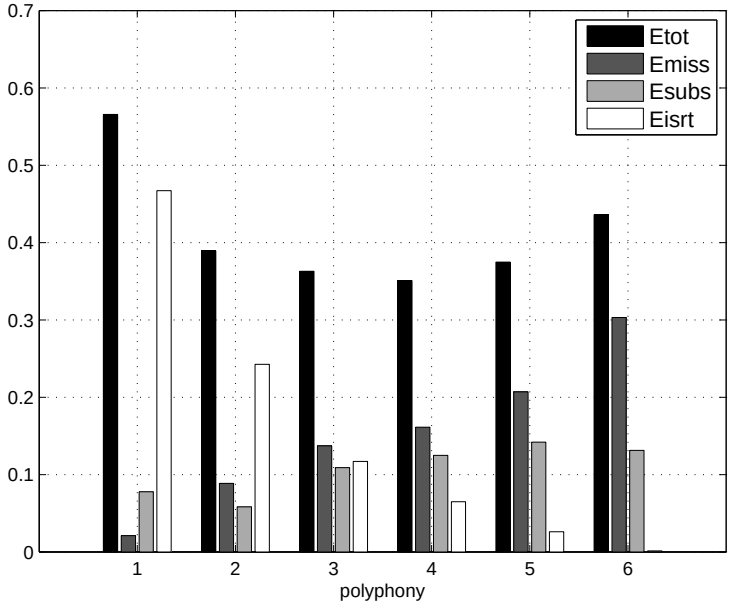


Figure 7.13: Error results for sources with different energy.

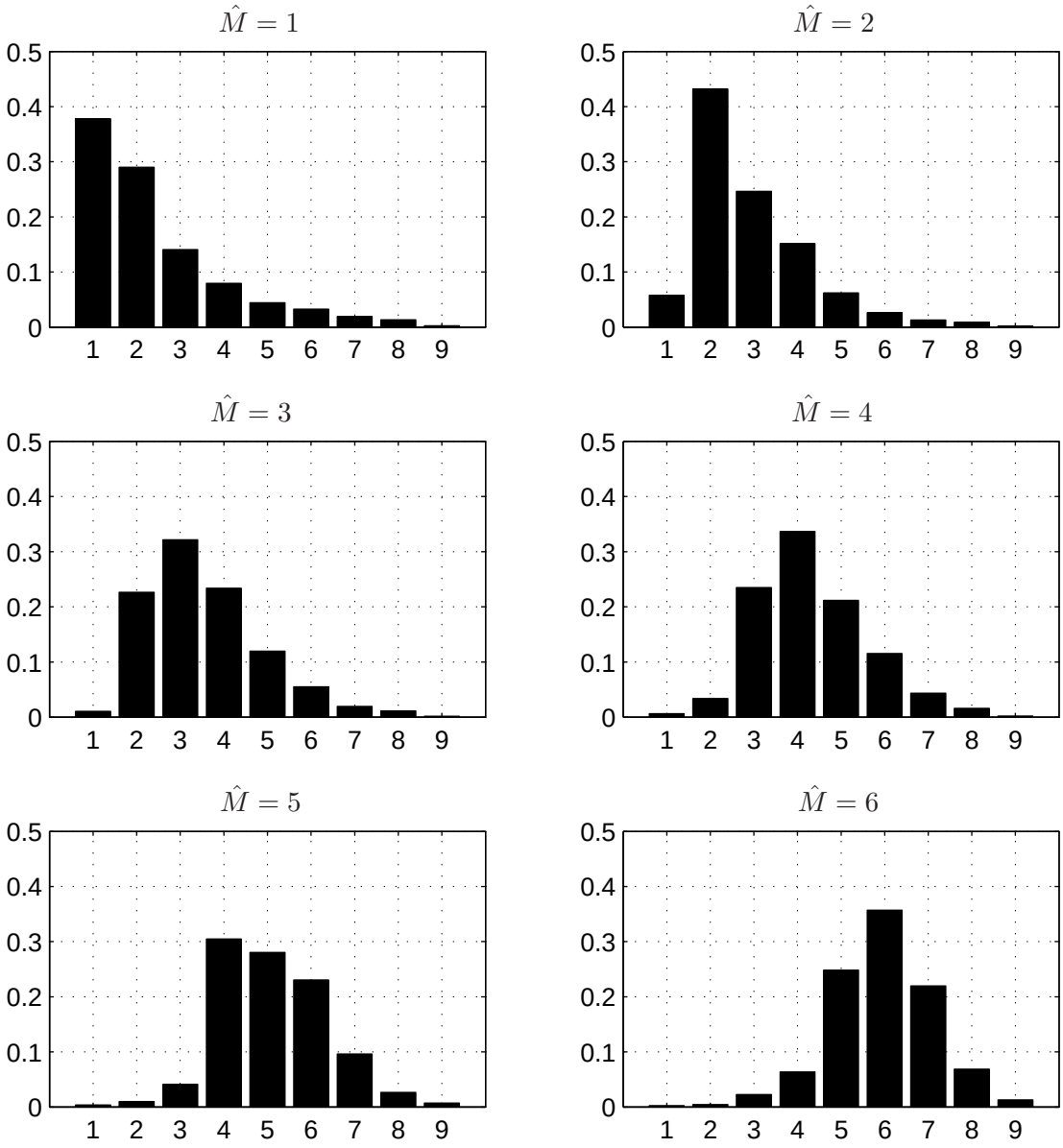


Figure 7.14: The likelihood of the estimated polyphony. Given the estimated polyphony $\hat{M}$, the likelihoods of the true polyphony M from 1 to 9 are illustrated.

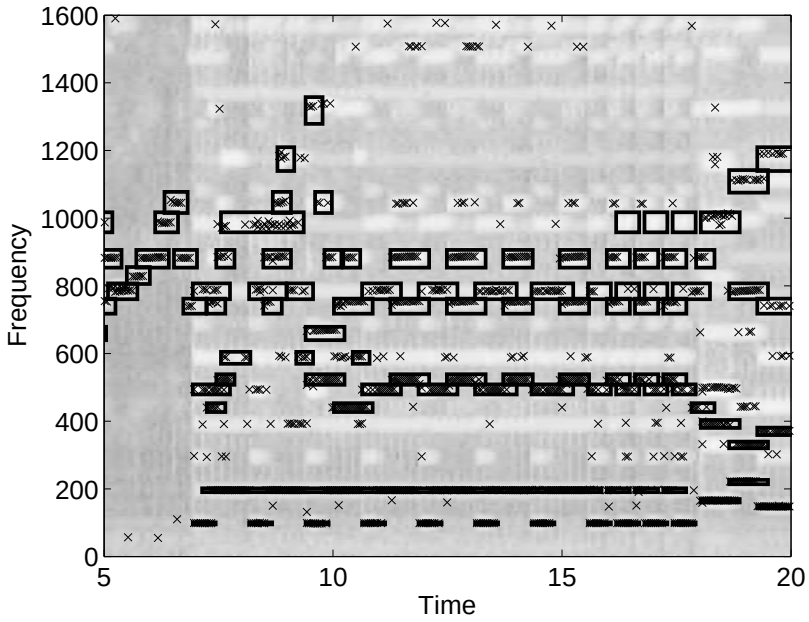


Figure 7.15: An example of the estimated F0s of a piece of synthesized music. The crosses are the estimated F0s, and the rectangle shows the time-frequency boundaries of the ground truth notes.

solo recording track. The ground truth of the synthesized music is also annotated by YIN for each separately rendered notes. An evaluated system is required to report the estimated F0s every $10ms$. There are 12 participants who have submitted 16 systems in total. The participants, their abbreviations and the proposed methods are listed in Table 7.4. To facilitate the comparison, the ranking of the overall accuracy is shown in Figure 7.16, and the overall accuracy for different number of instruments is shown in Figure 7.17. Notice that the number of instruments does not stand for the true polyphony but is still representative of the complexity of the testing pieces. The proposed multiple-F0 estimation system (Yeh, 2007) has been ranked in the second place. The submitted system, called the MIREX version, has been tuned to favor lower polyphony, achieving an average accuracy at 58.9%. Large degradation in the five-instrument mixtures has been found to be related to the *yet-to-improve* implementation of the polyphony inference algorithm.

In order to compare the MIREX version with the thesis version, the MIREX version is evaluated by the same synthesized music database used for the evaluation of the thesis version. The average accuracy rate of the thesis version is about 8% better than that of the MIREX version (see Figure 7.18). Above all, the thesis version improves significantly the accuracy in the estimation for the polyphony higher than 3.

team ID	team members	method
AC	A. Cont	NMF with sparsity constraint
CC	C. Cao, M. Li, J. Liu and Y. Yan	subharmonic sum + harmonic structure tracking
CY	C. Yeh	feature combination of F0 candidates
EV	E. Vincent, N. Bertin and R. Badeau	NMF with harmonicity constraint
KE	K. Egashira, H. Kameoka and S. Sagayama	Harmonic Temporal Structured Clustering
PE	G. Poliner and D. P. W. Ellis	Support Vector Machine
PI	A. Pertusa and J. M. Inesta	feature combination of F0 candidates
PL	P. Leveau	sparse decomposition with instrument models
RK	M. Ryynänen and A. Klapuri	auditory model + HMM tracking
SR	S. A. Raczynski, N. Ono and S. Sagayama	NMF with harmonicity constraint
VE	V. Emiya, R. Badeau and B. David	Maximum likelihood based on High Resolution analysis
ZR	JR. Zhou and J. D. Reiss	Resonator Time-Frequency Image + error pruning

Table 7.4: The participants of MIREX'07 Multiple-F0 estimation and their proposed methods.

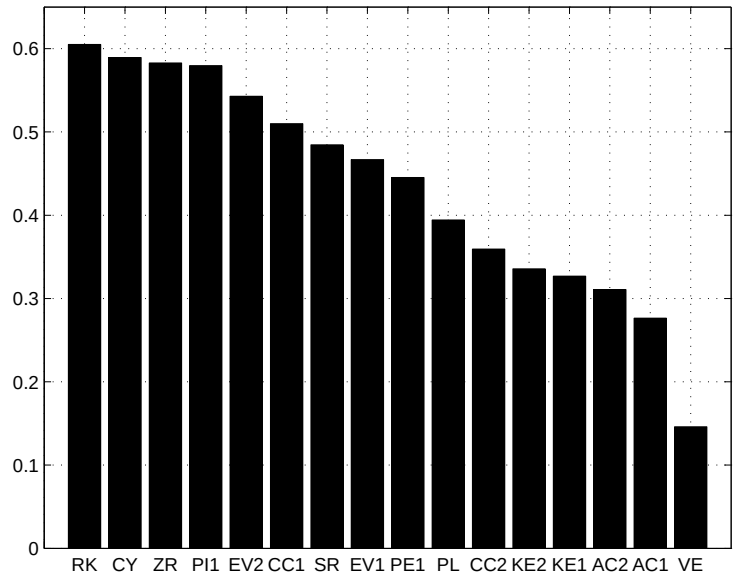


Figure 7.16: MIREX'07 result: the accuracy ranking of the evaluated systems. The proposed system CY is ranked in the second place.

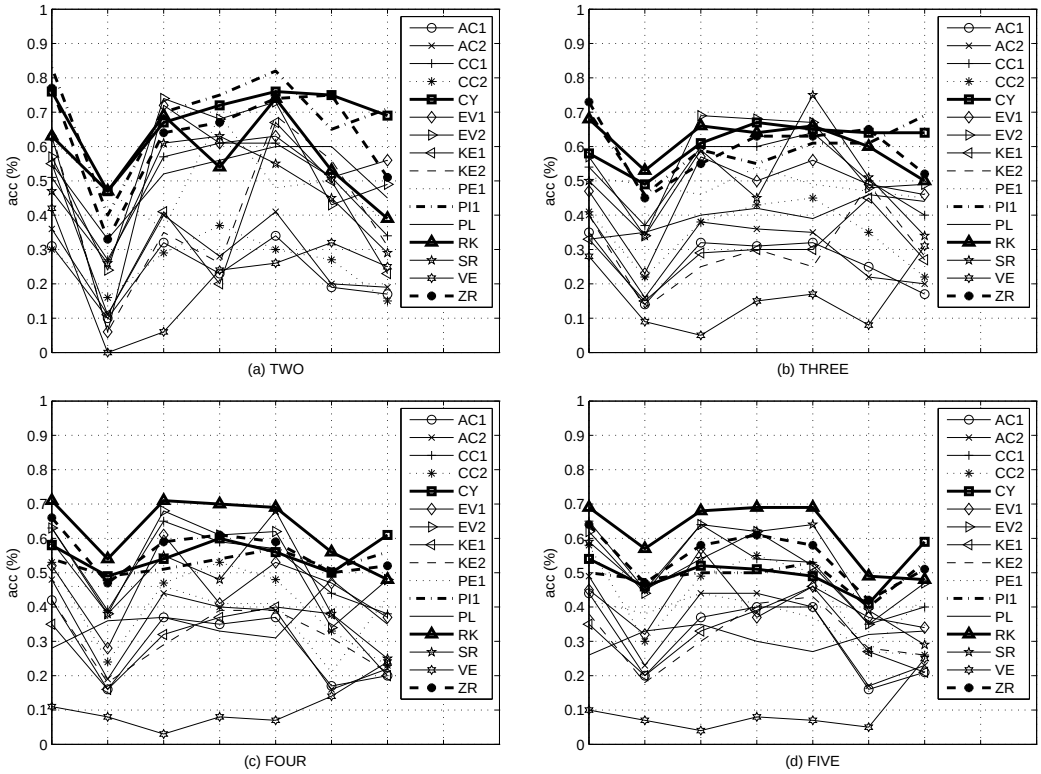


Figure 7.17: MIREX'07 result: the accuracy rates for each number of mixed instrument tracks.

