Mémoire de DEA ATIAM

Université de Paris VI - Université Aix Marseille II - Ecole National Supérieure des Télécommunications - Institut National Polytechnique de Grenoble - Institut de Recherche et de Coordination Acoustique Musique

Année Universitaire
2002 - 2003

Multiple Fundamental Frequencies Estimation

YEH, Chunghsin

ircam
Centre
Pompidou

Responsable de stage : Xavier RODET
Encadrant de stage : Axel RÖBEL

Remerciements

Lors de mes études effectuées à l'IRCAM, j'ai accumulé une dette envers les personnes suivantes, auxquelles j'aimerais exprimer ma profonde gratitude:

- Xavier Rodet, pour la générosité et gentillesse dont vous avez fait preuve en me donnant l'occasion de travailler au sein de l'équipe Analyse/Synthèse.
- Axel Roöbel, pour vos conseils précieux et pour vos encouragements, sans quoi ce travail n'aurais pas été possible .
- Miroslav Zivanovic, pour avoir partagé votre savoir à propos de la séparation de pics sinusoidaux des pics non-sinusoidaux.
- Matthias Krauledat, pour m'avoir aidé à tracer les codes sources.
- Geoffroy Peeters et tous les autres membres de l'équipe Analyse/Synthèse, pour leur gentillesse.
- Mes camarades du DEA ATIAM: Anne-Florence Dosser, Chloé Clavel, David Cournapeau, Filippo Bonini, Guillaume Denis, Jean-François Oliver, Maëva Garnier, Marie-Céline Bezet, Mathieu Guillaume, Massimo Muzzi, Natacha Paniez, Nicolas Durand, Olivier Macherey, Raphaël Duree, Thomas Pellegrini et Terence Caulkins, pour m'avoir accompagné et aidé pendant les cours et pour préparer les examens.
- Cyrille Defaye, pour votre assistance soutenue au cours de l'année.
- Ricardo Canzio, pour votre appui continue qui m'a permis de réaliser mes rêves.
- Mes parents, pour votre amour.

Acknowledgements

During the studies at IRCAM, I have incurred much debt from the following people to whom I would like to express my deep appreciation:

- Xavier Rodet, for your generosity and kindness to offer me an opportunity to work at the Analysis/Synthesis team.
- Axel Roöbel, for your invaluable advice and unfailing encouragement, which made this work possible.
- Miroslav Zivanovic, for sharing your knowledge of separating sinusoidal peaks from non-sinusoidal peaks.
- Matthias Krauledat, for helping me trace the program codes.
- Geoffroy Peeters and all the others at the Analysis/Synthesis team, for treating me kindly.
- All my classmates of DEA ATIAM: Anne-Florence Dosser, Chloé Clavel, David Cournapeau, Filippo Bonini, Guillaume Denis, Jean-François Oliver, Maëva Garnier, Marie-Céline Bezet, Mathieu Guillaume, Massimo Muzzi, Natacha Paniez, Nicolas Durand, Olivier Macherey, Raphaël Duree, Thomas Pellegrini and Terence Caulkins, for accompanying me and helping me through the courses and exams.
- Cyrille Defaye, for your special concerns all the time.
- Ricardo Canzio, for your continue support to make my dream come true.
- My parents, for your love.

Résumé

La fréquence fondamentale, dite $F0$, occupe une position clef dans les signaux musicaux et les signaux de parole du fait qu'elle soit fortement liée à la perception de la hauteur. Von Helmholtz pensait que $F0$ déterminait à elle seule la perception de hauteur, mais sa théorie fut critiqué un siècle après son apparition. Schouten fut le premier à prouver que la périodicité des harmoniques élevées jouait un rôle important dans la perception de hauteur. Plomp conclut des expériences menées par Ritsma que les harmoniques supérieures était responsables de la sensation de hauteur jusqu'à 1500Hz.

Les théories de l'analyse de hauteur auditive tendent à se distinguer en deux dimensions: la hauteur spectrale fondée sur l'information de lieu et la hauteur de périodicité fondée sur l'information de périodicité. Ainsi, les téchniques pour estimer la fréquence fondamentale peuvent être catégorisée en deux domaines: le domaine temporel et le domaine spectral. Les techniques temporelles incluent l'extraction de fondamentales-harmoniques utilisant des extracteurs pour générer des marqueurs de période, et l'analyse de la structure temporelle en observant l'enveloppe temporelle d'un signal. Les techniques spectraux incluent la méthode d'autocorrélation, les fonctions de distance, cepstrum, la compression spectrale et l'appariement spectral. Toutes ces techniques ont été appliquée à la détection de fréquence fondamentale simple, cependant la détection de fréquences fondamentales multiples requiert des techniques et des connaissances approfondies.

Les approches récentes fondées sur des modèles statistiques intègrent les paramètres qui font sens au niveau musical dans un système permettant d'estimer des $F0$s. *PreFEst* dévéloppée par M. Goto est capable d'extraire la mélodie et la basse d'une musique enregistrée. Les modèles graphiques Bayesian, proposés par Paul J. Walmsley, Simon J. Godsill et Peter J. W. Rayner, fournissent un cadre pour une modélisation graphique qui représente les dépendances statistiques entre des données observées et des paramètres d'un modèle. Le système de transcription automatique, proposé par Anssi P. Klapuri, qui n'utilise pas de modèles statistiques fournit un cadre complèt pour pouvoir transcrire la musique multiphonique et est également appliqué à la synthèse des signaux multiphoniques simples. Puisqu'il existe beaucoup de problèmes à résoudre dans l'estimation de $F0$s multiples, on simplifie la tâche courante de la manière suivante:

1) Les fréquences cibles $F0$s ne sont pas multiples les unes des autres.

2) Le nombre de $F0$s visées est supposé connu à l'avance.

3) On considère seulement les parties stationnaires dans les signaux musi-

caux.

`f0` est un programme dévéloppé pour estimer les fréquences fondamentales multiples. On résume ses processus pas par pas:

1) Les modèles spectraux idéaux se construisent par la superposition des pics spectraux sinusoidaux espacés de manière égale.

2) Elimination provisoire du plus grand pic observé dans le spectre de manière à pouvoir détecter les pics cachés par les formants.

3) Extraction de pics candidats à être $F0$s.

4) Assigner les regions de pondération pour tous les pics.

5) Filtrer les pic de bruits en observant les propriétés temporelles-fréquentielles: *bandwidth*(largeur de bande), *duration*(durée), *group delay*(délai de groupe), et *instantaneous frequency*(fréquence instantanée).

6) Elimination des candidats correspondant aux subharmoniques des $F0$s en utilisant HLR_{F0} qui mésure les variations des amplitudes spectrales d'une série de harmoniques. Les pics non désirés sont ensuite filtrés par Dev_{F0}, qui représente la variance de *mean time*(temps moyenné).

7) Regrouper les candidats $F0$s qui ont une relation d'harmonique. Dans chaque groupe, le candidat possédant la distance minimale par rapport au spectre observé est gardé à l'etape finale. En plus de HLR_{F0} et Dev_{F0}, $Shift_{F0}$ est un autre composant de la fonction de distance et il mésure les écarts dans le groupe de partiels correspondant à un $F0$.

8) Deux composants concernant les propriétés combinées sont introduit dans la fonction de distance finale. $ErrSpec_{F0}^{N_{note}}$ calcule l'erreur entre la somme des amplitudes spectrales d'un spectre combiné et celle d'un spectre observé. $Shift_{F0}^{N_{note}}$ mésure les valeurs $Shift_{F0}$ combinées. Des propriétés individuelles comme HLR_{F0} et Dev_{F0} sont additionées pour chaque combinaison des $F0$s candidats. La combinaison dont la fonction de distance est minimale représente le resultat de l'estimation.

Les paramètres introduits lors du processus entier sont générés et testés par l'algorithme d'évolution. Des échantillons sonores sont choisis aléatoirement parmi des échantillons correspondant à des instruments divers. Pour 100 échantillons choisis aléatoirement, `f0` a obtenu une moyenne d'erreur de 23.6%. Si l'on désire estimer des fréquences fondamentales multiples sur des extraits musicaux réels (non-simplifiés), une étude approfondie des irrégularités spectrales sera éssentielle. La détection des attaques de notes pourra également aider à classifier certains groupes de partiels. Une combinaison de connaissance musicale et de techniques de traitement de signal sera la clef pour pouvoir détecter des fondamentales multiples dans des extraits musicaux réels.

Abstract

Fundamental frequency, or $F0$, occupies a key position in musical signals and speech signals because it is strongly related to pitch perception. Von Helmhotz believed that $F0$ itself determines the percieved pitch and it was critisized one century after its appearance. Schouten was the first one to prove that the periodicity of higher harmonics plays an important role in pitch perception. Plomp concluded from Ritsma's experiments that higher harmonics rather than $F0$ dominate pitch sensation up to a frequency range about 1500Hz.

The theories of auditory pitch analysis tend to differ on two dimensions: spectral pitch based on place information and periodicity pitch based on periodicity information. Thus, fundamental frequency estimation techniques could be categorized into two domains: time domain and spectral domain. Time domain techniques include fundamental-harmonic extarction using basic extractors to generate periodic markers, and temporal structure analysis by observing temporal envelope of signals. Spectral domain techniques include autocorrelation, distance functions, cepstrum, spectral compression and spectral matching. All these techniques have been applied to single $F0$ estimation while multiple $F0$s estimation requires more techniques and knowledge.

Recent approaches based on statistical models integrate musically meaningful parameters into multiple $F0$s estimation systems. *PreFEst* developed by M. Goto is capable of extracting the melody and bass lines in complex music recordings. Bayesian graphical models, proposed by Paul J. Walmsley, Simon J. Godsill and Peter J. W. Rayner, provides a graphical modeling framework which represents the statistical dependencies between observed data and model parameters. Anssi P. Klapuri's automatic transcription system without constructing statistical models gives a complete framework of transcribing polyphonic music and is further applied to synthesize simple polyphonic signals. Since there are many problems with estimating multiple $F0$s, we simplify our current tasks in the follwing aspects:

1) The target $F0$s are not multiples of one another.

2) The number of $F0$s are assumed to be known in advance.

3) Only stationary parts of musical signals are considered.

`f0` is a program developed for estimating multiple $F0$s. The processes are summarized step by step:

1) Constructing ideal spectral models by summing up equally-spaced sinusoidal spectral peaks.

2) Removing the largest peak in the observed spectrum to detect the peaks possibly hidden in formants.

3) Extracting peaks as $F0$ candidates.

4) Assigning influential regions for peaks extracted.

5) Filtering noise peaks by observing *bandwidth*, *duration*, *group delay* and *instantaneous frequency* of each peak.

6) Eliminating the candidates corresponding to subharmonics of correct $F0$s by HLR_{F0} which is a measure of the spectral amplitude variation in a sequence of harmonic partials. Unwanted peaks are futher filtered using Dev_{F0}, the variance of mean time.

7) Grouping $F0$ candidates which are harmonically related. In each group, only the $F0$ candidate with the smallest distance is kept to the final stage. In addition to HLR_{F0} and Dev_{F0}, $Shift_{F0}$ is another component included in this distance function and it is a measure of the deviation of the group of partials belonging to one $F0$.

8) Two components concerning the combined properties of $F0$ candidates are introduced in the final distance function. $ErrSpec_{F0}^{N_{note}}$ calculates the error between the sum of spectral magnitudes of a combined spectrum and that of an observed spectrum. $Shift_{F0}^{N_{note}}$ measures the combined $Shift_{F0}$ values. Individual properties like HLR_{F0} and Dev_{F0} are summed for each combination of $F0$ candidates. The combination of the smallest distance gives the final result.