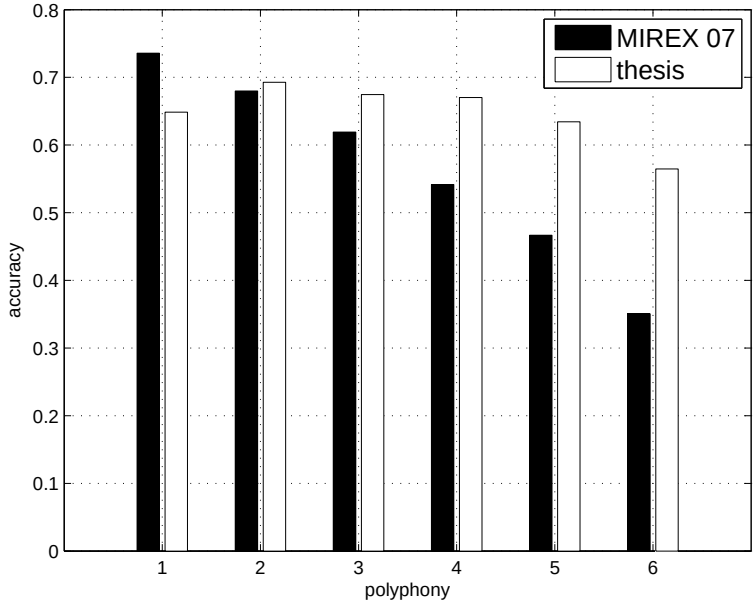


Figure 7.18: Comparison of the accuracy rates between the MIREX version and the thesis version.

7.4 Multiple F0 Tracking in Solo Recordings of Monodic Instruments

The algorithms developed so far have been focused on the frame-based analyses. In order to build continuous F0 trajectories, one often makes use of the information across the frames. Following the HMM proposed by Tokuda *et al.* (1999), Wu *et al.* (2003) models the F0 state space as a union state space Ω consisting of all the hypotheses of the concurrent number of sources

$$\Omega = \Omega_0 \cup \Omega_1 \cup \Omega_2 \cup \cdots \Omega_M \tag{7.12}$$

where Ω_m represents the state with m sources and M is the maximal polyphony. The concept of this tracking model is to evaluate all hypothetical combinations of $\{S_m\}_{m=1}^{M}$ and to search for the optimal sequence of the state space throughout the observation spaces. For the proposed multiple-F0 estimation system, the observations can be represented by the probabilities generated by all the hypothetical combinations. Although Wu *et al.* (2003) applied this model to the special case in which $M = 2$, this tracking mechanism is complicated for a general case. Despite the limitation of the performance for the polyphony higher than six, the presented system can be applied to a special case of multiple-F0 tracking: monodic instrument solo recordings.

Single-F0 estimators are often used to analyze the solo recordings of monodic instruments, assuming that there is only one F0 present in the observed signal. Since the recordings are often done in a reverberant environment, the use of single-F0 estimation algorithms is not appropriate because reverberation prolongs the note duration, resulting in a polyphonic signal. In consequence, a single-F0 estimator may tend to favor a subharmonic which explains both the current note and the reverberation of the preceding notes. Baskind and de Cheveigné (2003) applied a double-F0 estimator (an extension of YIN) to the task of F0 tracking for monodic instrument recordings. Significant improvement in robustness of F0 estimation of reverberant sounds has been reported, which encourages a multiple-F0 tracking approach to this problem. On the assumption that there is one monodic instrument playing, the observed signal can be modeled as a predominant harmonic source plus the reverberant parts of the preceding notes and the background noise. Accordingly, it is proposed to first decode the predominant-F0 track from a set of hypothetical combinations, and then to use the less-dominant F0 hypotheses to elicit the continuity of the notes in the predominant-F0 track.

The proposed F0 tracking system is composed of three parts (see Figure 7.19). In each analysis frame, multiple-F0 estimation provides the list of the top-ranked combinations of F0 hypotheses. F0 tracking can thus be considered the decoding of the optimal path through the trellis structure which consists of the hypothetical combinations across the frames (see Figure 7.20). Because it is difficult to define the transition probability between two hypothetical combinations that are of different polyphony hypotheses, it is proposed to decode first the predominant-F0 track based on the probability of individual F0 hypothesis. Then, the secondary F0s, which are assumed to be the reverberant parts, can be tracked by extending the notes in the predominant-F0 tracks. In this case, polyphony inference (see Section 7.1) is implicitly carried out by the tracking

mechanism.

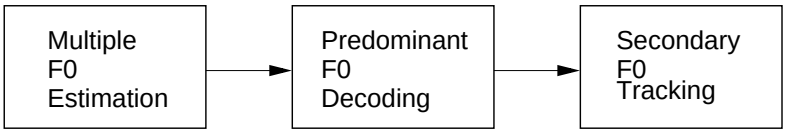


Figure 7.19: Overview of the F0 tracking system for monodic instrument solo recordings.

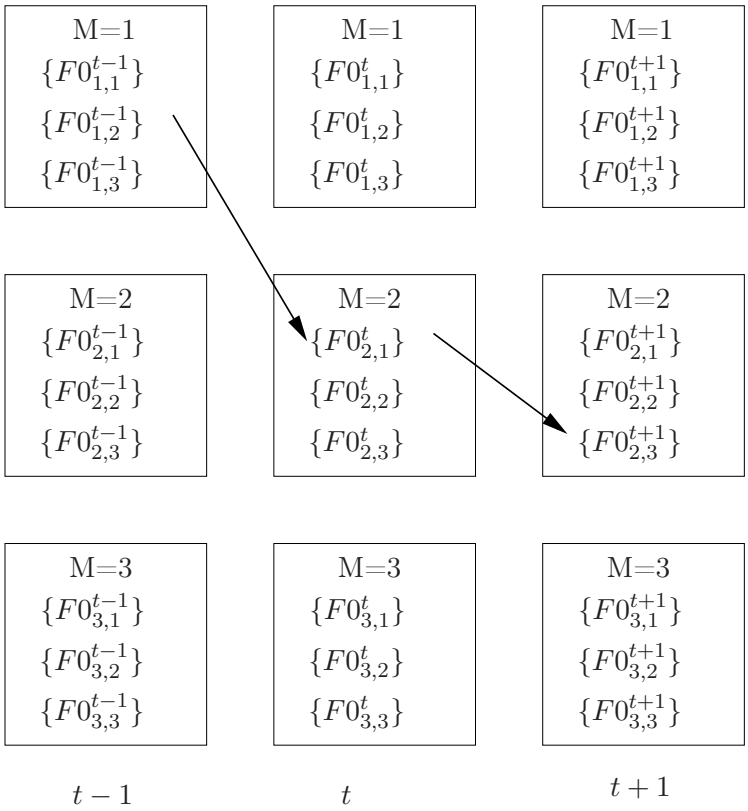


Figure 7.20: Decoding the optimal multiple-F0 path in the trellis structure of hypothetical combinations. Each hypothetical combination is denoted as $\{F0^{i}_{m,c}\}$ (where m ranges from 1 to the maximal polyphony) for the cth top-ranked candidate combination at frame i.

Predominant-F0 tracking

For solo recordings of monodic instruments, the predominant F0s are related to the monophonic melody line being played. As long as the reverberation of preceding notes is less dominant than the current note, taking the most significant F0 as the predominant F0 is generally accepted. Given the individual probabilities of F0 hypotheses (see eq.(7.2)) as observations, the best state sequence of predominant F0s is inferred by a two-stage tracking method:

1. **Forward connection between frames**
   In the first stage, the connection is made between the F0 hypotheses in two consecutive frames. For each F0 hypothesis, the frequency difference allowed for the connection is one

half tone. For every F0 hypothesis, the connection that produces the highest product of individual probabilities is kept for the next stage.

2. **Track construction**
   An F0 track can be defined by the locally connected F0s. However, there are often several "holes" in-between tracks, where F0 hypotheses should be filled in to establish a correct track. These holes represent the missing F0 hypotheses, which are often observed when the onset of one note disturbs the quasi-stationary parts of the other notes. This issue is addressed by linear prediction of F0 tracks, a technique similar to Lagrange *et al.* (2004)'s method. To reconstruct a broken track, a backward/forward search based on linear prediction is applied to each pair of adjacent tracks that are close in time and in frequency.

Secondary-F0 tracking

Once the predominant-F0 track is decoded, the secondary F0s can be tracked by extending the notes in the predominant F0 track. To track the reverberant parts of the predominant-F0 tracks, the combinations containing the current predominant F0 and the preceding predominant F0s are used to establish the secondary F0 tracks. As long as the effective salience of a secondary F0 is larger than a pre-defined threshold, it is considered a part of the reverberation.

Testing examples

To demonstrate the proposed tracking algorithm, two solo recordings are tested: bassoon and violin. For the bassoon solos, the proposed method is compared with the single-F0 estimator YIN. YIN produces subharmonic errors in the frames where the reverberant parts of the preceding notes are strong in energy (see Figure 7.21). This demonstrates the difficulty of F0 tracking for monodic solo recordings, which can be barely handled by a single-F0 estimator. In the second example, a violin solo, the proposed method yields promising estimates for the fast arpeggios of which the reverberant parts are well tracked, too. The challenge is to track successive notes playing in harmonic relations, especially in an octave relation. When the reverberation of the preceding note is still strong and the current note is, for instance, one octave above, the current note is mixed with the reverberation tail of the preceding note, which causes octave ambiguity.

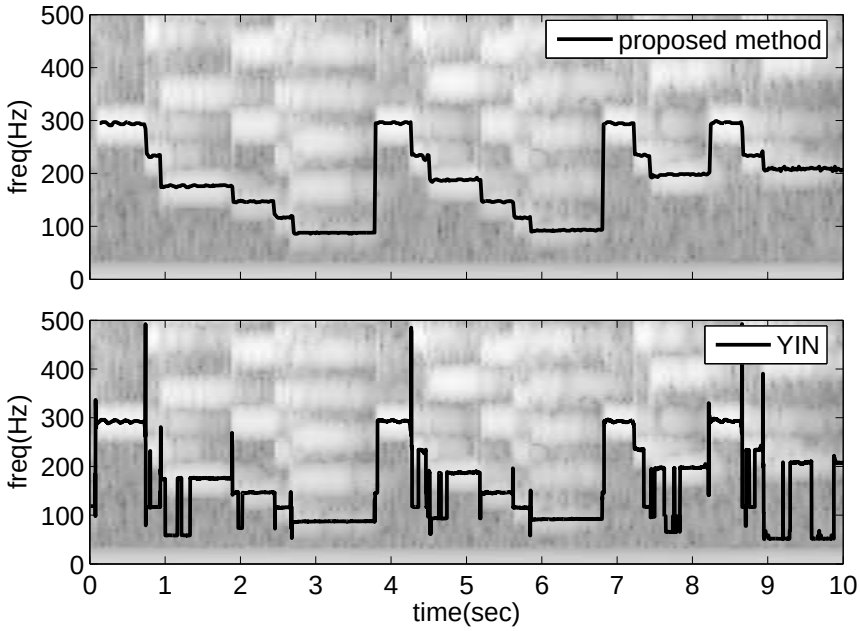


Figure 7.21: Comparison of the proposed predominant-F0 estimator and YIN, using a Mozart's bassoon solo recording for the test.

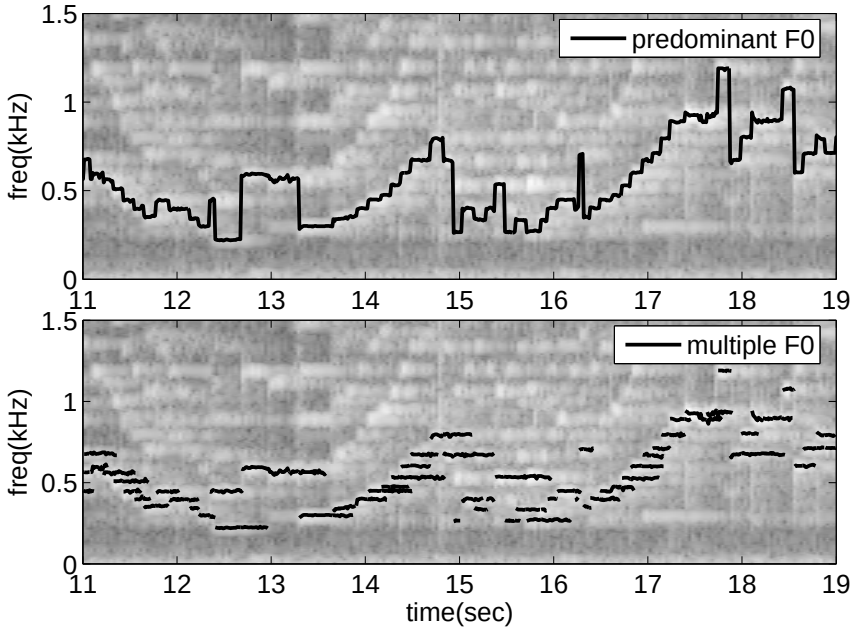


Figure 7.22: Multiple-F0 tracking test of a Bach's violin solo recording.

8

CONCLUSIONS AND PERSPECTIVES

8.1 Conclusions

A frame-based multiple-F0 estimation system has been presented to analyze music sound signals. The approach is based on joint evaluation of F0 hypotheses, following three guiding principles related to the physical properties of harmonic instrument sounds: harmonicity, spectral smoothness and synchronicity. The algorithms concerning noise estimation, joint evaluation of F0 hypotheses, and polyphony inference have been developed, which cope with the three fundamental models in the problem of multiple-F0 estimation: the noise model, the source model and the source interaction model. The proposed algorithms are designed to target the extraction of either NHRF0s (non-harmonically related F0s) or HRF0s (harmonically related F0s). The main contributions made in this thesis can be summarized with respect to four aspects: noise estimation, overlapping partial treatment, formulation of the guiding principles and, database construction.

Contrary to the usual assumptions that noise is white Gaussian, it is proposed to model the noise magnitude distribution by a succession of Rayleigh distributions, each of which is a function of frequency. An adaptive noise level estimation algorithm has been developed which neither includes additional information from the neighboring frames or pure noise segments, nor makes use of harmonic analysis. The bias of the estimated noise level are around 5%; the variance of the estimated noise level are around $25\% - 30\%$. The estimated noise level provides a probabilistic threshold for the classification of spectral peaks into sinusoids and noise. It is considered the most important feature for extracting the correct number of NHRF0s.

The development of the joint estimation algorithm has focused on the treatment of overlapping partials. The proposed overlap treatment reallocates the overlapping partials to one or several sources such that the ambiguity in the related HPS (hypothetical partial sequences) are removed. It is believed to be the key treatment for a reliable evaluation of hypothetical sources.

Based on the three guiding principles of harmonic instrument sounds, four score criteria have been proposed: harmonicity, mean bandwidth, spectral centroid and synchronicity. The synchronicity criterion has been derived from the single-frame observation, which is different from the usual technique which makes use of the information across frames. The four criteria are linearly combined to yield a score function, which ranks all possible combinations among F0 candidates. In the case in which the number of F0s is known, the performance of the joint estimation algorithm is encouraging and competitive to the existing methods. The efficiency concern has been discussed by the evaluation of three methods for F0 candidate selection. The method that iteratively extracts first NHRF0s and then the HRF0s has reduced more than one thousand times of combinations to calculate, compared to the method that simply uses a threshold of the harmonicity criterion. Nevertheless, the accuracy of multiple-F0 estimation is brought down by only $1-2\%$.

One essential problem in the research of multiple-F0 estimation, often not underlined in the previous studies, is that there is no polyphonic music database (corpus+ground truth) available. Therefore, a systematic method has been proposed for the creation of a polyphonic music database. The idea is to make use of the large numbers of existing MIDI files and music instrument sound samples to render synthesized polyphonic music. Care has been taken to split MIDI tracks to ensure that separate notes do not overlap after rendering. In this way, ground truth can be more reliably established by a single-F0 estimator. The proposed methodology is reproducible, extensible and interchangeable. Most importantly, the ground truth is verifiable.

A polyphony inference algorithm has been developed, based on the noise estimation algorithm and the joint estimation algorithm. The maximal polyphony is first estimated, and a consolidation process is then carried out which makes use of two criteria: the explained energy and the improvement in spectral smoothness. Two thresholds related to the criteria are learned from music instrument sound samples to avoid spurious detections. NHRF0s which explain less energy than the noise energy are not considered valid; HRF0s overly smoothing the envelopes are not considered valid. Evaluations have proved the proposed system to be one of the best among the existing systems. The average accuracy rate is about 65%.

8.2 Perspectives

The proposed F0 estimation system can be improved in two aspects: (1) simplification of the frame-based system and (2) inclusion of a tracking mechanism as post-processing. The joint evaluation part and the polyphony inference part could be combined in an **iterative combination/consolidation** manner. Given a list of F0 candidates, one may iteratively evaluate the validity of an added F0 hypothesis in a hypothetical combination. To develop an efficient and robust algorithm, a strategy to, for each iteration, replace less likely F0 hypotheses with more

probable ones is necessary.

The other way to improve the frame-based estimation system is to incorporate a tracking mechanism. F0 tracking in a joint manner is complicated. Hence, it might be simpler to establish the probable F0 trajectories first and then prune the spurious sources afterwards. With the individual probability of each F0 hypothesis and the frequency proximity between F0 hypotheses, it is possible to collect individual F0 hypotheses into **candidate trajectories**. Three issues need to be addressed to yield the final tracking result: (1) connection of separate trajectories (belonging to one source stream) due to missing F0s (2) verification of octave trajectories and (3) refinement of the onset/offset positions.

In this thesis, the impact of concurrent percussive instrument sounds has not been studied. Mallet percussion instruments evoke distinct pitch but they have special partial structures which require specific spectral models to deal with. The less-pitched percussion instruments like drums introduce strong transient components. When drum sounds overlap with concurrent harmonic sounds, the partials of the harmonic sounds are highly disturbed, which causes quite a few estimation errors for a frame-based analysis system. One way to analyze polyphonic music with drums is to perform drum identification and separation before multiple-F0 estimation is performed. If the interference of drum sounds can not be effectively attenuated, the inference of **ghost** F0 hypotheses is a possible solution to reconstruct the "broken" trajectories disturbed by drum sounds.

The finds of this study have a lot of potential for various applications. The major contributions of the study include at least the following: First, the estimated noise level is representative of the instantaneous noise spectrum, which can serve as a new feature for signal analysis. By smoothing the varying noise levels across frames, the noise spectrogram can be estimated, from which a **relative spectrogram** can be derived. The relative spectrogram can serve as a new representation of the signal with enhanced sinusoidal components. Noise level estimation can also be applied to transient detection because the noise level usually rises when strong attacks occur. By combining noise level estimation with harmonic analysis, a robust onset detection could be achieved. Similar techniques can also be applied to the voicing determination of speech signals. The unvoiced part of speech contains turbulence and is thus noise-like. Therefore, it should be detected when the HNR (harmonic-to-noise ratio) is low.

Second, the study of overlapping partials is helpful to polyphonic signal analysis and source separation. The proposed overlap treatment uses the interpolated amplitudes to estimate the overlapping partial amplitudes, which appears to work properly for the current task. An extensive study of the overlap model can initiate a more refined method for the inference of overlapping partial amplitudes. This can be useful for, for instance, a more precise separation of concurrent sources.

Third, the proposed method for creating polyphonic music database can be extended to other MIR (Music Information Retrieval) tasks, such as melody extraction, key estimation, beat tracking, tempo extraction, drum detection, chord detection, onset detection, score alignment, and source separation. Although the proposed method uses synthesized music, it can be easily integrated into multi-track recordings of monodic instruments and singing voices. As for the

ground truth of the beats and the tempos, they can be programmed in MIDI files to achieve a realistic groove. By analyzing the MIDI notes, the ground truth of chords can be extracted. This method is particularly valuable for the MIR tasks that require timing precision.

Most of the research on multiple-F0 estimation aims at using it as the core component within an automatic music transcription system which integrates low-level analyses into a high-level representation as a musical score. The development of an automatic music transcription system involves a lot more research topics: key estimation, tempo/meter estimation, instrument recognition and musicological model. In addition to automatic music transcription, a multiple-F0 estimation algorithm has many more practical applications, in the context of polyphonic signals, such as musical instrument recognition, chord estimation and source separation.

A

The Magnitude Distribution of White Gaussian Noise

Consider a narrow band noise process $n(t) = r(t)e^{j\phi(t)} = x(t)+jy(t)$. Assuming that the real part $x(t) = r(t)\cos(\phi(t))$ and the imaginary part $y(t) = r(t)\sin(\phi(t))$ are (1) Gaussian distributed with zero mean and variance σ, and (2) independent, the joint probability density function is

$$p(x, y) = p(x) \cdot p(y) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}$$

Using the change of variable $x^2 + y^2 = r^2$, $p(x, y)$ can be expressed by

$$p(r, \phi) = \frac{r}{2\pi\sigma^2} e^{-r^2/2\sigma^2}$$

The marginal distributions can thus be derived:

$$p(\phi) = \int_0^\infty p(r, \phi)dr = \frac{1}{2\pi}, \quad \text{for } \phi \in [0, 2\pi]$$
$$p(r) = \int_0^{2\pi} p(r, \phi)d\phi = \frac{r}{\sigma^2} e^{-r^2/2\sigma^2}, \quad 0 \le r < \infty$$

where the phase ϕ is uniformly distributed and the magnitude $r(t)$ is Rayleigh distributed.

B

Spectral Descriptors for Sinusoid/Non-Sinusoid Classification

Röbel and Zivanovic (2004) proposed four spectral peak descriptors to classify spectral peaks. The descriptors are designed to deal with non-stationary sinusoids. In this thesis, three descriptors, **Normalized Bandwidth Descriptor** (NBD), **Duration Descriptor** (DD) and **Frequency Coherence Descriptor** (FCD) are selected to classify sinusoidal/non-sinusoidal peaks (see Table B.1). NBD and DD of a spectral peak are derived from bandwidth and duration of a signal (Cohen, 1995). FCD is based on the reassignment operators (Auger and Flandrin, 1995). The thresholds of the descriptors can be adaptively determined (Zivanovic *et al.*, 2007).

Normalized Bandwidth

Consider a signal $s(t)$ with the related spectrum $S(\omega)$. The spectral content of a signal can be characterized by its mean frequency and its bandwidth. The **mean frequency** indicates the frequency at which the signal energy is concentrated:

$$\bar{\omega} = \frac{\int \omega \cdot |S(\omega)|^2 d\omega}{\int |S(\omega)|^2 d\omega} \tag{B.1}$$

descriptor	formula
normalized bandwidth	$\text{NBD} = \frac{\sum_k (k-\bar{\omega})^2 \lvert X(k)\rvert^2}{L \sum_k \lvert X(k)\rvert^2}$ where $\bar{\omega} = \frac{\sum_k k \lvert X(k)\rvert^2}{\sum_k \lvert X(k)\rvert^2}$
duration	$\text{DRD} = \sqrt{\frac{\lvert X(k)\rvert^2}{\sum_k \lvert X(k)\rvert^2} \sum_k [A'(k)^2 + (g_d(k) - \bar{t})^2]}$ where $g_d(k)$ and $A'(k)$ are the real and imaginary part of $\frac{X_t(k)X^*(k)}{\lvert X(k)\rvert^2}$ and $\bar{t} = \frac{\sum_k g_d(k) \lvert X(k)\rvert^2}{\sum_k \lvert X(k)\rvert^2}$
frequency coherence	$\text{FCD} = \underset{k}{\text{argmin}} \frac{X_d(k)X^*(k)}{\lvert X(k)\rvert^2}$ for $k \in$ all bins in one peak

Table B.1: Spectral descriptors for the spectral peak classification. The summation of k is with respect to all frequency bins within a spectral peak.

The bandwidth B gives an idea of how the energy of the signal is distributed around its mean frequency

$$B = \sqrt{\frac{\int (\omega - \bar{\omega})^2 \cdot |S(\omega)|^2 d\omega}{\int |S(\omega)|^2 d\omega}} \tag{B.2}$$

which is calculated in terms of the square root of the *second central moment of energy density*. The normalized bandwidth descriptor NBD is derived for a spectral peak from calculating the bandwidth around a single peak and normalizing it with the spectral width L [1] of the peak. NBD is meant to evaluate the energy concentration about the mean frequency.

Duration

The **mean time** indicates a signal's *center of gravity* in the time domain:

$$\bar{t} = \frac{\int t \cdot |s(t)|^2 dt}{\int |s(t)|^2 dt} \tag{B.3}$$

where $|s(t)|^2$ denotes the instantaneous power. The duration of a signal measures whether the energy density concentrates around its mean time:

$$T^2 = \frac{\int (t - \bar{t})^2 |s(t)|^2 dt}{\int |s(t)|^2 dt} \tag{B.4}$$

[1] L is defined between two neighboring local minima around the peak.

In order to calculate the duration in the frequency domain, it is necessary to express $\bar{t}$ in terms of $S(\omega)$.

$$\begin{aligned}
\bar{t} &= \frac{1}{\int |S(\omega)|^2 d\omega} \int\int\int t S^*(\omega) e^{j(\omega'-\omega)t} S(\omega') d\omega d\omega' dt \\
&= -\frac{1}{j\int |S(\omega)|^2 d\omega} \int\int\int S^*(\omega) \frac{\partial}{\partial\omega} e^{j(\omega'-\omega)t} S(\omega') dt d\omega' d\omega \\
&= -\frac{2\pi}{j\int |S(\omega)|^2 d\omega} \int\int S^*(\omega) \frac{\partial}{\partial\omega} \delta(\omega'-\omega) S(\omega') d\omega' d\omega \\
&= -\frac{2\pi}{j\int |S(\omega)|^2 d\omega} \int S^*(\omega) \frac{\partial}{\partial\omega} \frac{1}{2\pi} S(\omega) d\omega \\
&= \frac{1}{\int |S(\omega)|^2 d\omega} \int S^*(\omega) (-\frac{1}{j}\frac{d}{d\omega}) S(\omega) d\omega
\end{aligned} \tag{B.5}$$

Denoting $S(\omega) = A(\omega)e^{j\phi(\omega)}$, the mean time can be further expressed in terms of $S(\omega)$

$$\begin{aligned}
\bar{t} &= \frac{1}{\int |S(\omega)|^2 d\omega} \int S^*(\omega) j(A'(\omega)e^{j\phi} + j\phi'(\omega)S(\omega)) d\omega \\
&= \frac{1}{\int |S(\omega)|^2 d\omega} \int [-\phi'(\omega) + j\frac{A'(\omega)}{A(\omega)}] |S(\omega)|^2 d\omega \\
&= -\frac{1}{\int |S(\omega)|^2 d\omega} \int \phi'(\omega) |S(\omega)|^2 d\omega
\end{aligned} \tag{B.6}$$

and the duration can thus be expressed as

$$\begin{aligned}
T &= \frac{1}{\int |S(\omega)|^2 d\omega} \int S^*(\omega) (-\frac{1}{j}\frac{d}{d\omega} - \bar{t})^2 S(\omega) d\omega \\
&= \frac{1}{\int |S(\omega)|^2 d\omega} \int |(\frac{1}{j}\frac{d}{d\omega} + \bar{t}) S(\omega)|^2 d\omega \\
&= \frac{1}{\int |S(\omega)|^2 d\omega} \int |\frac{1}{j}\frac{A'(\omega)}{A(\omega)} + \phi'(\omega) + \bar{t}|^2 A^2(\omega) d\omega \\
&= \frac{1}{\int |S(\omega)|^2 d\omega} \int (\frac{A'(\omega)}{A(\omega)})^2 A^2(\omega) d\omega + \int (\phi'(\omega) + \bar{t})^2 A^2(\omega) d\omega \\
&= \frac{1}{\int |S(\omega)|^2 d\omega} \int (\frac{A'(\omega)}{A(\omega)})^2 A^2(\omega) d\omega + \int (g_d(\omega) - \bar{t})^2 A^2(\omega) d\omega
\end{aligned} \tag{B.7}$$

where $g_d(\omega) = -\phi'(\omega)$ is the **group delay**. The duration descriptor DD makes use of the above equation to define the duration of a spectral peak as shown in Table B.1.

Frequency Coherence

The frequency coherence descriptor FCD is derived from the frequency reassignment operators. The reassignment can be interpreted as the estimation of the instantaneous frequency (frequency reassignment) and the group delay (time reassignment) for each bin in the time-frequency plane. The reassignment spectrogram incorporates phase information which is not included in the traditional spectrogram and indicates more precisely the center of gravity in the time-frequency

plane.