The parameters introduced in the whole process are generated and tested by the evolutionary algorithm. Testing samples are randomly mixed by samples of a variety of musical instruments. For 100 randomly mixed samples, `f0` has been tested to perform an average error rate of 23.6%. To be able to estimate multiple $F0$s in real world music, a profound study on spectrum irregularities is essential. Detecting the onsets of notes also helps to classify certain groups of partials. That is, integrating musical knowledge with signal processing techniques is the key to estimate multiple $F0$s in real world music.

Contents

Chapter 1

Introduction

1.1 Pitch perception of complex tones

A simple description of the process of pitch perception could 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 could be interpreted by human brain as a pitch. The American National Standard Institute(ANSI) defines the term "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) adds that pitch is associated with frequency and is low or high according to whether this frequency is smaller or greater.

In Von Helmhotz's pitch theory, a pitch is considered to be determined by $F0$, the fundamental frequency, which was generally accepted for about one century. The first successful attack on the significance of $F0$ in low-pitch[1] sensation was made by Schouten[2]. He investigated that the pitch of the complex tone was the same as that prior to the elimination of $F0$. He described the sensation induced by a periodic sound wave with many partials:

> "The lower harmonics can be perceived individually and have almost the same pitch as when sounded separately. The higher harmonics, however, cannot be perceived separately but are perceived collectively as one component with a pitch determined by the periodicity of the collective waveform, which is equal to that of the fundamental tone."

The controversy that $F0$ is not essential for perceiving low pitch raises the question as to which harmonics are most important. From Ritsma's experiments in 1963, Plomp concluded that those partials below the *fifth* harmonics give a much better low-pitch sensation than the higher harmonics. Plomp further found that higher harmonics rather than $F0$ dominate pitch sensation up to a frequency of about 1500Hz.

[1] Depending upon the experimental conditions, the pitch of complex tones are referred to as residue pitch, low pitch, periodicity pitch, time separation pitch, repetition pitch, virtual pitch, etc.

The theories of auditory pitch analysis tend to differ in two dimensions[1]: 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. Human auditory system seems to perceive pitch through "pattern matching". We recognize a sound by its spectral pattern composed of a series of partials which characterize it. Even when some partials are too weak to be detected, human auditory system tends to reconstruct the missing partials and complete the pattern matching task.

1.2 Fundamental frequency estimation

Wolfgang Hess categorizes the PDA(pitch determination algorithms, or fundamental frequency estimation algorithms), into two domains[3]:

1) Time domain: If there is a time-domain signal at this point that has the same time base as the input signal, the PDA works in the time domain.

2) Spectral domain: The alternative is the lag domain if the input is a correlation function, or the frequency domain if the input is a Fourier spectrum or some functions derived from it. All these other than the time domain itself are labeled as spectral domains.

1.2.1 Time domain techniques

Fundamental-harmonic extraction

There are three basic extraction methods: the zero crossings analysis basic extractor(ZXABE), the nonzero threshold analysis basic extractor(TABE), and the TABE with hysteresis(two-threshold basic extractor). They can be characterized as event detectors. When the extractors detect the significant event they are designed for, a marker is generated. If the basic extractor operates correctly, the elapsed time between two consecutive markers represents a pitch period.

The technique of fundamental-harmonic extraction in the time domain requires the first partial to be present in the waveform. This restricts the application of these techniques to those cases where the signal is not band limited unless nonlinear preprocessing reconstructs the first partial. Another drawback of these techniques is their sensitivity to low frequency signal distortions.

Temporal structure analysis

Basically, one should be able to read the periodicity of the signal out of its temporal structure. In this respect we could model the process from which periodicity is determined visually from an oscillogram. There are mainly two types of methods based on this concept:

1) Modeling of signal envelope and searching for discontinuities which mark the beginning of individual periods.

2) Direct investigation of the temporal structure by algorithm; search and extraction of anchor points from which periodicity is derived.

To increase the overall efficiency of the system, a simplification of the temporal structure such as inverse filtering is usually applied.

1.2.2 Spectral domain techniques

Autocorrelation

Correlation is a measure of similarity. In the case of autocorrelation the input sequences are correlated with themselves, with the lag as the parameter of the autocorrelation function. If the signal is periodic or quasi periodic, there are great similarities, i.e., high correlation coefficients, when the lag equals one period or a multiple thereof.

Distance function

Contrary to the correlation techniques which are measures of similarity, distance functions detect periodicity by investigating the global deviation between two sequences. A strong minimum implies that the lag equals $T0 = 1/F0$.

Cepstrum

The term cepstrum was introduced by Bogert and has been accepted terminology for the inverse Fourier transform of the logarithm of the power spectrum of a signal. The pulse sequence originating from the periodic signal reappears in the cepstrum as a strong peak at the "quefrency"(lag) $T0$.

Spectral compression

Spectral compression, first proposed by Schroeder(1968), opens an avenu to accurate determination of $F0$ from higher harmonics without requiring the respective harmonic numbers to be known. The frequencies of selected peaks are noted as its entries. They are first divided by two, three, four,..., and so on and then to be added together. The histogram finally has a distinct maximum at $F0$.

The harmonic *product* spectrum and the harmonic *sum* spectrum are the generalizations of the principle of spectral compression. The harmonic product spectrum is computed when the log power spectrum is compressed and added. Similarly, the hamonic sum spectrum is defined when the compressed amplitude spectra are added instead of the compressed logarithmic spectra.

Spectral matching

Martin applies the principle of harminic pattern matching with a comb filter[11] to estimate pitch:

> The principle behind the comb method consists in the search for values of the spectrum situated at harmonic frequencies, and whose sum is a maximum for a given frequency interval. The intercorrelation of spectrum and comb amounts to the calculation of the sum of the spectral components corresponding to a given harmonic structure. The fundamental corresponding to the harmonic structure giving the largest sum is then taken to be the fundamental frequency of

the signal, as long as this sum differs sufficiently from the values obtained for other structures in the same spectrum.

1.3 Analysis/Synthesis team at IRCAM

The main objective of the Analysis/Synthesis team is to design and develop tools to help composers and musicians implement their creative ideas, and also to provide musicologists with efficient measures to analyze musical pieces. Under the direction of Xavier Rodet, the Analysis/Synthesis team has accomplished many important projects like "Chant" and "Farinelli" projects, and has kept exploring new techniques in creating high-quality analysis/synthesis tools. The research areas include signal modeling, fundamental frequency estimation, separation of signals, physical modeling, signal description, score and speech following, gestual control, to mention just a few. Creative softwares like **Super Vocodeur de Phase**, **AudioSculpt**, **Diphone Studio** and **Chant** have been provided for IRCAM forum members. Estimating single fundamental frequency has been studied profoundly and implemented in these tools and `f0` is the first project to estimate multiple $F0$s. Since the complexity in real world music is difficult to deal with, we start studying this problem under simplified conditions.

1.4 Organization of this thesis

In this chapter, we review some important concepts concerning pitch perception of complex tones. Several $F0$ estimation techniques are surveyed, too. In the second chapter, we discuss the recent approaches to estimating multiple $F0$s and clarify our current tasks. In the third chapter and the fourth chapter, we propose the algorithms for estimating single $F0$ and multiple $F0$s. The testing results will be shown and discussed in the fifth chapter. Finally, we give conclusions of the current algorithms and propose future work.

Chapter 2

Recent approaches and difficulties

Since multiple $F0$s estimation is a complicated and difficult task, we first survey three recent approaches to study this problem and then we define our current tasks.

2.1 *PreFEst*

Masataka Goto developed a method called *PreFEst* (Predominant-$F0$ Estimation method)[5], which is based on the EM(Expectation-Maximization) algorithm and which is able to detect the melody and bass lines in complex mixtures containing simultaneous sounds of various instruments. This method is summarized as follows:

- Estimating the $F0$ of the most predominant harmonic structure in the input mixing signals.
- Simultaneously taking into consideration all the possibilities of $F0$s and the input mixture which contains every possible harmonic structure with different weights.
- Regarding the input frequency components as a weighted mixture of harmonic structure tone models of all possible $F0$s and finding the $F0$ of the maximum-weight model corresponding to the most predominant harmonic structure.
- A multiple-agent architecture is introduced to consider the global temporal continuity of estimated $F0$s and the final $F0$ output is determined on the basis of the most dominant and stable $F0$ trajectory.

The probability model Goto constructed contains several important concepts and, however, has some limitations:

1) For each $F0$ tone model, the center frequencies of all partials are modeled by a Gaussian distribution, given a variance of about 16 Hz. This modeling

method takes into consideration that the harmonic partials usually deviate from the theoretical position[1].

2) Two kinds of tone models are evaluated for each $F0$ candidate. One is the tone model with constant peak magnitudes and the other is the tone model with 2/3 peak magnitudes at even harmonic partials. Introducing different tone models is a solution to modeling the variations between the relative levels of even and odd harmonics due to the nature of musical instruments or tone quality controlled by musicians.

3) Since *PreFEst* depends strongly on the spectral magnitudes, peaks corresponding to the $F0$s of simultaneous sounds tend to compete in the observed spectral probability density and are transiently selected. A multiple-agent architecture is used to find the most predominant as well as the most stable $F0$. Thus, a modeling method depending too much on the spectral magnitudes is not the best model.

4) In *PreFEst*, the analyzing region is divided into two: one for estimating the melody line and the other for estimating the bass line. Then, single $F0$ estimation is evaluated, based on a mixed spectrum, in each analyzing region. This tells us the reason that Goto's system works better for pop music: an usual mixing technique brings out the vocal part, thus the melody, and the bass part the most.

5) The features of overlapped peaks are not treated separately form those of independant peaks.

2.2 Bayesian graphical models

Proposed by Paul J.Walmsley, Simon J. Godsill and Peter J.W. Rayner, Baysian graphical models are a flexible tool for the modeling of musical signals[7]. To model musical signals, they employ a graphical modeling framework which represents the statistical dependencies between observed data and model parameters. Each frame of data is modeled as the sum of a number of musical notes. The notes in each frame are assumed conditionally independent of hyperparameters which demonstrate their underlying variations. One of the main assumptions at this stage is that the parameters of musical signals are highly correlated in time, *i.e.*, they vary slowly over several frames.

The Bayesian framework allows for incorporating *a priori* information into the model and also forms a basis for probabilistic model selection in the joint detection and estimation of musical signals. Markov Chain Monte Carlo methods are employed to produce maximum *a posteriori* parameter estimates which enable the flexible choice of *a priori* probability distributions, which would otherwise be analytically intractable.

This has been regarded as a powerful statistical approach and here are a few discussions of its advantages and disadvantages:

1) The choices of *a priori* distributions give a good example of constructing a statistical model. It is pointed out in[7] that the priors could be made

[1]Multiples of the fundamental freqeuncy

more informative if salient prior knowledge is available. However, for the variance of fundamental frequency(hyperparameter), it is difficult to produce a value that represents the prior belief.

2) Musical signals generally exhibit a rapid variation and thus frequencies may vary rapidly. Dealing with the transient parts of musical signals is always another challenge, especially when they are mixed with stationary parts. Detection of onsets of notes and quantization of durations of notes is necessary to produce a musically readable output.

3) The inharmonicity is not modeled in this method. The excitation of some instruments may have some degree of aperiodicity as a result of the chaotic oscillations causing the excitation, whereas the harmonic model assumes periodic oscillations.