An efficient computation of the reassignment operators has been proposed by Auger and Flandrin (1995), which is based on the STFT. A later review points out the relationship between the *phase-derived* reassignment operators and the *amplitude-derived* reassignment operators (Hainsworth and Macleod, 2003). Consider a STFT $X(\omega,t) = A(\omega,t)e^{j\phi(\omega,t)}$, the reassignment operators are related to the partial derivative of phase w.r.t. time (instantaneous frequency) and frequency (negative group delay):

$$\hat{\omega}(\omega,t) = \frac{\partial}{\partial t}\phi(\omega,t) = \omega + \Im\Big\{\frac{X_d(\omega,t)}{X(\omega,t)}\Big\} = \omega + \Im\Big\{\frac{X_d(\omega,t)X^*(\omega,t)}{|X(\omega,t)|^2}\Big\} \tag{B.8}$$

$$\hat{t}(\omega,t) = \frac{\partial}{\partial \omega}\phi(\omega,t) = t - \Re\Big\{\frac{X_t(\omega,t)}{X(\omega,t)}\Big\} = t - \Re\Big\{\frac{X_t(\omega,t)X^*(\omega,t)}{|X(\omega,t)|^2}\Big\} \tag{B.9}$$

where $X_d(\omega,t)$ is the STFT using the derivative of the window and $X_t(\omega,t)$ is the STFT using a time ramped version of the window. The frequency offset between a bin of DFT (Discrete Fourier Transform) and its instantaneous frequency (reassigned frequency), i. e., $\hat{\omega} - \omega$, is used to measure the frequency coherence of a spectral peak. Frequency coherence descriptor FCD is defined as the minimal frequency offset among all bins within a spectral peak (see Table B.1). By taking the derivative of amplitude w.r.t. time and frequency, the following equations can be obtained.

$$\frac{\partial}{\partial t}A(\omega,t) = \Re\Big\{\frac{X_d(\omega,t)}{X(\omega,t)}\Big\} \tag{B.10}$$

$$\frac{\partial}{\partial \omega}A(\omega,t) = \Im\Big\{\frac{X_t(\omega,t)}{X(\omega,t)}\Big\} \tag{B.11}$$

Notice that the duration descriptor involves the calculation of the derivative of amplitude $A(\omega,t)$ w.r.t. frequency ω (see eq.(B.7)).

C

Sinusoidal Parameter Estimation

The observed quasi-periodic signals, usually embedded in noise, can be modeled as a sum of several sinusoids plus noise. Short-time sinusoidal models can be either stationary or non-stationary. Sinusoidal parameter estimation can be achieved by either parametric methods or non-parametric methods. The accuracy of estimators is usually evaluated by comparing the estimation error variance to **Cramér-Rao Bounds** (CRB). CRB is a theoretical lower bound on the variance of an unbiased estimator. The closer the estimated error is to CRB, the more accurate the estimator is. The estimator's bias can be observed when the performance of the estimator is no longer improved under sufficiently high SNR conditions.

C.1 Short-time stationary sinusoids

A short-time stationary sinusoid model assumes stationarity in a short-time analysis window and thus parameterizes a sinusoid with constant amplitude, frequency and phase. The Maximum Likelihood Estimation (MLE) method assumes the noise to be independent additive white Gaussian noise (AWGN). For a single sinusoid under high SNR condition, the MLE of the frequency is found by choosing the frequency at which the spectral magnitude attains its maximum (Rife and Borstyn, 1974), i. e., the maximal spectral peak. In the DFT domain, peak-picking provides a good initial parameter estimation and the parameter refinement can be applied afterwards, such as the quadratic interpolation (Serra, 1989).

When there are more than one sinusoid involved, the difficulty arises for the MLE method due to the presence of cross-product terms between sinusoids. As long as the frequencies of

neighboring sinusoids are sufficiently separate, the cross-product terms can be neglected, thereby permitting an approximate MLE. In this way, the spectral peaks remain as good initial parameter estimates (Rife and Borstyn, 1976). However, parameter estimation for closely spaced sinusoids require more sophisticated analysis techniques, such as the **High Resolution** (HR) method (Badeau, 2005).

C.2 Short-time non-stationary sinusoids

In general, a non-stationary sinusoid can be expressed as

$$s[n] = a[n]e^{j\theta[n]} \quad \text{for } n = 0, \ldots, L-1 \tag{C.1}$$

where the amplitude $a[n]$ and the frequency $\theta'[n]$ vary in the short-time analysis frame of length L.

Frequency modulated sinusoids

A sinusoid with constant amplitude, $a[n] = A_0$, and linearly changing instantaneous frequency is called **chirp**. Accordingly, the phase is quadratic $\theta[n] = \beta_0 n^2 + \omega_0 n + \phi_0$ where ω_0 is the mean frequency, $2\beta_0$ is the frequency slope and ϕ_0 is the initial phase. The MLE methods for a single chirp (Djuric and Kay, 1990) and multiple chirps (Saha and Kay., 2001) have been derived. Non-parametric methods based on time-frequency analysis such as reassignment methods (Auger and Flandrin, 1995) can also serve to estimate the frequency slope (Röbel, 2002). Master and Liu (2003) proposed an analytic solution to chirp parameters based on DFT and Fresnel analysis, but a correction model is required for small frequency slopes.

Amplitude and frequency modulated sinusoids

Two models of sinusoids have been applied to the HR methods: (1) sinusoids with exponentially damped amplitudes (Kumaresan and Tufts, 1982) or (2) sinusoids with polynomial amplitudes (Badeau, 2005). In both cases, the frequencies of sinusoids are assumed constant. However, the problem becomes complicated when both amplitude modulation and frequency modulation (AM/FM) are involved in the model. For a single AM chirp, explicit expressions of estimates might differ due to different assumptions on the models (Friedlander and Francos, 1995; Zhou *et al.*, 1996; Ghogho *et al.*, 1999; Besson *et al.*, 1999). Compared to the parametric methods, the recently developed non-parametric methods have shown accurate results with better computational efficiencies (Abe and Smith, 2005; Wells and Murphy, 2006).

C.3 Selected methods for noise level estimation

In the proposed noise level estimation algorithm (see Chapter 4), efficient subtraction of sinusoids is important for a better noise level approximation, which requires a good sinusoidal parame-

ter estimator. For music sound signals, AM (*tremolo*) and FM (*vibrato*) are often observed. Therefore, a non-stationary model is more appropriate to generalize music signals. The method proposed by Abe and Smith (2005) is selected because of its efficiency and flexibility. Considering an exponential AM and a linear FM, the amplitude and the phase of a sinusoid are expressed as

$$a_{m,h}[n] = e^{\lambda_{m,h}+\alpha_{m,h}n} \tag{C.2}$$

$$\theta_{m,h}[n] = \beta_{m,h}n^2 + \omega_{m,h}n + \phi_{m,h} \tag{C.3}$$

where $e^{\lambda_{m,h}}$ is the instantaneous amplitude at the reference time index (usually at the center of the window), $\alpha_{m,h}$ is the AM rate, $\beta_{m,h}$ is half the FM rate (frequency slope), $\omega_{m,h}$ is the instantaneous frequency at the reference time index and $\phi_{m,h}$ is the initial phase.

The method of Abe and Smith (2005) is based on the quadratically interpolated FFT (QIFFT) estimator with bias correction. The QIFFT estimator models the logarithmic amplitude and the phase by means of parabolic functions. It is exact for the Gaussian window (Peeters, 2001). Due to AM and FM, bias is introduced to the estimates of amplitude, frequency and phase. Moreover, there exist biases to be compensated for other types of windows. Abe and Smith (2005) proposed to first estimate the parameters based on Gaussian window, and then adapt the estimates to different types of windows. The correction coefficients are numerically determined by multiple regression analysis.

Consider a damped chirp process $y[n] = Ae^{-\alpha n}e^{j(\beta n^2+\omega n+\phi)} + w[n]$ where $w[n]$ is white Gaussian process with variance σ_w^2, the CRBs are (Zhou *et al.*, 1996)

$$\mathrm{CRB}(\hat{A}) = \frac{\sigma_w^2}{2}\frac{\epsilon_2}{\epsilon_0\epsilon_2 - \epsilon_1^2} \tag{C.4}$$

$$\mathrm{CRB}(\hat{\alpha}) = \frac{\sigma_w^2}{2A^2}\frac{\epsilon_0}{\epsilon_0\epsilon_2 - \epsilon_1^2} \tag{C.5}$$

$$\mathrm{CRB}(\hat{\phi}) = \frac{\sigma_w^2}{2A^2}\frac{\epsilon_2\epsilon_4 - \epsilon_3^2}{D} \tag{C.6}$$

$$\mathrm{CRB}(\hat{\omega}) = \frac{\sigma_w^2}{2A^2}\frac{\epsilon_0\epsilon_4 - \epsilon_2^2}{L^2 D} \tag{C.7}$$

$$\mathrm{CRB}(\hat{\beta}) = \frac{\sigma_w^2}{2A^2}\frac{\epsilon_0\epsilon_2 - \epsilon_1^2}{L^4 D} \tag{C.8}$$

$$\tag{C.9}$$

in which

$$\epsilon_k = \sum_{n=0}^{L-1}(\frac{n}{N})^k e^{-2\alpha n/L} \tag{C.10}$$

$$D = \epsilon_0\epsilon_2\epsilon_4 - \epsilon_1\epsilon_4^2 - \epsilon_0\epsilon_3^2 + 2\epsilon_1\epsilon_2\epsilon_3 - \epsilon_2^3 \tag{C.11}$$

The estimator of Abe and Smith (2005) is tested for exponentially damped chirps with random parameters that are distributed uniformly in the ranges specified in Table C.1. Since the amplitude modulation parameter α is a variable in CRBs, the influences of α on the estimator

parameter	A	α	ϕ	$\omega/2\pi$	$\beta/2\pi$
range	0.5	$(0:0.05:0.3)/L$	$[-\pi \quad \pi]$	$[0.01 \quad 0.3]$	$[-0.5/L^2 \quad 0.5/L^2]$

Table C.1: Sinusoidal parameter distribution range. The parameter α is selected between 0 and $0.3/L$, with an increment of $0.05/L$. The parameters of ϕ, ω and β are uniformly distributed in the indicated range.

performance is first tested. Seven values of α are chosen in this test. For each α, one thousand AM/FM sinusoids are generated. The analysis window is Blackman. The results are shown in Figure C.1. It is observed that the frequency, the frequency slope and the phase estimators have consistent performance at different amplitude modulation rates. To compare the estimator performance using different cosine windows, the amplitude modulation rate is fixed at 0.3. The errors are plotted along with the CRBs in Figure C.2. In both tests, the residual energy compared to the additive noise energy is demonstrated, too.

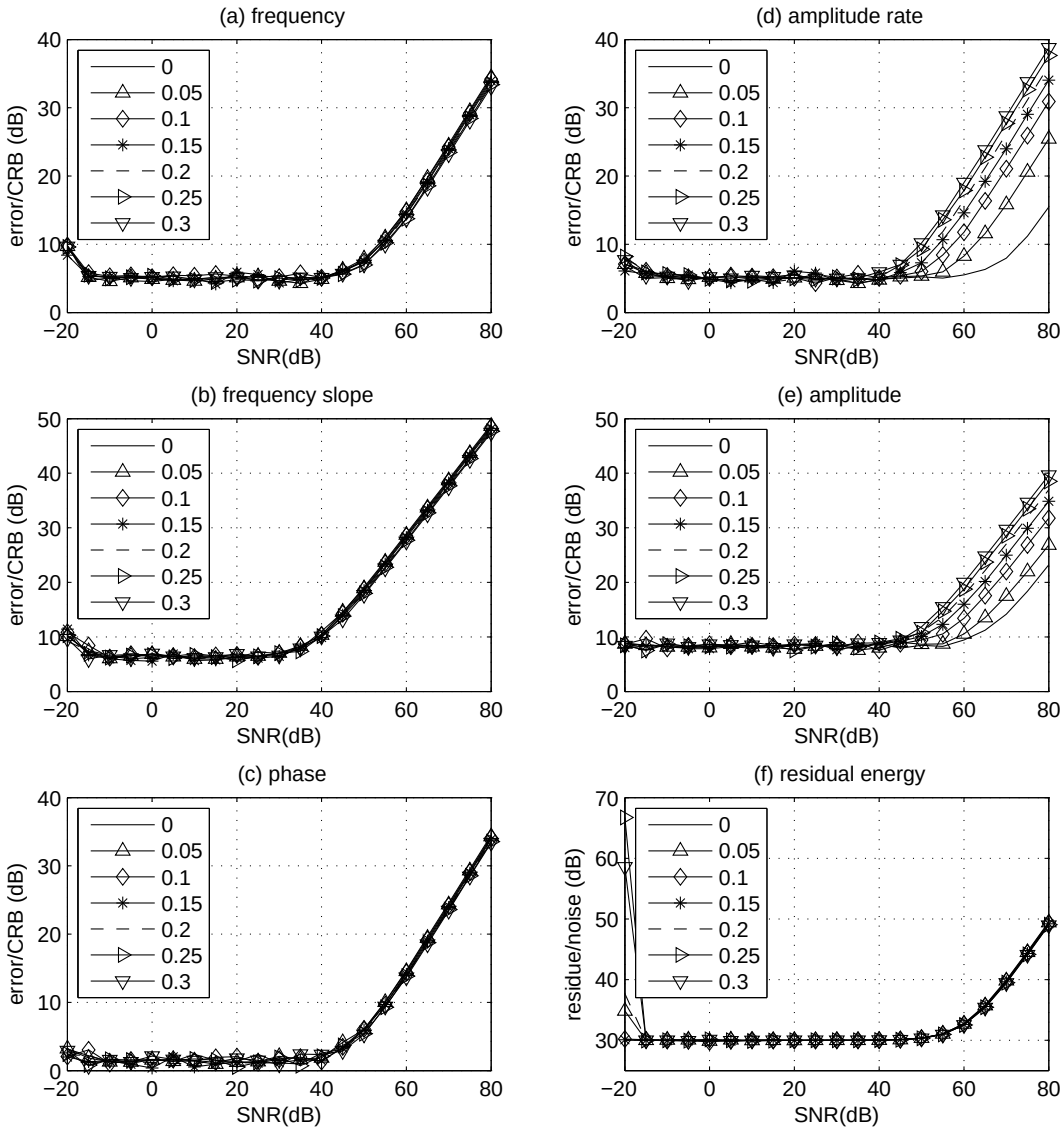


Figure C.1: (a)-(e): Abe&Smith estimator errors w.r.t. CRBs; (f): residual energy w.r.t. additive noise energy.

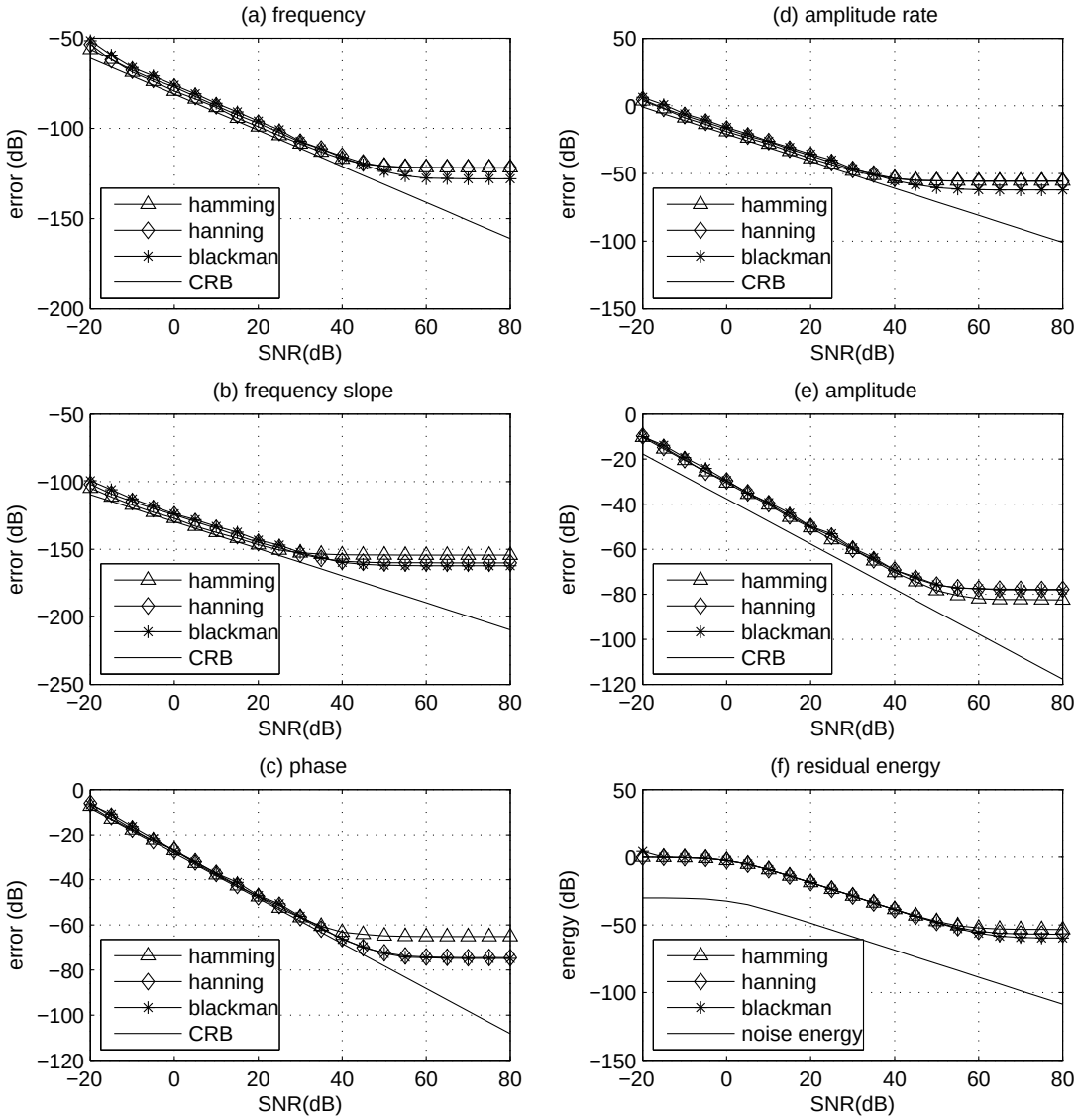


Figure C.2: Abe&Smith estimator using different cosine windows. (a)-(e): estimator errors and CRBs; (f): residual energy and additive noise energy.

D

The Expected Amplitude of Overlapping Partials

One of the major problems with multiple-F0 estimation is the handling of the overlapping partials. To facilitate the study of this problem, it is often assumed that the partials are sinusoids overlapping at the same frequency but of different phases. Considering K sinusoids overlap, the resulting sinusoid can be written as (Smith, 2007)

$$A\cos(\omega t+\phi) = A_1\cos(\omega t+\phi_1) + A_2\cos(\omega t+\phi_2) + \cdots + A_K\cos(\omega t+\phi_K) \tag{D.1}$$

where each sinusoid has amplitude A_k, phase ϕ_k and frequency ω. Expand the above equation using trigonometric identity

$$[A\cos(\phi)]\cos(\omega t) - [A\sin(\phi)]\sin(\omega t) = \Big[\sum_{k=1}^{K} A_k\cos(\phi_k)\Big]\cos(\omega t) - \Big[\sum_{k=1}^{K} A_k\sin(\phi_k)\Big]\sin(\omega t)$$

from which

$$A = \sqrt{\Big[\sum_{k=1}^{K} A_k\cos(\phi_k)\Big]^2 + \Big[\sum_{k=1}^{K} A_k\sin(\phi_k)\Big]^2} \tag{D.2}$$

Given the observed amplitude A, it is very difficult to infer the unknown parameters $(A_k)_{k=1}^K$ and $(\phi_k)_{k=1}^K$. To study the estimation of A, it is assumed that the amplitudes of all sinusoids are known (or can be represented by the partial amplitudes of source models). In the following,

the expected amplitude of two overlapping sinusoids is derived. An overlapping model is then suggested for more than two overlapping sinusoids.

Expected amplitude of two overlapping partials

Based on the assumption that the partials can be represented by sinusoids, the partial overlapping is represented by summing two sinusoids of the same frequency but of different amplitudes and phases:

$$s = s_1 + s_2 = A_1 \cos(\omega_1 t + \phi) + A_2 \cos(\omega_1 t) \tag{D.3}$$

The amplitude of s is

$$A = \sqrt{A_1^2 + A_2^2 + 2A_1 A_2 \cos\phi} \tag{D.4}$$

It is assumed that the phase difference ϕ is uniformly distributed between $-\pi$ and π, i. e., its probability density function is $f(\phi) = 1/2\pi$. The expected value of A can thus be calculated

$$\begin{aligned}
E(A) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \sqrt{A_1^2 + A_2^2 + 2A_1 A_2 \cos\phi} d\phi = \frac{1}{2\pi}\int_{-\pi}^{\pi} \sqrt{A_1^2 + A_2^2 + 2A_1 A_2 (1 - 2\sin^2\frac{\phi}{2})} d\phi \\
&= \frac{A_1 + A_2}{2\pi}\int_{-\pi}^{\pi} \sqrt{1 - \frac{4A_1 A_2}{(A_1 + A_2)^2}\sin^2\frac{\phi}{2}} d\phi \\
&= \frac{A_1 + A_2}{\pi}\int_{-\pi/2}^{\pi/2} \sqrt{1 - \frac{4A_1 A_2}{(A_1 + A_2)^2}\sin^2\theta} d\theta \\
&= \frac{2(A_1 + A_2)}{\pi}\int_{0}^{\pi/2} \sqrt{1 - \frac{4A_1 A_2}{(A_1 + A_2)^2}\sin^2\theta} d\theta \\
&= \frac{2(A_1 + A_2)}{\pi} E_p(\frac{2\sqrt{A_1 A_2}}{A_1 + A_2})
\end{aligned} \tag{D.5}$$

where

$$E_p(\mathrm{k}) = \int_0^{\pi/2} \sqrt{1 - \mathrm{k}^2 sin^2\theta} d\theta \tag{D.6}$$

is the **complete elliptic integral of the second kind**.

The variance of A can then be obtained

$$\begin{aligned}
\mathrm{var}(A) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} (\sqrt{A_1^2 + A_2^2 + 2A_1 A_2 \cos\phi} - E(A))^2 d\phi \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} (A_1^2 + A_2^2 + 2A_1 A_2 \cos\phi) d\phi - 2E(A)\frac{1}{2\pi}\int_{-\pi}^{\pi} \sqrt{A_1^2 + A_2^2 + 2A_1 A_2 \cos\phi} d\phi + E(A)^2 \\
&= A_1^2 + A_2^2 - E(A)^2
\end{aligned} \tag{D.7}$$

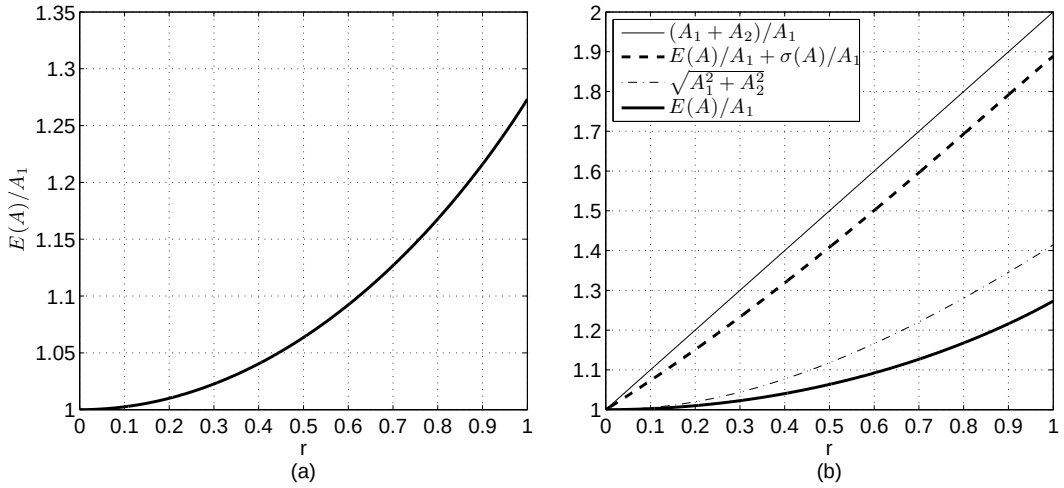


Figure D.1: (a) The expected overlap amplitude as a function of the amplitude ratio between two partials; (b) Comparison between the expected amplitudes and the other two overlap models.

Without loss of generality, substitute $A_2 = rA_1 \leq A_1$ (where $0 \leq r \leq 1$) into eq.(D.5) to have

$$\frac{E(A)}{A_1} = \frac{2(1+r)}{\pi} E_p(\frac{2\sqrt{r}}{1+r}) \tag{D.8}$$

This equation shows that as long as the amplitude ratio between two sinusoids is known, the increment of the overlapping amplitude relative to the stronger sinusoid can be deduced immediately (see Figure D.1 (a)). An **overlap model** for two overlapping partials, called the **expected overlap model**, is thus defined. The standard deviation appears to be rather large (compare the thick solid line and the thick dash line in Figure D.1 (b)), which implies the large uncertainty in the estimation of the amplitude of overlapping partials. Two assumptions usually made for overlapping partials are: the **additivity of the linear spectrum** $(A_1 + A_2)$ and the **additivity of power spectrum** $(\sqrt{A_1^2 + A_2^2})$. Both assumptions are in fact special cases of ϕ (see eq.(D.3)). The additivity of linear spectrum implies the maximum of the overlapping amplitude occurring when two sinusoids are *in phase*, that is , $\cos(\phi) = 1$. The additivity of power spectrum implies the cases when $\cos(\phi) = 0$, which is found close to the expected amplitude $E(A)$.

Expected amplitude of N overlapping partials

For the general case that N partials overlap, it is assumed that the expected amplitude can be deduced pair by pair using the proposed model. Given the amplitudes of all partials that overlap $\{A_1, A_2, \cdots, A_N\}$, the resulting amplitude of overlapping partials A' can be derived by the **overlap chain rule**:

Initialization of $A' \leftarrow A_1$
for $k = 2$ to N **do**
  $A' \leftarrow A' \bigoplus A_k$ `/* apply the overlap model for two partials */`
end
return A'

Evaluation

In order to compare the three overlap models, a test is designed to evaluate their modeling precision with the following algorithm:

Algorithm: Expected spectrum using overlap models

input : The observed signal of N mixing sources with fundamental frequencies $\mathrm{F0}_1 < \mathrm{F0}_2 < \cdots < \mathrm{F0}_N$ and known source models as partial amplitude sequences

output: The scaling factors $(c_k)_{k=1}^N$ for source models that minimize the modeling error

Initialization of c_1 by matching non-overlapped partials
for $k = 2$ *to* N **do**
  find non-overlapped partials: $\mathcal{U}$
  find partials overlapped with lower F0s but not higher F0s: $\mathcal{V}$
  do initialization of the expected amplitude of $\mathcal{V} \leftarrow$ **apply overlap model**
  if *exist* $\mathcal{U}$ **then**
    Initialization of c_k using $\mathcal{U}$ with LSE
  else
    Initialization of c_k using $\mathcal{V}$ with LSE
  end
  do optimize c_k using gradient descend for $\mathcal{U} + \mathcal{V} \leftarrow$ **apply overlap model**
end

The idea is to make use of the available source models such that their optimal scaling factors $(c_k)_{k=1}^N$ can be derived. The overlap model providing a more accurate estimate of overlapping amplitudes should give rise to a more accurate estimate of the scaling factors. As a consequence, the modeling error will be smaller. A synthesized music database is used (see Section 7.2.3) such that individual source signals are available, from which the true source models are extracted. Moreover, the overlap positions can be precisely inferred, which facilitates the calculation based on the overlap models. For each number of overlapping partials, the modeling error is normalized by the sum of the partial amplitudes. The results are shown in Figure D.2 for different numbers of overlapping partials. The modeling error for non-overlapped partials (the number of overlap = 1) is about 2%, which demonstrates that the scaling factors are well estimated. Of the three models, the proposed overlap model using the expected amplitude performs the best. The additivity of linear spectrum model has rather high errors compared to the other two models. In general, the additivity of power spectrum has similar performance to that of the expected overlap model.