In a recent paper[8], an improved Bayesian model has been proposed. The inharmonicity is modeled as an detuned factor of each partial. However, robust results are obtained restrictly for no more than three notes.

2.3 Automatic transcription of music

An automatic music transcription system developed by Anssi P. Klapuri gives a complete framework of transcribing polyphonic music signals. Here is an overview of this system:

- A simplified version of [9] is used to detect the onsets of notes: detecting onsets in the logrithmic magnitude envelopes at distinct frequency bands and then combining the results across channels.

- A concept of dealing with the features of a mixture of harmonic sounds is introduced[10]: the features of prime number harmonics of each $F0$ in polyphonic music signals are the most independent features.

- Controlling the emphasis of "weighted order statistic"(WOS) filters, which estimate the probabilities of the interference among partials belonging to different sources, on the observed features, a tradeoff between the features of independant partials and those of dependant partials could be made.

- Based on the Bounded Q transform[2], the $F0$ candidates are estimated by combining the offset likelihoods in each band to yield global pitch likelihoods and the maximum global likelihood is used to determine $F0$. After iteratively separating a series of harmonics of the maximum $F0$ and estimating the most predominant $F0$, multiple $F0$s are obtained in a succession of iterative calculations.

- To compensate for a subtracted spectrum at each iteration, a spectrum smoothing method is proposed. The idea is derived from psychoacoustics, since the human auditory system prefers to associate a series of partials to a single acoustic source if they have a smooth spectrum and decreasing amplitude as a function of frequency[1].

[2]Constant number of bins in each octave

Keith D. Martin has built an automatic polyphonic music transcription system within a blackboard framework integrating frond ends based on autocorrelation with musical knowledge[11]. The implementation of this system is summarized as follows:

- The front end signal processing is modeled after the log-lag correlogram of Daniel Ellis[12]. A log-lag correlogram has three axes: filter channel frequency, lag(inverse of pitch) on a logarithmic scale, and time. Ellis normalizes the output of each frequency/lag cell by the energy in that filter bank channels, yielding a *summary autocorrelation*. The contention to built a transcription system with a correlated-based front end is that it requires no instrument models.
- The local maxima of each summary autocorrelation frame form a series of subharmonics from which periodicity hypotheses could be extracted.
- Onset hypotheses are derived directly from the energy envelope of the signal.
- Note hypotheses consist of one or more periodicity hypotheses combined with one onset hypotheses.
- Five knowledge souces are arranged in the sequence of processing to label and rate the hypotheses at each stage. A "Prune Note" knowledge source emphasizes above all western musical intervals to obtain a musically meaningful output.

Martin believed that the correlogram/periodogram representation may offer an advantage over sinusoidal representations for detecting the presence of octaves. He also mentioned that integrating musical knowledge such as the number and the type of instruments is necessary to build a useful transcription system.

2.4 Difficulties in estimating multiple $F0$s

After suveying the above methods, we summarize the difficulties in estimating multiple $F0$s:

1) The number of $F0$s is difficult to estimate without integrating perceptual and musical knowledge. Goto's system doesn't deal with this problem. The transcription system based on Bayesian graphical models keeps the notes with an energy within 30dB of the signal energy. Klapuri has measured several features such as the signal-to-noise ratio as indications to stop the iterative estimation-separation process.

2) The irregularities of the spectrum should be modeled into the estimation system. Except Martin's system, the recent approaches mentioned above have included certain concerns about inharmonicity. Although the physical inharmonicity of musical instruments might be integrated into an estimation system, inharmonicity caused by different playing techniques is difficult to estimate.

3) Even though a series of harmonic partials corresponding to one $F0$ are correctly estimated, it remains a challenge to extract reliable features of all partials since many of them are disturbed by harmonic partials of other $F0$s. Klapuri controls the importance between independant features and dependant features with the WOS filters.

4) Estimating multiple $F0$s in frames where transient parts and stationary parts are mixed requires another technique by observing time-frequency properties.

5) While room acoustics is taken into account, the estimation task becomes even more complicated: The harmonic structure is "blurred" and the prolongation of notes causes more complex mixtures in the spectrum. Also, the room boosts certain low frequencies.

2.5 Task description

Since there are many problems with estimating multiple $F0$s, we simplify our current tasks in the follwing aspects:

1) The target $F0$s are not multiples of one another.

2) The number of $F0$s are assumed to be known in advance.

3) Only stationary parts of musical signals are considered.

Chapter 3

Current algorithms for estimating $F0$ in monophonic signals

`f0` is a program developed for estimating multiple fundamental frequencies. Although the previous version is developed to estimate single $F0$ in speech signals, many of its criteria are designed for estimating multiple $F0$s. In this chapter, the architecture of `f0` is explained and its single $F0$ estimation criteria will be discussed concerning multiple $F0$s in polyphonic signals. The testing results of the previous version could be found in [13].

3.1 Constructing spectral models

In estimating single fundamental frequency, `f0` takes all the frequencies in a predefined range as candidates. For each fundamental frequency candidate, a spectral model is constructed by equally placing spectral peak models obtained from the Fourier transform of harmonic sinusoids(Fig. 3.1).

3.2 Detecting peaks covered by the largest peak

Since the formant often covers its neighboring peaks, we remove the largest peak(the most possible formant) to resolve the hidden peaks. There are two techniques to estimate its frequency and the frequency slope: the quadratic interpolation method and the reassignment method. The largest peak is modeled by $A \cdot e^{j\theta} e^{j\phi(n)}$ of which $\phi(n)$ is simply the accumulation of the estimated peak frequencies. The optimal amplitude A and phase θ should meet the demand that the error between the original signal and the processed signal is minimized. Given a signal $s(n)$ and a window $w(n)$, we define the error R to be

$$\sum_n |s(n)w(n) - w(n) \cdot Ae^{j\theta} e^{j\phi(n)}|^2 \tag{3.1}$$

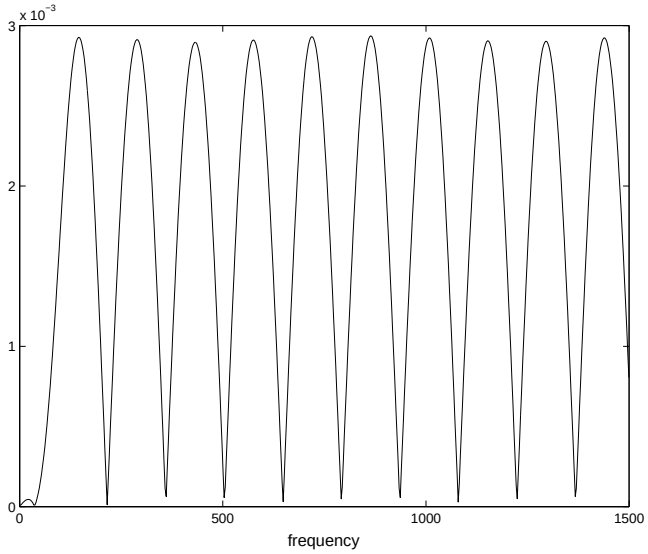

Figure 3.1: *F0*=144 Hz ideal model

and calculate the optimal amplitude and phase by setting $\frac{\partial R}{\partial \theta} = 0$ and $\frac{\partial R}{\partial A} = 0$. Then we have

$$\begin{aligned} \theta &= \frac{1}{2} angle\{\frac{\sum_n [|w(n)|^2 s(n) e^{-j\phi(n)}]}{\sum_n [|w(n)|^2 s^*(n) e^{j\phi(n)}]}\} \\ A &= \frac{\sum_n \{|w(n)|^2 \mathcal{R}e[s(n) e^{-j\theta} e^{-j\phi(n)}]\}}{\sum_n |w(n)|^2} \end{aligned} \tag{3.2}$$

The method which removes less the total energy is chosen to detect hidden peaks. A comparison of the spectrum before and after formant removal, $Spec$ and $Spec_{error}$ respectively, is shown in Fig. 3.2.

3.3 Peak extraction

The peaks in the observed spectrum are extracted based on three principles:

1) In $Spec_{error}$, peaks with magnitudes larger than those of their neighboring frequency bins are chosen.

2) Each peak chosen should meet the constraint that the minimum difference bewteen its corresponding magnitudes of $Spec$ and $Spec_{error}$ is 4 dB.

3) The peaks resolved but situated no further than half the width of a model peak from the largest peak are excluded. The largest peak is finally added back to the group of the extracted peaks.

After extracting all the peaks, their magnitudes are assigned to the corresponding magnitudes in the original spectrum.

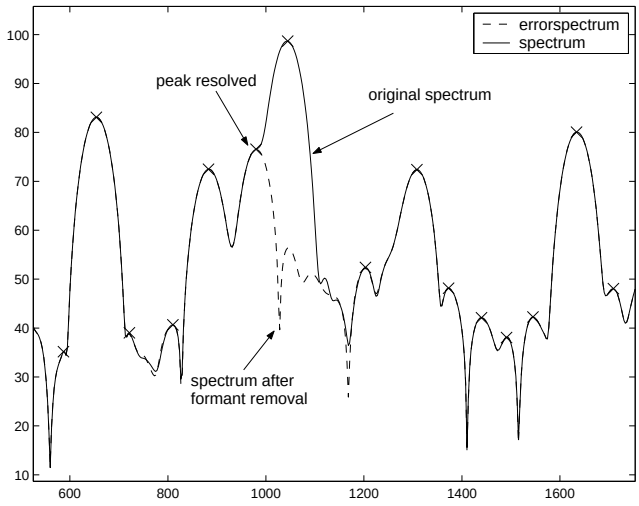

Figure 3.2: Spectrum before and after formant removal

3.4 Peak influential region assignment

To observe the spectral properties around each peak, it is necessary to define an influential region for each peak. The influential region of each peak is the combination of its influential region to the left, m, with that to the right, M. If one peak is well separated from its neighboring peak[1], we assign $m = BW$(or $M = BW$), where BW is one half of the mainlobe size of the analysis window. If one peak is not well separated from its left neighboring peak, we assign its left influential region using the following rule:

$$\begin{aligned} &\text{if } distleft_i > BW \\ &m = \frac{Peak_i - Peak_{i-1}}{Peak_i + Peak_{i-1}} \cdot BW + \frac{Peak_{i-1}}{Peak_{i-1} + Peak_i} \cdot distleft_i \\ &\text{otherwise} \\ &m = \frac{Peak_i}{Peak_{i-1} + Peak_i} \cdot distleft_i \end{aligned} \tag{3.3}$$

where $Peak_i$ denotes the spectral magnitude of the ith observed peak, and $distleft_i$ denotes the distance to its left neighboring peak. And the same way can be applied to defining its influential region to the right, M. In this way, a larger peak will assign a larger influential region.

3.5 The distance function for estimating single $F0$

The main principle of the distance function is to evaluate the designed criteria which compare the properties of each candidate with those of the observed spec-

[1]Its distance to a neighboring peak is larger than the mainlobe width of the analyzing window

trum. After adjusting different weightings on these criteria, we sum the weighted components, for each candidate, to obtain the final "distance", of which the minimum corresponds to the most probable $F0$ among others. For estimating single $F0$, it is formulated as follows:

$$Dist_{F0} = \frac{1}{\sum_{j=1}^{4} p_j}(p_1 \cdot Corr_{F0} + p_2 \cdot Shift_{F0} + p_3 \cdot HLR_{F0} + p_4 \cdot Dev_{F0}) \quad (3.4)$$

where p_1, p_2, p_3 and p_4 are weightings to adjust the importance of each of the evaluated criterion, and $Corr_{F0}$, $Shift_{F0}$, HLR_{F0} and Dev_{F0} are the four components which will be explained in the following paragraphs.