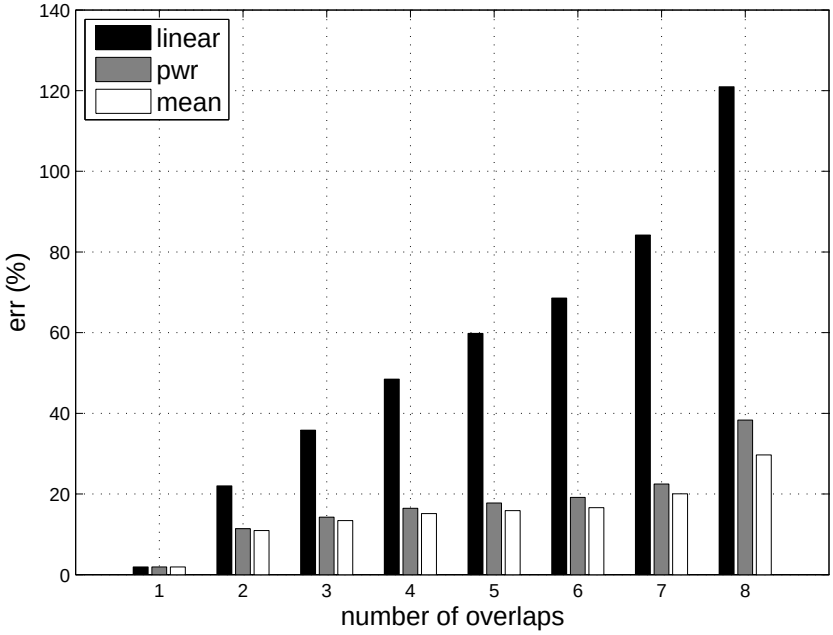


Figure D.2: Comparison of the three overlap models for the estimation of the amplitude of overlapping partials.

E

A New F0 Salience Function based on Inter-Peak Beating

(Harris, 1963) suggested locating all groups of *pitch harmonics* by means of identifying equally-spaced spectral peaks on which the salience of a group is built. The group with the most dominant salience is selected to estimate the F0. This method belongs to the **spectral interval** type F0 estimators (Klapuri, 2004). For instance, it is easy to identify the existence of F0 around 1380Hz in Figure E.1(a) because the successive peaks separated by this frequency difference are rather distinct and dominant in energy.

For polyphonic signals, however, partials belonging to different sources may form a group of harmonics which results in **ghost F0s**. For example, two strong partials at 250Hz and 360Hz from two sources may cause their frequency difference at 110Hz to appear as a good candidate. However, the partials are not situated within the tolerance intervals of the harmonic frequencies at 220Hz, 330Hz, 440Hz, etc. One way to avoid ghost F0 candidates is to cast further constraints on the **spectral location** of each partial. In the following, we propose a new salience function, called **inter-peak beating** (IPB), which integrates the spectral interval information and the spectral location information.

The spectral peaks can be classified as sinusoidal peaks or noise peaks by the adaptive noise level estimation (see Chapter 4). The noise peaks that do not result in significant beatings can be ignored during the grouping of equally-spaced spectral peaks. Given an F0 hypothesis, the sinusoidal peaks of the frequency difference within 3% of the F0 can be grouped into consecutive

partials. The **partial beating vector** is defined as

$$\mathrm{PBV}(h) = \min(a(h+1), a(h)), \text{ for } h = 1 \cdots H_b \tag{E.1}$$

where the minimum between two consecutive partials is selected for each pair and H_b is limited to 10. The beating vector is weighted by their **frequency proximity**

$$\mathrm{FP}(h) = \begin{cases} 1 - \frac{f_h - f_{h-1}}{\alpha F0} & \text{if } f_h - f_{h-1} < \alpha F0, \\ 0 & \text{otherwise.} \end{cases} \tag{E.2}$$

and then summed to obtain

$$\mathrm{IPB} = \sum_{h=1}^{H_b} \mathrm{PBV} \cdot \mathrm{FP} \tag{E.3}$$

for each F0 hypothesis. α defines the tolerance interval (see Section 5.1.1). In this way, IPB integrates the two criteria: harmonicity and partial beating into one salience function (see Figure E.1).

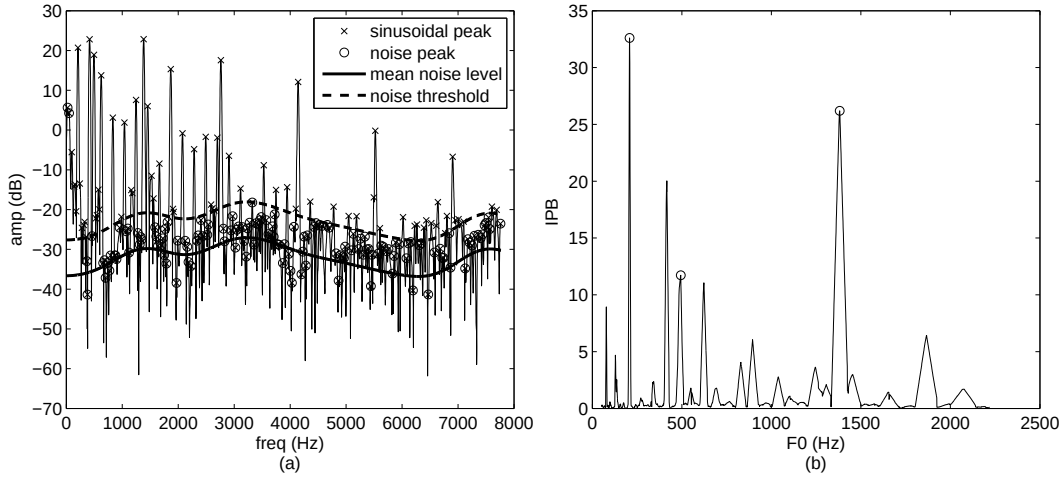


Figure E.1: F0 salience function based on Inter-Peak Beating: (a) An observed spectrum with the estimated noise level; (b) IPB. The circles indicate the correct F0s at around 208Hz, 494Hz and 1382Hz.

Bibliography

Abdallah, S. A. and Plumbley, M. D. (**2004**). "Polyphonic transcription by non-negative sparse coding of power spectra", in *Proc. of the 5th International Conference on Music Information Retrieval (ISMIR 2004)*, 318–325, Barcelona, Spain. (Cited on page 24.)

Abe, M. and Smith, J. O. (**2005**). "AM/FM rate estimation for time-varying sinusoidal modeling", in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, volume 3, iii/201–iii/204, Philadelphia, PA, USA. (Cited on pages 42, 118, and 119.)

Abe, T., Kobayashi, T., and Imai, S. (**1995**). "Harmonics tracking and pitch extraction based on instantaneou sfrequency", in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '95)*, volume 1, 756–759, Detroir, MI, USA. (Cited on page 66.)

Auger, F. and Flandrin, P. (**1995**). "Improving the readability of time-frequency and time-scale representations by the reassignment method", *IEEE Trans. on Signal Processing* **43**, 1068–1089. (Cited on pages 113, 116, and 118.)

Bach, F. R. and Jordan, M. I. (**2005**). "Discriminative training of hidden Markov models for multiple pitch tracking", in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, volume 5, 489–492, Philadelphia, PA, USA. (Cited on page 23.)

Badeau, R. (**2005**). *Méthodes à Haut Résolution pour l'Estimation et le Suivi de Sinusoides modulées. Application aux Signaux de Musique*, Ph.D. thesis, Ecole Nationale Supérieure des Télécommunications. (Cited on page 118.)

Badeau, R., David, B., and Richard, G. (**2004**). "Selecting the modeling order for the ESPRIT high resolution method: an alternative approach", in *Proc. of IEEE, International Conference on Acoustics, Speech and Signal Processing(ICASSP'042004)*, volume 2, ii–1025 – ii–1028, Montreal, Canada. (Cited on page 39.)

Baskind, A. (**2003**). *Modèles et Méthodes de Description Spatiale de Scènes Sonores*, Ph.D. thesis, Université Paris 6. (Cited on page 84.)

Baskind, A. and de Cheveigné, A. (**2003**). "Pitch-tracking of reverberant sounds, application to spatial description of sound scenes", in *Proc. of the AES 24th International Conference on Multichannel Audio*. (Cited on pages 17 and 102.)

Baumann, U. (**2001**). "A procedure for identification and segregation of multiple auditory objects", *Computational Models of Auditory Function* 269–280. (Cited on page 26.)

Beauchamp, J. W., Maher, R. C., and Brown, R. (**1993**). "Detection of musical pitch from recorded solo performances", in *94th AES Convention*, Berlin, Germany. (Cited on page 17.)

Bello, J. P. (**2003**). *Towards the Automated Analysis of Simple Polyphonic Music: A Knowledge-based Approach*, Ph.D. thesis, Department of Electronic Engineering, Queen Mary, University of London. (Cited on page 83.)

Bello, J. P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., and Sandler, M. B. (**2005**). "A tutorial on onset detection in music signal", *IEEE Trans. on Speech and Audio Processing* **13**, 1035–1047. (Cited on page 17.)

Benade, A. H. (**1976**). *Fundamentals of Musical Acoustics*, Dover Publications, Inc., New York. (Cited on page 6.)

Besson, O., Ghogho, M., and Swami, A. (**1999**). "Parameter estimation for random chirp signals", *IEEE Trans. on Signal Processing* **47**, 3208–3219. (Cited on page 118.)

Bogaards, N., Röbel, A., and Rodet, X. (**2004**). "Sound analysis and processing with AudioSculpt 2", in *Proc. of Internatinal Computer Music Conference (ICMC'04)*, Miami. (Cited on page 84.)

Bregman, A. S. (**1990**). *Auditory Scene Analysis*, The MIT Press, Cambridge, Massachusetts. (Cited on pages 9, 25, and 34.)

Brown, J. C. (**1991**). "Calculation of a constant Q spectral transform", *Journal of the Acoustical Society of America* **89**, 425–434. (Cited on page 27.)

Brown, J. C. (**1992**). "Musical fundamental frequency tracking using a pattern recognition method", *Journal of the Acoustical Society of America* **92**, 1394–1402. (Cited on page 13.)

Burred, J. J., Röbel, A., and Rodet, X. (**2006**). "An accurate timbre model for musical instruments and its application to classification", in *Proc. 1st Workshop on Learning the Semantics of Audio Signals (LSAS'06)*, Athens, Greece. (Cited on page 16.)

Cemgil, A., Kappen, B., and Barber, D. (**2006**). "A generative model for music transcription", *IEEE Trans. on Speech and Audio Processing* **13**, 679–694. (Cited on page 25.)

Chafe, C. and Jaffe, D. (**1986**). "Source separation and note identification in polyphonic music", in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '86)*, volume 11, 1289–1292. (Cited on pages 25 and 27.)

Chien, Y.-R. and Jeng, S.-K. (**2002**). "An automatic transcription system with octave detection", in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02)*, volume 2, 1865–1868, Orlando, FL, USA. (Cited on pages 26 and 27.)

Cohen, I. (**2003**). "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging", *IEEE Trans. on Speech and Audio Processing* **11**, 466–475. (Cited on page 38.)

Cohen, L. (**1995**). *Time-Frequency Analysis*, Prentice Hall. (Cited on pages 58 and 113.)

Cont, A. (**2006**). "Realtime multiple pitch observation using sparse non-negative constraints", in *Proc. of the 7th International Symposium on Music Information Retrieval (ISMIR'06)*. (Cited on page 24.)

Daudet, L. (**2004**). "Sparse and structured decompositions of audio signals in overcomplete spaces", in *Proc. of the 7th International Conference on Digital Audio Effects (DAFx-04)*, Naples,Italy. (Cited on page 17.)

Daudet, L. (**2006**). *Computer Music Modeling and Retrieval*, Chap. "A review on techniques for the extraction of transients in musical signals", Springer Lecture Notes in Computer Science series, Springer Publishing Company. (Cited on page 16.)

David, B., Richard, G., and Badeau, R. (**2003**). "An EDS modelling tool for tracking and modifying musical signals", in *Stockholm Music Acoustics Conference 2003*, 715–718, Stockholm, Sweden. (Cited on page 39.)

Davy, M. and Godsill, S. (**2003**). "Bayesian harmonic models for musical signal analysis", in *Bayesian Statistics 7: Proc. of the Seventh Valencia International Meeting*, edited by J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, Oxfort University Press, Valencia, Spain. (Cited on page 25.)

de Cheveigné, A. (**1993**). "Separation of concurrent harmonic sounds: fundamental frequency estimation and a time-domain cancellation model of auditory processing", *Journal of the Acoustical Society of America* **93**, 3271–3290. (Cited on pages 22 and 23.)

de Cheveigné, A. and Baskind, A. (**2003**). "F0 estimation of one or several voices", in *8th European Conference on Speech Communication and Technology (Eurospeech'03)*, 833–836, Geneva, Switzerland. (Cited on page 26.)

de Cheveigne, A. and Kawahara, H. (**1999**). "Multiple period estimation and pitch perception model", *Speech Communication* **27**. (Cited on page 23.)

de Cheveigné, A. and Kawahara, H. (**2001**). "Comparative evaluation of F0 estimation algorithms", in *Eurospeech 2001 Scandinavia*, 2451–2454, Aalborg, Denmark. (Cited on page 11.)

de Cheveigné, A. and Kawahara, H. (**2002**). "YIN, a fundamental frequency estimator for speech and music", *Journal of the Acoustical Society of America* **111**, 1917–1930. (Cited on pages 4, 11, and 90.)

de Krom, G. (**1993**). "A cepstrum-based techniques for determing a harmonics-to-noise ratio in speech signals", *Journal of Speech Hearing Research* **36**, 254–266. (Cited on pages 38 and 39.)

Dixon, S. (**2000**). "On the computer recognition of solo piano music", in *Australasian Computer Music Conference*, 31–37, Brisbane, Australia. (Cited on pages 26 and 83.)

Djuric, P. M. and Kay, S. M. (**1990**). "Parameter estimation of chirp signals", *IEEE Trans. on Signal Processing* **38**, 2118–2126. (Cited on page 118.)