For each $F0$ model, we divide the observed peaks into peaks explained and peaks not explained by its harmonic model peaks according to:

$$s_{F0,i} = \frac{|f_{peak_i} - F0 \cdot round(\frac{f_{peak_i}}{F0})|}{\alpha \cdot F0} \quad (3.5)$$

where α controls the range of explaining one peak, and f_{peak_i} is the frequency of the ith observed peak. A peak is explained if $s_{F0,i} < 1$. This criterion for choosing explained peaks is a key component in the evaluation of the distance function.

Discussion I

Since the calculation of the distance components is considered for the peaks explained by each model, the peak-explaining range should be properly defined for multiple $F0$s estimation since peaks related to different $F0$s are mixed in the spectrum. A setting of $\alpha = 0.4$ in the previous version of `f0` works well for monophonic signals even though a higher $F0$ candidate does include more unrelated peaks. But for polyphonic signals, α should be modified such that a $F0$ model not only includes all the related peaks but also excludes the unrelated ones as many as possible. A theoretical α should be about $0.0293(\sqrt[24]{2} = 1.0293)$ to be able to distinguish among notes a half tone apart, and thus the theoretical range of explanation should be around $2 \cdot 0.029 \cdot h \cdot F0$ where $h \cdot F0$ represents the frequency of the hth model peak. Around the *18*th partial, the neighboring harmonics are to be explained$(0.029 \cdot 18 > 0.5)$, too. Therefore, the limit of the range of explanation should be set as $min(2 \cdot 0.029 \cdot h \cdot F0, F0)$.

The first component – $Corr_{F0}$

$Corr_{F0}$ is designed as an estimation of the "sinusoidality" by comparing the similarity between the peaks explained and an ideal model peak. By taking into consideration only the peaks that are explained, this criterion penalizes especially the higher harmonics of a correct $F0$ since they could never explain the whole spectrum as much as the correct $F0$ does.

The correlation between each peak explained and the window spectrum is calculated within the 6dB bandwidth, δ, of the analysis window. For the ith observed peak explained by one $F0$ model, $Peak_{F0,i}(k)$, its center is first lined with the center of the window spectrum $Win(k)$ and we assign the correlation value

$$c_{F0,i} = \frac{\sum_{k \in \delta} Win(k) \cdot Peak_{F0,i}(k)}{\sqrt{\sum_{k \in \delta} |Win(k)|^2 \cdot \sum_{k \in \delta} |Peak_{F0,i}(k)|^2}} \quad (3.6)$$

to all frequency bins within its influential region, Δ_i. Notice that $Win(k)$ and $Peak_{F0,i}(k)$ are of complex values. Thus, an explained peak of non-constant phase will decrease $c_{F0,i}$ and it attenuates $c_{F0,i}$ of non-sinusoidal peaks. For the influential regions without any peak explained, the correlation values are assigned to 0. Then the first distance component is defined as:

$$Corr_{F0} = \sum_k [1 - c_{F0}(k)^{\delta 1}] \cdot |Spec(k)| \tag{3.7}$$

where $\delta 1 = 0.5$.

Discussion II

In estimating $F0$ of a monophonic signal, the local minima of the first distance component occur around $\frac{m}{n} \cdot F0$ of which m and n are positive integers. That is, the higher-harmonics and subharmonics of one $F0$ tend to explain well of the observed peaks. The complex product of $Win(k)$ and $Peak_{F0,i}(k)$ attenuates the correlation of the peaks with smaller magnitudes, non-symmetrical shapes or varying phases, such that the explanation of a well-formed peak has a higher advantage than that of a badly-formed peak of the same amplitude.

The second component-$Shift_{F0}$

The model with a smaller total displacement from its explaining peak sequence implies better explanation.

For the ith observed peak, it is explained by one $F0$ model peak if $s_{F0,i} < 1$. Otherwise, we set $s_{F0,i}$ to 1. Then $s_{F0,i}$ is weighted by

$$h_{F0,i} = min(\frac{1}{n \cdot F0}, \frac{1}{f_{peak_i}}) \tag{3.8}$$

By choosing the minimum between the two weightings, the shift values of the model peak with frequency larger than $n \cdot F0 (n \in \mathcal{N})$ are attenuated. In single $F0$ estimation, $n = 1$ is evaluated to perform well for speech signals. This weighting is based on the fact that higher harmonics usually deviate more from its theoretical position due to inharmonicity and thus their shift values are less reliable.

The second distance component is then formulated as

$$Shift_{F0} = \sum_k \frac{c_{F0,i}^{\delta 1} \cdot h_{F0,i}^{\delta 2} \cdot s_{F0,i} \cdot |Spec(k)|}{\sum_k [c_{F0,i}^{\delta 1} \cdot h_{F0,i}^{\delta 2} \cdot |Spec(k)|]} \tag{3.9}$$

where $s_{F0,i} = 1$ for the peaks not explained. $\delta 1$ and $\delta 2$ are parameters to be evaluated by the evolutionary algorithm on the database "Bagshaw"[13]. The weighting of the spectral magnitudes is meant to emphasize $s_{F0,i}$ values of predominant peaks. The weighting of $c_{F0,i}^{\delta 1}$ attenuates $s_{F0,i}$ of less-correlated peaks and the weighting of $h_{F0,i}^{\delta 2}$ attenuates the shift values for higher frequency components.

Discussion III

The exponential weighting factor $\delta 2$ emphasizes more on $F0$ itself and should be treated with care. For signals with relatively small magnitude at a correct

$F0$ (as found in the sounds of several musical instruments), $Shift_{F0}$ fails if $\delta 2$ is set too large and the effect of magnitude weighting will be too strong. To avoid problems that strong higher-harmonics might perform better than $F0$ in $Shift_{F0}$, we could either use $n \approx 5$(according to the discussion in section 1.1) or choose a $\delta 2$ which emphasizes the shift vaules of some of the first partials.

For polyphonic signals, summing up all explained peaks in calculating $Shift_{F0}$ might tend to include the peaks better explained by other $F0$ candidates if α is not properly set. A constant range of $2 \cdot 0.4 \cdot F0$ as used in the previous work[13] apparently includes many unrelated peaks.

The third component-HLR_{F0}

The partials of a correct $F0$ usually form a smoother spectrum envelope vector than those of its subharmonic partials do. The third distance component aims to attenuate the competing performance of subharmonics of a correct $F0$ in the preceeding two distance components.

For the hth model peak to explain certain observed peaks, $a_{F0,h}$ is defined as its corresponding magnitude in the observed spectrum. For the model peak to explain none of the observed peaks, its magnitude is assigned to 0. One spectrum envelop vector a_{F0} is constructed by the sequential ensemble of these magnitudes. Then we flip a_{F0} around zero and combine it with the original a_{F0} to form a new vector A_{F0}. After high-pass filtering A_{F0} obtains A'_{F0}, we define the third distance component as

$$HLR_{F0} = \frac{\sum_{h=1}^{2H} |A'_{F0,h}|^2}{\sum_{h=1}^{2H} |A_{F0,h}|^2} \tag{3.10}$$

where H denotes the number of harmonic partials of one $F0$ model. An example comparing the variation of spectrum envelop between $F0$ and $F0/2$ is shown in Fig.3.3.

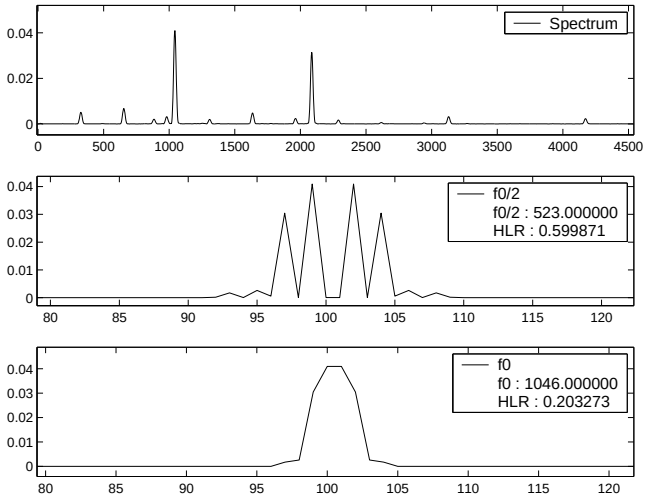

Figure 3.3: The variation of spectral amplitude vector A_{F0}

Discussion IV
Although this component seems to work well for monophonic signals, a more reasonable assignement of $a_{F0,h}$ is the magnitude of the nearest explained peak. However, it is possible that an unrelated peak is chosen as the nearest peak. Then, the result will be less reliable.

The fourth component-Dev_{F0}

Partials of one instrument should have similar time evolution of amplitudes. By observing the variance of mean time for the harmonic group of each peak, we could obtain smaller variance for those $F0$s with most sinusoidal-like partials and could be used to punish the candidates corresponding to noise peaks.

The deviation of mean time is formulated as

$$Dev_{F0} = \sqrt{Var_{F0}} = \sqrt{\sum_{i \in \text{allpeaks}} \{[\hat{t}_{F0,i} - \sum_{j \in \text{allpeaks}} (w_{F0,j} \cdot \hat{t}_{F0,j})]^2 \cdot w_{F0,i}\}} \tag{3.11}$$

where $\hat{t}_{F0,i}$ denotes the mean time of the ith observed peak. Mean time is an indication of the center of gravity in the time domaine. Its values other than zero indicate that the corresponding $F0$ deviates from the center of the analysis window. For the jth peak, $w_{F0,j}$ is a weighting factor defined as

$$w_{F0,j} = \frac{|Peak_{F0,j}|^2 \cdot (1 - s_{F0,j})}{\sum_{j \in \text{allpeaks}} [|Peak_{F0,j}|^2 \cdot (1 - s_{F0,j})]} \tag{3.12}$$

and is used to attenuate the weightings of those peaks deviating too much from the window center. For the peaks not explained by one $F0$, they are excluded since $s_{F0,j} = 1$. Dev_{F0} is then normalized by half of the window size to be used as the fourth distance component.

Discussion V
For modulated signals, Dev_{F0} will not always give a minimum among other candidates.

3.6 Final judgement on *F0*

Combining the four distance components multiplied by different weightings which are evaluated by the evolutionary algorithms, the $F0$ with the smallest "distance" is chosen as the final output of the single fundamental frequency estimation.

Chapter 4

Algorithms proposed for estimating multiple $F0$s

The creteria in evaluating single $F0$ are designed under the hypotheses of monophonic cases, and they evaluate the individual properties of each $F0$ candidate. To estimate multiple $F0$s, the distance function should be modified to observe not only the individual properties of each $F0$ candidate but also the combined properties of $F0$ candidates. To this end, especially the first two distance components will need to be changed to evaluate the properties of the explained peaks by a combination of $F0$ candidates. To reduce the number of possible combinations among $F0$ candidates, we propose a sequence of strategies to eliminate unwanted $F0$ candidates such as the subharmonics corresponding to correct $F0$s and the higher-harmonics corresponding to correct $F0$s.

4.1 First selection of $F0$ candidates

In general, the peaks in the spectrum are regarded as top $F0$ candidates. Thus, we start with observing four time-frequency descriptors of all the peaks and use them as measures to get rid of non sinusoidal-like peaks.

The four descriptors observed are **duration**, **bandwidth**, **group delay** and **instantaneous frequency**. Different thresholds for the four properties are set in advance. For each peak, if two of the four descriptor values fall within the preset thresholds, it is chosen as a $F0$ candidate. The following explanations of the four descriptors are based on Cohen's book[14].

Duration

Consider the signal expressed in the frequency domain $S(\omega) = |S(\omega)| \cdot e^{j\psi(\omega)}$.We denote the spectral region around the ith observed peak as $IReg_i$. The mean time of the ith observed peak, $\langle t_i \rangle$, could be calculated as follows:

$$\langle t_i \rangle = -\int_{\omega \in IReg_i} \psi'(\omega) |S(\omega)|^2 d\omega \tag{4.1}$$

which indicates where the energy of the ith peak is concentrated in the window. As explained in Cohen's book, we could calculate the duration of the ith peak

as

$$T_i^2 = \int_{IReg_i} \Big\{ [S'(\omega)]^2 + [\langle t_i \rangle + \psi'(\omega)]^2 |S(\omega)|^2 \Big\} d\omega \tag{4.2}$$

which describes to what extend the energy of the ith peak is concentrated around the average.

Bandwidth

The energy density spectrum tells us which frequencies existe during the total duration of the signal. If $|S(\omega)|^2$ represents the energy density in frequency, we have the average frequency for the ith peak

$$\langle \omega_i \rangle = \int_{IReg_i} \omega \cdot |S(\omega)|^2 d\omega \tag{4.3}$$

and bandwidth[1]

$$B_i^2 = \int_{IReg_i} (\omega - \langle \omega_i \rangle)^2 \cdot |S(\omega)|^2 d\omega \tag{4.4}$$

which is an indication of the energy spread in frequencies for the duration of the ith peak. We further normalize B_i by dividing the width of $IReg_i$.

Group delay

As expressed in eq.(4.1), we could consider group delay, $-\psi(\omega)'$, as an indication of the average time for a particular frequency. In the reassignment methods proposed in [15], group delay could be expressed in terms of the short time Fourier transform, $STFT$, of the signal:

$$-\psi'(\omega) = -\mathcal{R}e\{\frac{STFT_{h_T}(t,\omega) \cdot STFT_h^*(t,\omega)}{|STFT_h(t,\omega)|^2}\} \tag{4.5}$$

where $h(t)$ is the analysis window and $h_T(t) = t \cdot h(t)$.

Instantaneous frequency

Consider the signal expressed in complex form $s(t) = A(t) \cdot e^{j\phi(t)}$ where $A(t)$ is its amplitude and $\phi(t)$ is its phase. If $|s(t)|^2$ represents the energy density in time, mean frequency can be further expressed as

$$\langle \omega \rangle = \int \phi'(t) |s(t)|^2 dt \tag{4.6}$$

and we obtain the instantaneous frequency $\omega_i(t) = \phi'(t)$.

Similar to the calculation of group delay, the instantaneous frequency could be expressed in terms of the short time Fourier transform of the signal:

$$\omega_i(t) = \mathcal{I}m\{\frac{STFT_{h_D}(t,\omega) \cdot STFT_h^*(t,\omega)}{|STFT_h(t,\omega)|^2}\} \tag{4.7}$$

where $h_D = \frac{\partial h(t)}{\partial t}$.

[1] Root mean square bandwidth

4.2 Second selection

The second selection aims at eliminating the candidates corresponding to subharmonics of correct $F0$s. The third distance component HLR_{F0} plays an important role at this stage.

In the previous version of `f0`, the amplitude vector A_{F0} is high-pass filtered at a prefixed frequency and this does not provide the variation information below the prefixed frequency. Therefore, we modified HLR_{F0} in a way that the *bandwidth* of the amplitude vector A_{F0} is evaluated instead of high-pass filtering it at a prefixed frequency. A threshold is set to preserve those $F0$s of small HLR_{F0} values, that is, more energy concentrates in the low frequency region. This threshold will be optimized by the evolutionary algorithm.

After selecting $F0$ candidates with smooth spectral envelopes, we further filter those candidates of which the explaining partials deviate too much from the center of the analysis window. A threshold is set to filter the candidates with $Dev_{F0} > 0.1$.

4.3 Harmonic grouping

To reduce the number of $F0$ candidates corresponding to higher-harmonics of correct $F0$s, we categorize the candidates into different harmonic groups and in each group only the $F0$ of the smallest distance value is chosen. The distance function applied here excludes $Corr_{F0}$ in the previous version and is formulated as:

$$Dist_{F0} = \frac{1}{\sum_{j=2}^{4} p_j}(p_2 \cdot Shift_{F0} + p_3 \cdot HLR_{F0} + p_4 \cdot Dev_{F0}) \tag{4.8}$$

Since $F0$s in musical signals are somewhat harmonically related, there is ambuiguity of grouping a harmonic partial related to more than one $F0$. Consider two notes with fundamental frequencies $F0_1$ and $F0_2$, both explaining the other candidate $F0_3 \approx n \cdot F0_1 \approx m \cdot F0_2$, i.e., $\frac{F0_1}{F0_2} \approx \frac{m}{n}$ where m and n are both positive integers. Using $s_{F0,i}$ as an indication of how well one peak is explained by one $F0$, we assign this peak at $F0_3$ to the $F0$ group with the smallest $s_{F0,i}$ value.

However, annoying problems may occur in situations where:

1) If one lower-frequency peak(either a subharmonic of a correct $F0$ or simply a noise peak) is not filtered out in the previous stages, there is possibility that it includes more than one correct $F0$ in the same group.

2) Modulated $F0$s with largely deviated partials produce less reliable $Shift_{F0}$ values.

To ensure $Shift_{F0}$'s performance in the situation mentioned above, we could reassign a $F0$ candidate to its right-neighboring frequency(in the same peak influential region) with the minimum sum of $Shift_{F0}$ and HLR_{F0}. This correction could ensure each $F0$ candidate to explain the most of its corresponding partials in the observed spectrum because inharmonicity appears as a rising tendancy in the series of partials.

An example of the distance components is shown in Fig.4.1. This is a mixture of three notes played by the cello, the piano and the oboe with $F0$s at 328Hz, 885Hz and 1047Hz. The cross signs represent $F0$ candidates and those $F0$s circled are the candidates selected by HLR_{F0} and Dev_{F0} at the previous stage.

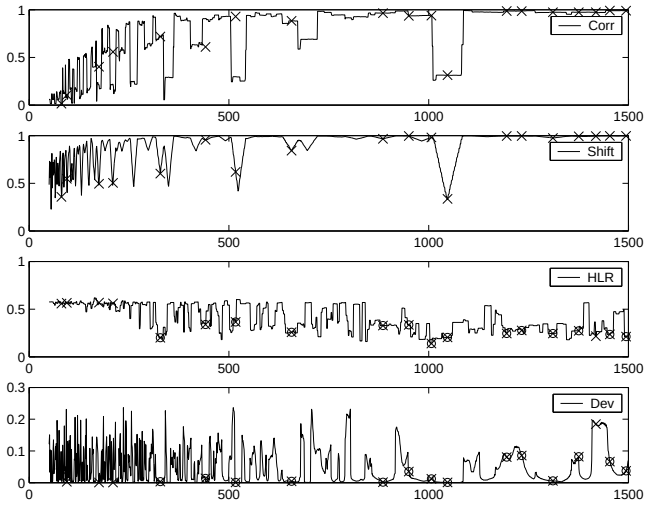

Figure 4.1: Distance components

4.4 The distance function for estimating multiple $F0$s

Assume the number of notes is known to be N_{note} and the number of final $F0$ candidates is M_{cadt}. There are $\begin{pmatrix} M_{cadt} \\ N_{note} \end{pmatrix}$ combinations, one of which will explain best the observed spectrum. To compare the combined properties of the final candidates, we construct a new distance function as

$$Dist_{F0}^{N_{note}} = \frac{1}{\sum_{j=5}^{8} p_j}(p_5 \cdot ErrSpec_{F0}^{N_{note}} + p_6 \cdot Shift_{F0}^{N_{note}} + \sum_{N_{note}} [p_7 \cdot HLR_{F0} + p_8 \cdot Dev_{F0}]) \tag{4.9}$$

where HLR_{F0} and Dev_{F0} are as explained before. Different from HLR_{F0} and Dev_{F0} which are calculated individually for each $F0$ candidate, $ErrSpec_{F0}^{N_{note}}$ and $Shift_{F0}^{N_{note}}$ compare the properties of a combined model spectrum with those of the observed spectrum.

The first component – $ErrSpec_{F0}^{N_{note}}$

A combination of all correct $F0$s should explain the most of the spectral magnitudes in a mixed spectrum. $ErrSpec_{F0}^{N_{note}}$ is an error rate of the spectral

magnitude sum of a combined model spectrum, compared to that of the observed spectrum. It is formulated as:

$$ErrSpec_{F0}^{N_{note}} = 1 - \frac{\sum_k |Spec_{F0}^{N_{note}}(k)|}{\sum |Spec(k)|} \tag{4.10}$$

where $Spec_{F0}^{N_{note}}$ is obtained by combining the spectral magnitudes of N_{note} model spectrums[2]. For the spectral regions explained by more than one $F0$ candidates, the average values are assigned to the corresponding magnitudes.

A graphical example is shown in Fig. 4.2. This is a polyphonic signal with $F0_1$ = 328Hz(cello), $F0_2$ = 885Hz(piano) and $F0_3$ = 1047Hz(oboe), which correspond to the notes E4, A5 and C5, respectively. Notice the calculation in $ErrSpec_{F0}^{N_{note}}$ uses the linear spectral magnitude, while the graphical example uses log magnitude to show more clearly how a model spectrum fits into the observed spectrum.

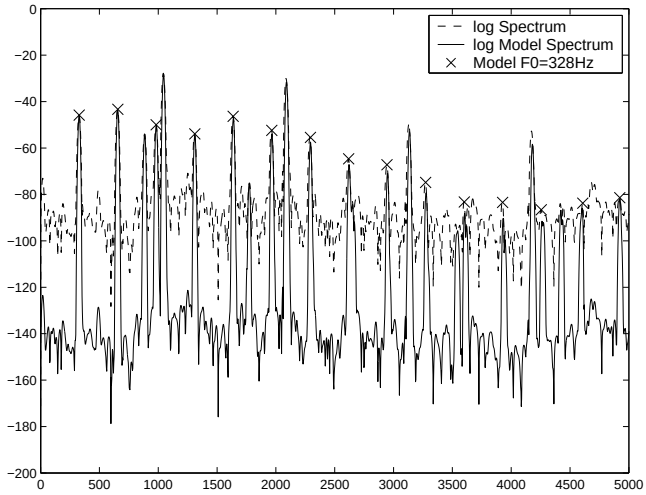

Figure 4.2: Log spectrum of the combined model

The second component – $Shift_{F0}^{N_{note}}$

Assume $\{F0_{t^1}, F0_{t^2}, \ldots, F0_{t^{N_{note}}}\}$ is one of the possible combinations of all $F0$ candidates. To explain the ith observed peak by a combined sequence of model peaks, the minimum shift value among $\{s_{F0_j,i}\}_{j=t^1,t^2,\ldots,t^{N_{note}}}$ is assigned to the shift value, $s_{F0,i}^{N_{note}}$, of a combined model spectrum. The weighting factor in eq.(3.8) is modified to be

$$h_{F0,i}^{N_{note}} = min\Big(\frac{1}{n \cdot max(F0_{j=t^1,t^2,\ldots,t^{N_{note}}})}, \frac{1}{f_{peak_i}}\Big) \tag{4.11}$$

[2]Individual model spectrum is obtained by multiplying the observed spectral magnitude with an ideal spectral model as shown in Fig. 3.1.

around the ith peak. n is used to attenuate $Shift_{F0}^{N_{note}}$ after the nth harmonics. Then we define the second component as

$$Shift_{F0}^{N_{note}} = \sum_{k} \frac{c_{F0,i}^{\delta 1} \cdot (h_{F0,i}^{N_{note}})^{\delta 2} \cdot s_{F0,i}^{N_{note}} \cdot |Spec(k)|}{\sum_k [c_{F0,i}^{\delta 1} \cdot (h_{F0,i}^{N_{note}})^{\delta 2} \cdot |Spec(k)|]} \tag{4.12}$$

where

$$s_{F0,i}^{N_{note}} = \begin{cases} min\big((s_{F0_j,i})_{j=t^1,t^2,\ldots,t^{N_{note}}}\big), & \text{for peaks explained} \\ 1, & \text{otherwise} \end{cases} \tag{4.13}$$

and $\delta 1 = 0.5$.