Doval, B. and Rodet, X. (**1991**). "Estimation of fundamental frequency of musical sound signals", in *Proc. IEEE-ICASSP 91*, 3657–3660, Toronto. (Cited on pages 13 and 32.)

Dubois, C. and Davy, M. (**2007**). "Joint detection and tracking of time-varying harmonic components: a flexible Bayesian approach", *IEEE Trans. on Audio, Speech and Language Processing* **15**, 1283–1295. (Cited on page 66.)

Duifhuis, H. and Willems, L. (**1982**). "Measurement of pitch in speech: An implementation of Goldstein's theory of pitch perception", *Journal of the Acoustical Society of America* **71**, 1568–1580. (Cited on pages 13 and 53.)

Every, M. R. and Szymanski, J. E. (**2004**). "A spectral-filtering approach to music signal separation", in *Proc. of the 7th International Conference on Digital Audio Effects (DAFx-04)*, 197–200, Naples,Italy. (Cited on page 15.)

Fernandez-Cid, P. and Casajus-Quiros, F. (**1998**). "Multi-pitch estimation for polyphonic musical signals", in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '98)*, volume 6, 3565–3568, Seattle, WA, USA. (Cited on pages 26 and 27.)

Fletcher, N. F. and Rossing, T. D. (**1998**). *The Physics of Musical Instruments*, 2nd. edition, Springer-Verlag, New York. (Cited on page 5.)

Friedlander, B. and Francos, J. M. (**1995**). "Estimation of amplitude and phase parameters of multicomponent signals", *IEEE Trans. Acoustics, Speech and Signal Processing* **43**, 917–926. (Cited on page 118.)

Ghogho, M., Nandi, A. K., and Swami, A. (**1999**). "Cramér-Rao bounds and maximum likelihood estimation for random amplitude phase-modulated signals", *IEEE Trans. on Signal Processing* **47**, 2905–2916. (Cited on page 118.)

Goldstein, J. L. (**1973**). "An optimum processor theory for the central formation of the pitch of complex tones", *Journal of the Acoustical Society of America* **54**, 1496–1516. (Cited on page 13.)

Goto, M. (**2000**). "A robust predominant-F0 estimation method for real-time detection of melody and bass lines in CD recordings", in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '00)*, volume 2, II–757–760, Istanbul, Turkey. (Cited on pages 25 and 66.)

Goto, M. (**2003**). "RWC Music Database: Music Genre Database and Musical Instrument Sound Database", in *Proc. of the 4th International Conference on Music Information Retrieval (ISMIR 2003)*, 229–230, Baltimore, Maryland, USA. (Cited on pages 83 and 87.)

Goto, M. (**2006**). "AIST annotation for the RWC Music Database", in *Proc. of the 7th International Conference on Music Information Retrieval (ISMIR 2006)*, Victoria, Canada. (Cited on page 87.)

Goto, M., Hashiguchi, H., Nishimura, T., and Oka, R. (**2002**). "RWC Music Database: Popular, Classical, and Jazz Music Databases", in *Proc. of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, 287–288, Paris, France. (Cited on pages 83 and 87.)

Hainsworth, S. and Macleod, M. (**2003**). "Time-frequency reassignment: a review and analysis", Technical Report CUED/FINFENG /TR.459, Cambridge University Engineering Department. (Cited on page 116.)

Hainsworth, S., Macleod, M. D., and Wolfe, P. J. (**2001**). "Analysis of reassigned spectrograms for musical transcription", in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'01)*, 22–23, Mohonk, NY, USA. (Cited on page 39.)

Harold A. Conklin, J. (**1999**). "Generation of partials due to nonlinear mixing in a stringed instrument", *Journal of the Acoustical Society of America* **105**, 536–545. (Cited on page 16.)

Harris, C. M. (**1963**). "Pitch extraction by computer processing of high-resolution Fourier analysis data", *Journal of the Acoustical Society of America* **35**, 339–343. (Cited on pages 13 and 129.)

Hartmann, W. M. (**1998**). *Signals, Sound and Sensation*, Springer-Verlag New York Inc. (Cited on page 9.)

Hayes, M. H. (**1996**). *Statistical Digital Signal Processing and Modeling*, 1st edition, John Wiley & Sons, Inc. (Cited on page 26.)

Hermansky, H. and Morgan, N. (**1994**). "RASTA processing of speech", *IEEE Transaction on Speech and Signal Processing* **2**, 587–589. (Cited on page 22.)

Hess, W. (**1983**). *Pitch Determination of Speech Signals*, Springer-Verlag, Berlin Heidelberg. (Cited on pages 10, 11, and 18.)

Jensen, K. (**1999**). *Timbre Models of Musical Sounds*, Ph.D. thesis, Department of Computer Science, University of Copenhagen. (Cited on page 16.)

Johnson, N. L., Kotz, S., and Balakrishnan, N. (**1994**). *Continuous Univariate Distributions*, 2nd. edition, John Wiley & Sons, Inc, New York. (Cited on page 40.)

Jordan, M. I. (**2004**). "Graphical models", *Statistical Science* **19**, 140–155. (Cited on page 23.)

Jot, J.-M., Cerveau, L., and Warusfel, O. (**1997**). "Analysis and synthesis of room reverberation based on a statistical time-frequency model", in *103th AES Convention*, New York. (Cited on page 9.)

Kameoka, H., Nishimoto, T., and Sagayama, S. (**2004**). "Separation of harmonic structures based on tied Gaussian mixture model and information criterion for concurrent sounds", in *Proc. of IEEE, International Conference on Acoustics, Speech and Signal Processing (ICASSP'04)*, volume 4, 297–300, Montreal, Canada. (Cited on page 26.)

Kameoka, H., Nishimoto, T., and Sagayama, S. (**2005**a). "Audio stream segregation of multi-pitch music signal based on time-space clustering using Gaussian kernel 2-dimensional model", in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, volume 3, iii/5–iii/8, Philadelphia, PA, USA. (Cited on page 25.)

Kameoka, H., Nishimoto, T., and Sagayama, S. (**2005**b). "Harmonic-temporal-structured clustering via deterministic annealing EM algorithm for audio feature extraction", in *Proc. of the 6th International Conference on Music Information Retrieval (ISMIR'05)*, 115–122. (Cited on pages 25, 26, and 84.)

Kameoka, H., Nishimoto, T., and Sagayama, S. (**2007**). "A multipitch analyzer based on harmonic temporal structured clustering", *IEEE Transaction on Audio, Speech and Language Processing* 982–994. (Cited on pages 27 and 66.)

Kaprykowsky, H. and Rodet, X. (**2006**). "Globally optimal short-time dynamic time warping, application to score to audio alignment", in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '06)*, volume 5, Toulouse, France. (Cited on page 84.)

Karjalainen, M. and Tolonen, T. (**2000**). "A computationally efficient multipitch analysis model", *IEEE Trans. on Speech and Audio Processing* **8**, 708–716. (Cited on pages 23 and 56.)

Kashino, K., Nakadai, K., Kinoshita, T., and Tanaka, H. (**1998**). *Computational Auditory Scene Analysis*, Chap. "Application of the Bayesian probability network to music scene analysis", 115–137, Lawrence Erlbaum. (Cited on pages 26 and 83.)

Kashino, K. and Tanaka, H. (**1993**). "A sound source separation system with the ability of automatic tone modeling", in *Proc. of International Computer Music Conference (ICMC'93)*, 248–255, Tokyo, Japan. (Cited on pages 25 and 83.)

Keijzer, M. and Babovic, V. (**2000**). *Lecture Notes in Computer Science*, volume 1802/2004, Chap. "Genetic programming, ensemble methods and the bias/variance tradeoff - introductory investigations", 76–90, Springer-Verlag. (Cited on page 46.)

Keren, R., Zeevi, Y. Y., and Chazan, D. (**1998**). "Automatic transcription of polyphonic music using the multiresolution Fourier transform", in *9th Mediterranean Electrotechnical Conference, 1998 (MELECON'98)*, volume 1, 654–65, Israel. (Cited on page 27.)

Kitahara, T., Goto, M., Komatani, K., Ogata, T., and Okuno, H. G. (**2007**). "Instrument identification in polyphonic music: feature weighting to minimize influence of sound overlaps", *EURASIP Journal on Advances in Signal Processing* **2007**, Article ID 51979, 15 pages. (Cited on page 83.)

Klapuri, A. (**1998**). "Number theoretical means of resolving a mixture of several harmonic sounds", in *Proc. of the European Signal Processing Conference (EUSIPCO)*, Rhodes, Greece. (Cited on page 15.)

Klapuri, A. (**2004**). *Signal Processing Methods For the Automatic Transcription of Music*, Ph.D. thesis, Tampere University of Technology. (Cited on pages 52, 66, and 129.)

Klapuri, A. (**2005**). "A perceptually motivated multiple-F0 estimation method", in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'05)*. (Cited on pages 22, 26, and 27.)

Klapuri, A. (**2006**). "Multiple fundamental frequency estimation by summing harmonic amplitudes", in *Proc. of the 7th International Conference on Music Information Retrieval (ISMIR'06)*, Vienna, Austria. (Cited on pages 23, 26, 53, and 64.)

Klapuri, A. P. (**2003**). "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness", *IEEE Transactions on Speech and Audio Processing* **11**, 804–816. (Cited on pages 22, 33, 63, 64, 83, and 85.)

Kobzantsev, A., Chazan, D., and Zeevi, Y. (**2005**). "Automatic transcription of piano polyphonic music", in *Image and Signal Processing and Analysis 4th Int'l Symposium on Image and Signal Processing and Analysis (ISPA'05)*, Zagreb, Croatia. (Cited on page 27.)

Kumaresan, R. and Tufts, D. W. (**1982**). "Estimating the parameters of exponentially damped sinusoids and pole-zero modeling in noise", *IEEE Trans. Acoustics, Speech and Signal Processing* **30**. (Cited on page 118.)

Kuttruff, H. (**1991**). *Room Acoustics*, 4th edition, Spon Press, New York. (Cited on pages 6 and 8.)

Lagrange, M., Marchand, S., and Rault, J.-B. (**2004**). "Using linear prediction to enhance the tracking of partials", in *Proc. of IEEE, International Conference on Acoustics, Speech and Signal Processing(ICASSP'04)*, volume 4, iv–241 – iv–244, Montreal, Canada. (Cited on page 104.)

Lagrange, M., Marchand, S., and Rault, J.-B. (**2005**). "Tracking partials for the sinusoidal modeling of polyphonic sounds", in *Proc. of IEEE, International Conference on Acoustics, Speech and Signal Processing(ICASSP'05)*, volume 3, iii/229–iii/232, Philadelphia, PA, USA. (Cited on page 39.)

Lahat, M., Niederjohn, R. J., and Krubsack, D. A. (**1987**). "A spectral autocorrelation method for measurement of the fundamental frequency of noise-corrupted speech", *IEEE Trans. Acoustics, Speech and Audio Processing* **ASSP-35**. (Cited on page 12.)

Lea, A. (**1992**). *Auditory Model of Vowel Perception*, Ph.D. thesis, University of Nottingham. (Cited on page 22.)

Li, Y. and Wang, D. (**2007**). "Pitch detection in polyphonic music using instrument tone models", in *Proc. of IEEE, International Conference on Acoustics, Speech and Signal Processing (ICASSP'07)*, volume 2, II.481–484, Honolulu, HI, USA. (Cited on page 83.)

Licklider, J. C. R. (**1951**). "A duplex theory of pitch perception", *Cellular and Molecular Life Sciences (CMLS)* **7**, 128–134. (Cited on page 23.)

Loureiro, M. A., de Paula, H. B., and Yehia, H. C. (**2004**). "Timbre classification of a single musical instrument", in *Proc. of the 5th International Conference on Music Information Retrieval (ISMIR'04)*, Barcelona, Spain. (Cited on page 16.)

Lyon, R. F. (**1984**). "Computational models of neural auditory processing", in *Proc. of IEEE, International Conference on Acoustics, Speech and Signal Processing (ICASSP'84)*, volume 9, 41–44. (Cited on page 23.)

Maher, R. C. (**1990**). "Evaluation of a method for separating digitized duet signals", *Journal of Audio Eng. Soc.* **38**, 956–979. (Cited on page 54.)

Maher, R. C. and Beauchamp, J. W. (**1994**). "Fundamental frequency estimation of musical signals using a two-way mismatch procedure", *Journal of the Acoustical Society of America* **95**, 2254–2263. (Cited on page 23.)

Marolt, M. (**2004**). "A connectionist approach to automatic transcription of polyphonic piano music", *IEEE Trans. Multimedia* 439–449. (Cited on pages 27 and 83.)

Martin, K. D. (**1996**). "Automatic transcription of simple polyphonic music: robust front end processing", *MIT Media Laboratory Perceptual Computing Section Technical Report* . (Cited on page 25.)

Martin, P. (**1982**). "Comparison of pitch detection by cepstrum and spectral comb analysis", in *Proc. of IEEE, International Conference on Acoustics, Speech and Signal Processing(ICASSP'82)*, volume 7, 180–183. (Cited on page 13.)

Martin, R. (**1994**). "Spectral subtraction based on minimum statistics", in *Proc. of European Signal Processing Conference (EUSIPCO'94)*, 1182–1185. (Cited on page 38.)

Martin, R. (**2001**). "Noise power spectral density estimation based on optimal smoothingand minimum statistics", *IEEE Trans. on Speech and Audio Processing* **9**, 504–512. (Cited on page 38.)

Master, A. S. and Liu, Y.-W. (**2003**). "Nonstationary sinusoidal modeling with efficient estimation of linear frequency chirp parameters", in *Proc. of IEEE, International Conference on Acoustics, Speech and Signal Processing (ICASSP'03)*, volume 5, 656–659. (Cited on page 118.)

McIntyre, M. E., Schumacher, R. T., and Woodhouse, J. (**1983**). "On the oscillations of musical instruments", *Journal of the Acoustical Society of America* **74**, 1325–1345. (Cited on page 17.)

Mellinger, D. K. (**1991**). *Event Formation and Separation in Musical Sound*, Ph.D. thesis, Department of Computer Science, Stanford University. (Cited on page 25.)