4.5 Final judgement on multiple $F0$s

After the distance values of all combinations of $F0$ candidates are calculated, the combination with the smallest "distance" is chosen as the final output of the multiple fundamental frequencies estimation.

Chapter 5

Testing results and discussions

The musical instrument samples used are produced by the electronic music studios of Iowa University. The samples chosen to be tested include samples of several musical instruments: alto saxphone(with vibrato or without vibrato), contrabass(bowed or plucked), bassoon, Bb clarinet, cello(bowed or plucked), flute, french horn, oboe, piano(non-anechoic) and tenor trombone. By randomly combining these samples, we construct a database of polyphonic music samples to test our algorithms. Some specifications in this test are:

1) The range of notes chosen is from C2(65Hz) to B6(1980Hz). In combining a N_{note} voices polyphonic samples, N_{note} of twelve(C, Db, D, Eb, E, F, Gb, G, Ab, A, Bb, B) different note names are chosen first to ensure no octaves-related notes are mixed in one sample. For each note chosen, a musical instrument is randomly selected. Notice that instruments played by different techniques are regarded as different instruments in this test. Then, notes of different amplitudes(*ff, mf, pp*) and different octaves are randomly chosen to be combined into a polyphonic sample.

2) The stationary parts of the samples are preselected to combine a polyphonic sample. In each combination of notes, samples are normalized to have an equal sum of spectral magnitudes before mixing.

3) The $F0$ estimation range is set from 50Hz to 2000Hz and the maximum analyzing frequency is $5k$Hz. A large window size of 4500 samples and $2 \cdot 8192$ points FFT are used to dectect a $F0$ as low as 50Hz and to obtain a higher frequency resolution to resolve close $F0$ peaks.

4) The number of samples used in this test is 100, including equal amount of samples for one-voice signals(25 notes), two-voice mixing signals(50 notes), three-voice mixing signals(75 notes) and four-voice mixing signals(100 notes).

We rewrite eq.(4.8)

$$Dist_{F0} = \frac{1}{\sum_{j=2}^{4} p_j}(p_2 \cdot Shift_{F0} + p_3 \cdot HLR_{F0} + p_4 \cdot Dev_{F0})$$

and eq.(4.9)

$$Dist_{F0}^{N_{note}} = \frac{1}{\sum_{j=5}^{8} p_j}(p_5 \cdot ErrSpec_{F0}^{N_{note}} + p_6 \cdot Shift_{F0}^{N_{note}} + \sum_{N_{note}} [p_7 \cdot HLR_{F0} + p_8 \cdot Dev_{F0}])$$

The parameters to be evaluated by the evolutionary algorithms are p_2, p_4, p_5, p_6, p_8, with $p_3 = 15$ and $p_7 = 1$ fixed beforehand. The other three parameters to be evaluated are α in eq.(3.5) which controls the range of explaining peaks for each model peak. n in eq.(3.8) as well as in eq.(4.11), and $\delta 2$ in eq.(3.9) as well as in eq.(4.12), both emphasize the weightings on shift values of the first several partials .

5.1 Testing results

The following tables show some of the results obtained. The first column lists the sets of parameters for the two distance functions. The second column lists n which determines the attenuation of the shift values after the nth partials. The third column lists the threshold used for HLR_{F0} in the second selection stage. The columns "ONE", "TWO", "THREE" and "FOUR" show the numbers of wrongly estimated notes in the four data sets. The last column lists the error rates.

$(p_2, p_4, p_5, p_6, p_8)$	n	thrshd3	α	$\delta 2$	ONE	TWO	THREE	FOUR	error
$(35, 30, 14, 35, 7)$	3	0.49	0.029	1.5	1/25	6/50	13/75	44/100	25.6%
$(23, 23, 16, 35, 8)$	3	0.47	0.035	1.5	0/25	4/50	17/75	50/100	28.4%
$(15, 19, 12, 40, 6)$	2	0.45	0.027	1.5	1/25	8/50	14/75	51/100	29.6%
$(19, 33, 10, 40, 6)$	1	0.47	0.035	2.5	0/25	8/50	17/75	50/100	30.0%
$(19, 35, 20, 50, 10)$	1	0.4	0.035	4.5	0/25	10/50	18/75	52/100	32.4%

The randomly mixed musical samples often include samples of which $F0$s are multiples of one another. Lower notes tend to group higher notes because higher harmonics of lower notes usually situate in the same peak-explaining range of higher notes. Especially for the data set "FOUR", there are quite a few samples mixed by $F0$s that are likely to be grouped together. After tracing the errors using the first set of parameters and do not consider the errors due to $F0$s which are multiples of one another, we obtain an error rate **17.7%**[1].

Another result is shown in the next table. In this second test, several modifications are made:

1) In harmonic grouping, additional better $F0$ candidates are kept of which the number is the floor integer of the number of $F0$s(in one group) divided by 5. The range of explaining peaks is also limited to a maximum of $2 \cdot 0.4 \cdot F0$ instead of $2 \cdot 0.5 \cdot F0$.

2) The amplitude $a_{F0,h}$ of each model peak is assigned the magnitude of the nearest peak in each peak-explaining region.

[1] The error rates for the four data sets are 4%, 8%, 17% and 27%, respectively.

$(p_2, p_4, p_5, p_6, p_8)$	n	thrshd3	α	$\delta 2$	ONE	TWO	THREE	FOUR	error
$(23, 21, 20, 25, 11)$	4	0.51	0.029	1	1/25	3/50	14/75	41/100	23.6%
$(33, 21, 18, 20, 12)$	4	0.51	0.029	2	1/25	4/50	13/75	42/100	24.0%
$(33, 7, 18, 45, 18)$	4	0.49	0.033	1.5	1/25	5/50	16/75	42/100	25.6%
$(23, 13, 20, 20, 16)$	3	0.49	0.029	1.5	1/25	5/50	14/75	41/100	25.6%
$(33, 21, 14, 20, 12)$	4	0.51	0.031	2	1/25	4/50	16/75	44/100	26.0%

After tracing the errors using the first set of parameters and do not consider the errors due to $F0$s which are multiples of one another, we obtain an error rate **16.1%**[2].

5.2 Discussions

5.2.1 The parameters

1) In the previous version of `f0`[13], $h_{F0,i}$ emphasizes more on $F0$ itself with $n = 1$ and $\delta 2 = 4.5$. In this test, better performances are obtained when $n > 1$, that is, we emphasize the first several partials. This coincides with our preceeding discussions since the brass instruments often have a weaker $F0$ and it is reasonable to weight more on the first several partials.

2) The weighting $\delta 2$ seems to work better for values smaller than 4.5 obtained from testing speech signals in the previous work. The balance between n and $\delta 2$ could be optimized for each data set but a best value for all requires testing for more polyphonic signals.

3) The value of α larger than the theoretical value 0.029 sometimes improve the performance. This might be due to the fact that when there is a strong modulation such as in bowed strings, higher partials tend to deviate more than one half tone from the theoretical frequencies and a larger α could ensure related peaks to be included. However, for musical signals mixed with more than four notes, a larger α will cause more unrelated peaks to be explained by each $F0$ candidate.

5.2.2 Examination of the proposed algorithms

First selection

Since the window size is as large as 4500, the number of unwanted peaks, such as noise peaks and sidelobe peaks, are not effectively reduced. A compromise of reducing the size of the analysis window is not feasible because errors do occur when two adjacent notes(one half tone apart) of different instruments are not resolved. Before applying a larger window, we should improve the robustness of separating sinusoidal-like peaks from non sinusoidal-like peaks.

Second selection

As explained in the previous chapter, this selection aims to filter subharmonics of correct $F0$s. However, the threshold set for HLR_{F0} is case sensitive and

[2]The error rates for the four data set are 4%, 6%, 17% and 23%, respectively.

causes correct $F0$s to be filtered in a few tests. For reed instruments such as clarinets, the even harmonics are weaker than odd harmonics and thus there are more variations in the spectral amplitude vector A_{F0} which is not fair compared with other instruments.

Harmonic grouping

Proper harmonic grouping strongly depends on the performance of the first two selections. A lower-frequency noise peak will group correct $F0$s together and we will always lose some correct $F0$s because only the best candidate in one group is kept to the next stage.

Since we are dealing with $F0$s that are not multiples of one another, this strategy works to reduce the number of $F0$ candidates a lot. However, in music, $F0$s are mostly multiples of one another and it will be difficult to examine how many correct $F0$s are grouped together.

The distance function

$ErrSpec_{F0}^{N_{note}}$ fails in cases where a predominant partial exists and competes with the correct $F0$. The spectral model is constructed from equally-spaced hamonics which often deviate from the observed partials. Once a correct $F0$ model explains a small part of that predominant partial, it fails to compete with some higher-harmonic $F0$ candidate which explains more energy of the predominant partial.

As long as correct $F0$s are kept to compete at the last stage, the distance function performs well in current consideration. Notice that all samples are normalized before being mixed such that $ErrSpec_{F0}^{N_{note}}$ and $Shift_{F0}^{N_{note}}$ work well. For real musical signals in which notes of different levels are mixed, it is difficult to extract a relatively weak $F0$ since spectral magnitudes are introduced in the two combined properties.

Chapter 6

Conclusions and future work

Estimating multiple fundamental frequencies in musical signals has always been an important research topic in music technology. Considerations for single $F0$ estimation no longer meet the demand of estimating multiple $F0$s. A "blackboard" system is necessary to include signal processing knowledge as well as musical knowledge such that we could manage to model the estimation system. In `f0`, many signal processing techniques are applied to detecting hidden peaks, to estimating correct peak frequencies, to constructing ideal models, to observing time-frequency properties, etc. On the other hand, musical knowlege is not yet integrated while it is the key to accomplishing this complicated task. More information should be taken into account:

- **Spectrum irregularities**: Spectrum irregularities include inharmonicity, relative strength between even harmonics and odd harmonics, formant characteristics and modulated partials.
- **Onsets of notes**: The onsets of notes play an important role in identifying groups of partials and estimating the number of notes.

Integrating signal processing techniques with musical knowledge, it would be possible to classify peaks into independent partials, overlapped partials and noise peaks. Independent partials provide cues to overlapped partials. Overlapped partials should be resolved to obtain more partials. A possible modification of harmonic group assignment is to examine the partials in two ways: from low to high and from high to low. Onset information will be essential to estimate $F0$s hidden in overlapped partials. That is, we should observe the spectral evolution instead of concentrating on the current frame only.

In polyphonic musical signals, notes are usually of different amplitudes. Weighting the linear spectrum in the distance components attenuates relatively weak $F0$s while emphasizes relatively strong partials belonging to other $F0$s. Weighting a log-scaled spectral magnitudes might be a way to attenuate the degree of difference in the linear spectrum.

There are many parameters in the distance functions to be tested by the evolutionary algorithm. However, a more logical way to automatically adjust the parameters could save lots of testing work and also ensure an optimization for all cases. A statisical modeling method will be a good choice because the

parameters introduced in the model could be optimized based on the EM algorithm, for example. The difficulty lies in applying reasonable prior probability distributions to the parameters considered.

Although many problems with estimating $F0$s of real world polyphonic signals are not yet solved, we do have the following achievements:

1) HLR_{F0} is an effective strategy to punish subharmonics of correct $F0$s.

2) Reducing a great amount of $F0$ candidates by harmonic grouping eases the calculation work in the final stage.

3) The distance function performs 84% correct estimation on the database mixed from a variety of musical instrument samples.

Appendix A

EM algorithm

A.1 Bayesian statistics

Bayesian statistical methods provide a complete paradigm for both statistical inference and decision making under uncertainty. The Bayesian paradigm is based on an interpretation of probability as a *conditional measure of uncertainty.* Statistical inference about a quantity of interest is described as the modification of the uncertainty about its value in the light of evidence, and Bayes' theorem specifies how this modification should be made.

The Bayes' theorem is written as:

$$p(\theta|X,\mathcal{I}) = \frac{p(X|\theta,\mathcal{I})p(\theta|\mathcal{I})}{p(X|\mathcal{I})} \tag{A.1}$$

where $\mathcal{I}$ represents all prior information and assumptions about the model, $p(X|\mathcal{I})$ is the *evidence*, which may generally be regarded as a nomalising factor, $p(\theta|\mathcal{I})$ is the *prior* probability density of the parameters before the data are observed, and $p(\theta|X,\mathcal{I})$ is the *posterior* probability density. Bayes' theorem can, therefore, be summarized in the form [16] :

$$posterior = \frac{likelihood \times prior}{normalization\ factor} \tag{A.2}$$

In Bayesian parameter estimation, both the prior and posterior distributions represent, not any measurable property of the parameter, but only our own state of knowledge about it. The width of the distribution indicates the range of values that are consistent with our prior information and data, therefore, which honestly compels us to admit as possible values.

A.2 Prior probability

A central element of any statistical analysis is the specification of a probability model which is assumed to describe the mechanism generating the observed data as a fuction of parameters, sometimes named the *state of nature*, about whose value only limited information is available. Thus, the main problem encountered is to describe all uncertainties by means of probability distributions. In

particular, unknown parameters in probability models must have a joint probability distribution which describes the availible information about their values. This is often regarded as a characteristic element of the Bayesian approach. Parameters are treated as random variables, which is not a description of their variability but a description of the uncertainty about their values.

An important particular case arises when either no relevant prior information is readily available, or that information is subjective and an objective analysis is desired. This is addressed by *reference analysis* which uses information-theoretical concepts to derive appropriate reference posterior distributions, defined to encapsulate inferential conclusions on the quantities of interest solely based on the supposed model and the observed data. It is assumed that probability distributions may be described through their density functions, interperting the propability of an event as a conditional measure of uncertainty, on a $[0, 1]$ scale, about the occurence of the event in some specific conditions. The limiting extreme values 0 and 1, which are typically inaccessible in applications, respectively describe the impossibility and the certainty of the ocurrence of the event.

A.3 General EM algorithm

The Expectation-Maximization algorithm given by Dempster et al.(1977) has gained considerable popularity for soving MAP estimation problems. It is an iterative optimization technique specifically designed for probabilistic models.

EM makes a local approximation that is a lower bound to the objective function. Starting from a current guess of a set of parameters, we choose the new guess to maximize the lower bound that will be an improvement over the previous guess, unless the gradient at the current guess was zero. Therefore the idea is to alternate between computing a lower bound(the "E-step") and maximizing this bound(the "M-step"), until a point of zero gradient is reached.

A.3.1 EM as lower bound maximization

Maximum A-Posteriori(MAP) estimation concerns the maximization of the function

$$f(\theta) = p(\mathbf{X}, \theta) \tag{A.3}$$

where $\mathbf{X}$ is the matrix of observed data. If $f(\theta)$ is a simple function, then its maximum can often be found analytically, `e.g.`, by equating its gradient to zero. Equivalently, we can maximize the logorithm of the joint distribution which is proportional to the posterior [17]:

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \ \log p(\mathbf{X}, \theta) = \underset{\theta}{\operatorname{argmax}} \ \log \sum_{\mathrm{h}} p(\mathbf{X}, \mathrm{h}, \theta) \tag{A.4}$$

The idea behind EM is to start with a guess θ^t for the parameters θ, compute an easily computed lower bound $B(\theta; \theta^t)$ to the function $\log p(\theta, \mathbf{X})$, and then maximize the bound. After iterative computation, this procedure will converge to a local maximizer θ^* of the objective function, provided that the bound improves at each iteration.

Maximizing the above equation is difficult to deal with since it involves the *logarithm of a big sum.* Thus, we construct a tractable lower bound $B(\theta;\theta^t)$ that contains a *sum of logarithms* instead.

To derive the bound, we write $\log p(\mathbf{X},\theta)$ as [18]

$$\log p(\mathbf{X},\theta) = \log \sum_{\mathrm{h}} p(\mathbf{X},\mathrm{h},\theta) = \log \sum_{\mathrm{h}} f^t(\mathrm{h}) \frac{p(\mathbf{X},\mathrm{h},\theta)}{f^t(\mathrm{h})} \tag{A.5}$$

where $f^t(\mathrm{h})$ is an arbitrary probability distribution. The bound used by EM is the following form of Jensen's inequality:

$$\sum_j g(j) a_j \geq \prod_j g(j)^{a_j} \tag{A.6}$$

provided

$$\sum_j a_j = 1,\ a_j \geq 0,\ g(j) \geq 0 \tag{A.7}$$

Thus, we define the bound

$$B(\theta;\theta^t) \triangleq \sum_{\mathrm{h}} f^t(\mathrm{h}) \log \frac{p(\mathbf{X},\mathrm{h},\theta)}{f^t(\mathrm{h})} \leqq \log \sum_{\mathrm{h}} f^t(\mathbf{h}) \frac{p(\mathbf{X},\mathrm{h},\theta)}{f^t(\mathrm{h})} \tag{A.8}$$

A.3.2 Finding an optimal bound

One further step is to find the best bound that touches the objective function $\log p(\mathbf{X},\theta)$ at the current guess θ^t. Intuitively, finding the best bound at each iteration will guarantee an improved estimate θ^{t+1} when we locally maximize the bound with respect to θ. Since $B(\theta;\theta^t)$ is a lower bound, the optimal bound at θ^t can be found by maximizing

$$B(\theta^t;\theta^t) \triangleq \sum_{\mathrm{h}} f^t(\mathrm{h}) \log \frac{p(\mathbf{X},\mathrm{h},\theta^t)}{f^t(\mathrm{h})} \tag{A.9}$$

with respect to the distribution $f^t(\mathrm{h})$. Introducing a Lagrange multiplier λ to enforce the constraint $\sum_{\mathrm{h}} f^t(\mathrm{h}) = 1$, the objective becomes

$$G(f^t) = \lambda[1 - \sum_{\mathrm{h}} f^t(\mathrm{h})] + \sum_{\mathrm{h}} f^t(\mathrm{h}) \log p(\mathbf{X},\mathrm{h},\theta^t) - \sum_{\mathrm{h}} f^t(\mathrm{h}) \log f^t(\mathrm{h}) \tag{A.10}$$

taking the derivative

$$\frac{\partial G}{\partial f^t(\mathrm{h})} = -\lambda + \log p(\mathbf{X},\mathrm{h},\theta^t) - \log f^t(\mathrm{h}) - 1 \tag{A.11}$$

and solving $f^t(\mathrm{h})$ we obtain

$$f^t(\mathrm{h}) \triangleq \frac{p(\mathbf{X},\mathrm{h},\theta^t)}{\sum_{\mathrm{h}} p(\mathbf{X},\mathrm{h},\theta^t)} = \frac{p(\mathbf{X},\mathrm{h},\theta^t)}{p(\mathbf{X},\theta^t)} = p(\mathrm{h}|\mathbf{X},\theta^t) \tag{A.12}$$

By examining the value of the resulting optimal bound at θ^t, we see that it indeed touches the objective function:

$$B(\theta^t;\theta^t) = \sum_{\mathrm{h}} p(\mathrm{h}|\mathbf{X},\theta^t) \log \frac{p(\mathbf{X},\mathrm{h},\theta^t)}{p(\mathrm{h}|\mathbf{X},\theta^t)} = \log p(\mathbf{X},\theta^t) \tag{A.13}$$

applying $p(\mathbf{X},\mathrm{h},\theta^t) = p(\mathbf{X},\theta^t)p(\mathrm{h}|\mathbf{X},\theta^t)$

A.3.3 Maximizing the bound

In order to maximize $B(\theta;\theta^t)$ with respect to θ, we can write

$$\begin{aligned} B(\theta;\theta^t) &\triangleq \mathcal{E}_{f^t}[\log p(\mathbf{X},\mathrm{h},\theta)] + \mathcal{H} \\ &= \mathcal{E}_{f^t}[\log p(\mathbf{X},\mathrm{h}|\theta)] + \log p(\theta) + \mathcal{H} \\ &= Q^t(\theta) + \log p(\theta) + \mathcal{H} \end{aligned} \tag{A.14}$$

where $\mathcal{E}_{f^t}[\cdot]$ denotes the expectation with respect to $f^t(\mathrm{h}) \triangleq p(\mathrm{h}|\mathbf{X},\theta^t)$, $Q^t(\theta)$ is the expected complete log-likelihood defined as $Q^t(\theta) \triangleq \mathcal{E}_{f^t}[\log p(\mathbf{X},\mathrm{h}|\theta)]$, and $p(\theta)$ is the prior on the parameters and $\mathcal{H} \triangleq -\mathcal{E}_{f^t}[\log f^t(\mathrm{h})]$. Since $\mathcal{H}$ does not depend on θ, it is equivalent to maximizing the bound with respect to θ using the first two items only:

$$\theta^{t+1} = \underset{\theta}{\operatorname{argmax}}\ B(\theta;\theta^t) = \underset{\theta}{\operatorname{argmax}}\ [Q^t(\theta) + \log p(\theta)] \tag{A.15}$$

At each iteration, the EM algorithm first finds an optimal lower bound $B(\theta;\theta^t)$ at the current guess θ^t, and then maximizes this bound to obtain an improved estimate θ^{t+1}. Because the bound is expressed as an expectation, the first step is called the "expectation-step" or E-step, whereas in the M-step we are optimizing $Q^t(\theta)$ with respect to the free variable θ to obtain the new estimate θ^{t+1}. It is proven that the EM algorithm converges to a local maximum of $\log p(\mathbf{X},\theta)$, and thus equivalently maximizes the log-posterior $\log p(\theta|\mathbf{X})$. In practice, the E-steps and M-steps alternate repeatly until the difference changes by an arbitrarily small amount in the case of convergence of log-likelihood values.