Meyer, J. (**1972**). "Directivity of the bowed stringed instruments and its effect on orchestral sound in concert halls", *Journal of the Acoustical Society of America* **51**, 1994–2009. (Cited on page 8.)

Min, K., Chien, D., Li, S., and Jones, C. (**1988**). "Automated two speaker separation system", in *Proc. of IEEE, International Conference on Acoustics, Speech and Signal Processing (ICASSP'88)*, volume 1, 537–540, New York, NY, USA. (Cited on page 24.)

Molla, S. and Torrésani, B. (**2004**). "Determining local transientness of audio signal", *IEEE Signal Processing Letters* **11**, 625–628. (Cited on page 17.)

Monti, G. and Sandler, M. (**2002**). "Automatic polyphonic piano note extraction using fuzzy logic in a blackboard system", in *Proc. of the 5th International Conference on Digital Audio Effects (DAFx-02)*, 33–38, Hamburg, Germany. (Cited on page 83.)

Moorer, J. A. (**1977**). "On the transcription of musical sound by computer", *Comuter Music Journal* **1**, 32–38. (Cited on page 21.)

Nguyen, L. P. and Imai, S. (**1977**). "Vocal pitch detection using generalized distance function associated with a voice-unvoice decision logic", *Bull. P.M.E.* **39**, 11–21. (Cited on page 11.)

Noll, A. M. (**1967**). "Cepstrum pitch determination", *Journal of the Acoustical Society of America* **41**, 293–309. (Cited on page 12.)

Oppenheim, A. V., Willsky, A. S., and Nawab, S. H. (**1997**). *Signals & Systems*, 1997 edition, Prentice-Hall Inc. (Cited on page 4.)

Ortiz-Berenguer, L. I., Casajús-Quirós, F. J., and Torres-Guijarro, S. (**2005**). "Multiple piano note identification using a spectral matching method with derived patterns", *Journal of Audio Eng. Soc.* **53**, 32–43. (Cited on page 22.)

Parsons, T. W. (**1976**). "Separation of speech from interfering speech by means of harmonic selection", *Journal of the Acoustical Society of America* **60**, 911–918. (Cited on pages 15, 22, and 53.)

Patterson, R. D. and Holdsworth, J. (**1990**). "A functional model of neural activity patterns and auditory images", *Advances in speech, hearing and auditory images* . (Cited on page 27.)

Peeters, G. (**2001**). *Modèle et Modification du Signal Sonore Adapté à ses Caractéristiques Locales*, Ph.D. thesis, Université Paris 6. (Cited on page 119.)

Peeters, G. (**2003**). "A large set of audio features for sound description (similarity and classification) in the CUIDADO project", Technical Report. (Cited on pages 58 and 59.)

Peeters, G. (**2006**). "Music pitch representation by periodicity measures based on combined temporal and spectral representations", in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '06)*, volume 5, V–53 – V–56, Toulouse, France. (Cited on page 24.)

Peeters, G. and Rodet, X. (**1998**). "Sinusoidal characterization in terms of sinusoidal and non-sinusoidal components", in *Proc. of the 1st International Conference on Digital Audio Effects (DAFx-98)*, Barcelona, Spain. (Cited on page 39.)

Poliner, G. E. and Ellis, D. P. (**2006**). "A discriminative model for polyphonic piano transcription", *EURASIP JASP* . (Cited on pages 83 and 91.)

Qi, Y. and Hillman, R. E. (**1997**). "Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals", *Journal of the Acoustical Society of America* **102**, 537–543. (Cited on pages 39, 41, and 69.)

Raczynski, S. A., Ono, N., and Sagayama, S. (**2007**). "Multipitch analysis with harmonic nonnegative matrix approximation", in *Proc. of 8th International Symposium on Music Information Retrieval*. (Cited on page 24.)

Rayleigh, J. W. S. (**1945**). *The Theory of Sound, volume 2*, 2nd. edition, Dover Publications, New York. (Cited on page 8.)

Rife, D. and Borstyn, R. (**1974**). "Single tone parameter estimation from discrete-time observations", *IEEE Trans. on Information Theory* **20**, 591–598. (Cited on page 117.)

Rife, D. C. and Borstyn, R. R. (**1976**). "Multiple tone parameter estimation from discrete-time observations", *The Bell System Technical Journal* **55**, 1389–1410. (Cited on page 118.)

Ris, C. and Dupont, S. (**2001**). "Assessing local noise level estimation methods: application to noise robust ASR", *Speech Communication* 141–158. (Cited on pages 38 and 39.)

Rivet, B., Girin, L., and Jutten, C. (**2007**). "Log-Rayleigh distribution: a simple and efficient statistical representation of log-spectral coefficients", *IEEE Transaction on Audio, Speech and Language Processing* **15**, 796–802. (Cited on page 44.)

Röbel, A. (**2002**). "Estimating partial frequency and frequency slope using reassignment operators", in *Proc. of International Computer Music Conference (ICMC'02)*, 122–125, Göteborg. (Cited on page 118.)

Röbel, A. (**2003**a). "A new approach to transient processing in the phase vocoder", in *Proc. of the 6th International Conference on Digital Audio Effects (DAFx-03)*, 344–349, London. (Cited on pages 33 and 59.)

Röbel, A. (**2003**b). "Transient detection and preservation in the phase vocoder", in *Proc. of International Computer Music Conference (ICMC'03)*, 247–250, Singapore. (Cited on page 17.)

Röbel, A. (**2006**). "Onset detection in polyphonic signals by means of transient peak classification", in *International Symposium for Music Information Retrieval-MIREX (ISMIR/MIREX'06)*. (Cited on page 84.)

Röbel, A. and Rodet, X. (**2005**). "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation", in *Proc. of the 8th International Conference on Digital Audio Effects (DAFx-05)*, 30–35, Madrid, Spain. (Cited on page 44.)

Röbel, A. and Zivanovic, M. (**2004**). "Signal decomposition by means of classification of spectral peaks", in *Proc. of International Computer Music Conference (ICMC'04)*, Miami, Florida. (Cited on pages 39, 42, and 113.)

Rodet, X., Escribe, J., and Durigon, S. (**2004**). "Improving score to audio alignment: Percussion alignment and Precise Onset Estimation", in *Proc. of Internatinal Computer Music Conference (ICMC'04)*, Miami. (Cited on page 84.)

Rodet, X. and Jaillet, F. (**2001**). "Detection and modeling of fast attack transients", in *Proc. of International Computer Music Conference (ICMC'01)*, 30–33, Havana, Cuba. (Cited on pages 16 and 17.)

Ross, M. J., Shaffer, H. L., Cohen, A., Freudberg, R., and Manley, H. J. (**1974**). "Average magnitude difference function pitch extractor", *IEEE Trans. Acoustics, Speech Processing* **22**, 353– 362. (Cited on page 10.)

Roux, J. L., Kameoka, H., Ono, N., de Cheveigné, A., and Sagayama, S. (**2007**). "Single and multiple pitch contour estimation through parametric spectrogram modeling of speech in noisy environments", *IEEE Trans. on Speech, Audio and Language Processing* **15**, 1135–1145. (Cited on page 83.)

Ryynänen, M. and Klapuri, A. (**2005**). "Polyphonic music transcription using note event modeling", in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'05)*, Mohonk, NY, USA. (Cited on pages 27 and 84.)

Sagayama, S., Takahashi, K., Kameoka, H., and Nishimoto, T. (**2004**). "Specmurt anasylis: a piano-roll-visualization of polyphonic music signal by deconvolution of log-frequency spectrum", in *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing (SAPA'04)*. (Cited on page 24.)

Saha, S. and Kay., S. (**2001**). "A noniterative maximum likelihood parameter estimator ofsuperimposed chirp signals", in *Proc. of IEEE, International Conference on Acoustics, Speech and Signal Processing (ICASSP'01)*, volume 5, 3109–3112, Salt Lake City, UT, USA. (Cited on page 118.)

Schroeder, M. R. (**1968**). "Period histogram and product spectrum: new methods for fundamental frequency measurement", *Journal of the Acoustical Society of America* **43**, 829–834. (Cited on page 12.)

Schwefel, H.-P. (**1995**). *Evolution and Optimum Seeking*, Wiley & Sons, New York. (Cited on page 64.)

Serra, X. (**1989**). *A System for Sound Analysis/Transformation/Synthesis based on a Deterministic plus Stochastic Decomposition*, Ph.D. thesis, Stanford University. (Cited on page 117.)

Sha, F. and Saul, L. K. (**2005**). "Real-time pitch determination of one or more voices by nonnegative matrix factorization", *Advances in Neural Information Processing Systems* **17**, 1233–1240. (Cited on page 24.)

Shields, J. V. C. (**1970**). *Separation of Added Speech Signals by Digital Comb Filtering*, Ph.D. thesis, Massachusetts Institute of Technology. (Cited on page 21.)

Slaney, M. and Lyon, R. F. (**1990**). "A perceptual pitch detector", in *Proc. of IEEE, International Conference on Acoustics, Speech and Signal Processing (ICASSP'90)*, volume 1, 357–360, Albuquerque, NM, USA. (Cited on page 23.)

Sluyter, R., Kotmans, H., and Claasen, T. (**1982**). "Improvements of the harmonic-sieve pitch extraction scheme and an appropriate method for voiced-unvoiced detection", in *Proc. of IEEE, International Conference on Acoustics, Speech and Signal Processing(ICASSP'82)*, volume 7, 188–191. (Cited on page 13.)

Smaragdis, P. and Brown, J. (**2003**). "Non-negative matrix factorization for polyphonic music transcription", in *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'03)*, New Paltz, NJ, USA. (Cited on page 24.)

Smith, J. O. (**2007**). *Introduction to Digital Filters with Audio Applications*, W3K Publishing. (Cited on page 123.)

Sreenivas, T. V. and Rao, P. V. S. (**1981**). "Functional demarcation of pitch", *Signal Processing* **3**, 277–284. (Cited on pages 13 and 56.)

Sterian, A. D. (**1999**). *Model-Based Segmentation of Time-Frequency Images for Musical Transcription*, Ph.D. thesis, Department of Electronic Engineering, University of Michigan. (Cited on pages 26 and 83.)

Stevens, S. S. (**1970**). "Neural events and psychophysical law", *Science* **170**, 1043–1055. (Cited on page 56.)

Stuart, A. and Ord, J. K. (**1998**). *Kendall's Advanced Theory of Statistics, Vol. 1: Distribution Theory*, 6th. edition, Oxford University Press, New York. (Cited on page 43.)

Tokuda, K., Masuko, T., Miyazaki, N., and Kobayashi, T. (**1999**). "Hidden Markov models based on multi-space probability distributionfor pitch pattern modeling", in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '99)*, volume 1, 229–232. (Cited on page 102.)

Vincent, E. (**2004**). *Modèles d'Instruments pour la Séparation de Sources et la Transcription d'Enregistrements Musicaux*, Ph.D. thesis, Université Paris VI. (Cited on page 25.)

Virtanen, T. (**2003**a). "Algorithm for the separation of harmonic sounds with time-frequency smoothness constraint", in *Proc. of the 6th International Conference on Digital Audio Effects (DAFx-02)*, London, UK. (Cited on page 15.)

Virtanen, T. (**2003**b). "Sound source separation using sparse coding with temporal continuity objective", in *Proc. of International Computer Music Conference (ICMC'03)*. (Cited on page 24.)

Viste, H. and Evangelista, G. (**2002**). "An extension for source separation techniques avoiding beats", in *Proc. of the 5th International Conference on Digital Audio Effects (DAFx-02)*, 71–75, Hamburg, Germany. (Cited on page 15.)

Walmsley, P. J., Godsill, S. J., and Rayner, P. J. W. (**1999**). "Polyphonic pitch tracking using joint Bayesian estimation of multiple frame parameters", in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'99)*, Mohonk, NY, USA. (Cited on page 25.)

Weinreich, G. (**1997**). "Directional tone color", *Journal of the Acoustical Society of America* **101**, 2338–2346. (Cited on page 8.)

Weintraub, M. (**1986**). "A computational model for separating two simultaneous sound", in *Proc. of IEEE, International Conference on Acoustics, Speech and Signal Processing (ICASSP'86)*, volume 1, 3.1.1–3.1.4, Tokyo, Japan. (Cited on page 23.)

Wells, J. J. and Murphy, D. T. (**2006**). "High-accuracy frame-by-frame non-stationary sinusoidal modeling", in *Proc. of the 9th International Conference on Digital Audio Effects (DAFx-06)*, 253–258, Montreal, Canada. (Cited on page 118.)

Wu, M., Wang, D., and Brown, G. (**2003**). "A multipitch tracking algorithm for noisy speech", *IEEE Transactions on Speech and Audio Processing* **11**, 229–241. (Cited on pages 23, 27, 83, and 102.)

Yeh, C. (**2007**). "Multiple F0 estimation for MIREX 2007", The 3rd Music Information Retrieval Evaluation eXchange (MIREX'07). (Cited on page 99.)

Yeh, C., Röbel, A., and Rodet, X. (**2005**). "Multiple fundamental frequency estimation of polyphonic music signals", in *Proc. of IEEE, International Conference on Acoustics, Speech and Signal Processing (ICASSP'05)*, volume 3, iii/225 – iii/228, Philadelphia. (Cited on page 83.)

Yeh, C., Röbel, A., and Rodet, X. (**2006**). "Multiple F0 tracking in solo recordings of monodic instruments", in *120th AES Convention*, Paris, France. (Cited on pages 17 and 85.)

Zhou, G., Giannakis, G. B., and Swami, A. (**1996**). "On polynomial phase signals with time-varying amplitudes", *IEEE Trans. on Signal Processing* **44**, 848–861. (Cited on pages 118 and 119.)

Zhou, R. (**2006**). *Feature Extraction of Musical Content For Automatic Music Transcription*, Ph.D. thesis, École Polytechnique Fédérale de Lausanne. (Cited on pages 24 and 72.)

Zivanovic, M., Roebel, A., and Rodet, X. (**2007**). "Adaptive threshold determination for spectral peak classification", in *Proc. of the 10th International Conference on Digital Audio Effects (DAFx-07)*, 47–54, Bordeaux, France. (Cited on page 113.)