Appendix B

PreFEst: Predominant-$F0$ estimation

B.1 Weighted-mixture of tone models

Goto formulates the tone model as [5]

$$\begin{aligned} p(x|F,m,\mu^{(t)}(F,m)) &= \sum_{h=1}^{\mathrm{H}_i} p(x,h|F,m,\mu^{(t)}(F,m)) \\ &= \sum_{h=1}^{\mathrm{H}_i} c^{(t)}(h|F,m) G(x; F+1200\log_2 h, \mathrm{W}_i) \end{aligned} \tag{B.1}$$

where x is a conversion from frequency F to *cents*, m ($1 \leq m \leq \mathrm{M}_i$) is the number of tone models, i denotes the melody line(i=m) or bass line(i=b), H_i is the number of harmonics considered, W_i^2 is the variance of the Gaussian distribution $G(\cdot)$ and $\mu^{(t)}(F,m) = \{c^{(t)}(h|F,m)|h=1,\ldots,\mathrm{H}_i\}$ with $c^{(t)}(h|F,m)$ determines the relative amplitude of the h-th harmonic component satisfying $\sum_{h=1}^{\mathrm{H}_i} c^{(t)}(h|F,m) = 1$.

Assume that the observed probability density function from the spectrum was generated from a weighted mixture of all possible tone models:

$$p(x|\theta) = \int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{m=1}^{\mathrm{M}_i} w^{(t)}(F,m) p(x|F,m,\mu^{(t)}(F,m)) dF \tag{B.2}$$

Goto parameterizes the model with the set of parameters $\theta^{(t)} = \{w^{(t)}, \mu^{(t)}\}$ with

$$w^{(t)} = \{w^{(t)}(F,m)|\mathrm{Fl}_i \leq F \leq \mathrm{Fh}_i,\ m=1,\ldots,\mathrm{M}_i\} \tag{B.3}$$

$$\mu^{(t)} = \{\mu^{(t)}(F,m)|\mathrm{Fl}_i \leq F \leq \mathrm{Fh}_i,\ m=1,\ldots,\mathrm{M}_i\} \tag{B.4}$$

where Fl_i and Fh_i denotes the lower and upper limits of the possible $F0$ range and $w^{(t)}(F,m)$ is the weight of a tone model and satisfies

$$\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{m=1}^{\mathrm{M}_i} w^{(t)}(F,m) dF = 1 \tag{B.5}$$

Then the weight $w^{(t)}(F,m)$ can be interpreted as the probability density function of $F0$ as

$$p_{F0}^{(t)}(F) = \sum_{m=1}^{\mathrm{M}_i} w^{(t)}(F,m) \tag{B.6}$$

B.2 Introducing a prior distribution

A prior distribution $p_{0i}(\theta^{(t)})$ of the parameter $\theta^{(t)}$ is chosen by Goto as follows:

$$p_{0i}(\theta^{(t)}) = p_{0i}(w^{(t)})p_{0i}(\mu^{(t)}) \tag{B.7}$$

where

$$\begin{aligned} p_{0i}(w^{(t)}) &= \frac{1}{Z_w} e^{-\beta_{wi}^{(t)} D_w(w_{0i}^{(t)}; w^{(t)})} \\ p_{0i}(\mu^{(t)}) &= \frac{1}{Z_\mu} e^{-\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{m=1}^{\mathrm{M}_i} \beta_{\mu i}^{(t)} D_\mu(\mu_{0i}^{(t)}(F,m); \mu^{(t)}(F,m)) dF} \end{aligned} \tag{B.8}$$

with

$$\begin{aligned} D_w(w_{0i}^{(t)}; w^{(t)}) &= \int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{m=1}^{\mathrm{M}_i} w_{0i}^{(t)}(F,m) \log \frac{w_{0i}^{(t)}(F,m)}{w^{(t)}(F,m)} dF \\ D_\mu(\mu_{0i}^{(t)}(F,m); \mu^{(t)}(F,m)) &= \sum_{h=1}^{\mathrm{H}_i} c_{0i}^{(t)}(h|F,m) \log \frac{c_{0i}^{(t)}(h|F,m)}{c^{(t)}(h|F,m)} \end{aligned} \tag{B.9}$$

defined as Kullback-Leibler's information and Z_w, Z_μ are the normalization factors. We could combine the above two sets of equations to get

$$\begin{aligned} p_{0i}(w^{(t)}) &= \frac{1}{Z_w} e^{-\beta_{wi}^{(t)} \int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{m=1}^{\mathrm{M}_i} w_{0i}^{(t)}(F,m) \log \frac{w_{0i}^{(t)}(F,m)}{w^{(t)}(F,m)} dF} \\ p_{0i}(\mu^{(t)}) &= \frac{1}{Z_\mu} e^{-\beta_{\mu i}^{(t)} \int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{m=1}^{\mathrm{M}_i} \sum_{h=1}^{\mathrm{H}_i} c_{0i}^{(t)}(h|F,m) \log \frac{c_{0i}^{(t)}(h|F,m)}{c^{(t)}(h|F,m)} dF} \end{aligned} \tag{B.10}$$

For the prior distribution of the shape of tone models, Goto uses

$$c_{0i}^{(t)}(h|F,m) = \alpha_{i,m} g_{m,h} G(h; 1, \mathrm{U}_i) \tag{B.11}$$

where $\alpha_{i,m}$ is a normalization factor, U_i is the variance in *cents* and

$$g_{m,h} = \begin{cases} 2/3, & m{=}2 \text{ and } h \text{ is even} \\ 1, & \text{otherwise} \end{cases} \tag{B.12}$$

B.3 MAP estimation using EM algorithm

The problem to be solved is to estimate the model parameter $\theta^{(t)}$ on the basis of the prior distribution $p_{oi}(\theta^{(t)})$ when we observe $p_\Psi^{(t)}(x)$, the probability density function of the band-pass filtered frequency components $BPF_i(x)$, and it is defined as [6]

$$p_\Psi^{(t)}(x) = \frac{BPF_i(x)\Psi_p^{(t)}(x)}{\int_{-\infty}^{\infty} BPF_i(x)\Psi_p^{(t)}(x)dx} \tag{B.13}$$

where $\Psi_p^{(t)}$ is the power distribution function calculated with unit *cents*.

In E-step, the bound is formulated as

$$Q_{\mathrm{MAP}}(\theta^{(t)}|\theta_{old}^{(t)}) = Q(\theta^{(t)}|\theta_{old}^{(t)}) + \log p_{0i}(\theta^{(t)}) \tag{B.14}$$

and

$$\begin{aligned}
&Q(\theta^{(t)}|\theta_{old}^{(t)}) \\
&= \int_{-\infty}^{\infty} p_{\Psi}^{(t)}(x)\mathcal{E}_{F,m,h}[\log p(x,F,m,h|\theta^{(t)})]dx \\
&= \int_{-\infty}^{\infty}\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i}\sum_{m=1}^{\mathrm{M}_i}\sum_{h=1}^{\mathrm{H}_i} p_{\Psi}^{(t)}(x)p(F,m,h|x,\theta_{old}^{(t)})\log p(x,F,m,h|\theta^{(t)})dFdx \\
&= \int_{-\infty}^{\infty}\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i}\sum_{m=1}^{\mathrm{M}_i}\sum_{h=1}^{\mathrm{H}_i} p_{\Psi}^{(t)}(x)p(F,m,h|x,\theta_{old}^{(t)}) \\
&\qquad\qquad \cdot \log\left\{w^{(t)}(F,m)p(x,h|F,m,\mu^{(t)}(F,m))\right\}dFdx \\
&= \int_{-\infty}^{\infty}\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i}\sum_{m=1}^{\mathrm{M}_i}\sum_{h=1}^{\mathrm{H}_i} p_{\Psi}^{(t)}(x)p(F,m,h|x,\theta_{old}^{(t)}) \\
&\qquad\qquad \cdot \log\left\{w^{(t)}(F,m)c^{(t)}(h|F,m)G(x;F+1200\log_2 h,\mathrm{W}_i)\right\}dFdx
\end{aligned} \tag{B.15}$$

where

$$p(F,m,h|x,\theta_{old}^{(t)}) = \frac{w_{old}^{(t)}(F,m)p(x,h|F,m,\mu_{old}^{(t)})(F,m)}{p(x|\theta_{old}^{(t)})} \tag{B.16}$$

Then, the iterative calculation is performed to maximize $Q_{\mathrm{MAP}}(\theta^{(t)}|\theta_{old}^{(t)})$ with respect to $\theta^{(t)} = \{w^{(t)}(F,m), \mu^{(t)}(F,m)\}$.

Appendix C

Spectrum characteristics of musical instruments

Here are some notes, taken from [19] and [20], about the spectral behaviors of some musical instruments. The spectral characteristics of musical instruments should be taken into consideration to improve the performance of a multiple $F0$s estimation system.

C.1 Piano

It is a well-known fact that a piano sounds better with stretched octaves and there are physical reasons and psychological reasons for preferring stretched octaves.

The physical reasons are related to the inharmonicity of the string partials:

$$f_h \approx h \cdot F0 \cdot \sqrt{\frac{1 + h^2 B}{1 + B}} \tag{C.1}$$

where B is the inharmonicity coefficient. In order to minimize beats between each pair of octaves, the octaves should be stretched.

The psychoacoustical reasons are found through the results of several experiments that listeners judge either sequential or simultaneous octaves as true octaves when the interval is about 0.6% greater than a 2:1 frequency ratio.

C.2 Bowed string instruments

Real strings are stiff and the Helmhotz corner is not perfectly sharp. Corner rounding leads to a slight flattening of pitch as the bow force increases.

A broad resonance region that enhances certain harmonics lying in a fixed frequency range is called a formant. The resonance curve shapes the spectrum of vibrating strings and thus provides cues to identify complex tones of musical instrument. The three resonance modes are: air modes, top modes and body modes. The air resonance of a double bass, for example, is around 60Hz and its body resonance is around 100Hz.

C.3 Woodwind reed instruments

An idealized cylindrical pipe with close-open ends sounds an octave below the pitch of a similar pipe open at both ends. In other words, the overtones of a close-open-ends pipe are odd harmonics of its fundamental frequency. The clarinet of which the mouthpiece with the reed behaves as a close end has a resonance curve boosting at odd harmonics, while dipping at even harmonics. However, an oboe has a full harmonic spectrum with a relative weak $F0$ and a slope of -12dB/Oct. above the cutoff frequency.

C.4 Brass instruments

The steady-state spectral envelope of a brass instrument is characterized by a cuttoff frequency below which spectral magnitudes are approximately equal or they increase gradually with frequency, and above which the amplitude decreases sharply. The rate of rise below cutoff is typically 2 to 4dB/Oct., and the rate of fall above cutoff is typically -15 to -25dB/Oct..

As the instrument is played more loudly, a greater fraction of the radiated power is contained in the partials near and above cutoff. The slope below cutoff increases and the slope above cutoff decreases as the intensity level is increased.

Ancell has measured the silimlar acoustic behavior of several mutes on brass instruments. Most of the mutes show a Helmholtz resonance associated with their internal cavity at a frequency in the range 200-300Hz. This resonance causes a broad dip in the radiated spectrum of the instrument in the region of the first and second harmonics over most of its compass and makes the sound thin and reedy. The resonances and antiresonances of all kinds of mutes lie above about 1000Hz and impart particular tonal qualities to the sound.

C.5 Mallet percussion instruments

A bar of uniform thickness with free ends vibrates in a series of normal modes whose frequncies are approximately in the ratios: $(3.011)^2 : 5^2 : 7^2 : 9^2 : 11^2 : \cdots$. Removing materials from a bar affects certain modal frequencies and thus could be applied to tuning the individual partials. It is customary to tune one marimba, xylophone or vibrphone, such that the first two partials are 4 times and about 10 times, respectively, the magnitude of the fundamental frequency. However, adding resonating tubes with open-close ends, for example, might boost the odd-numbered harmonics and regularize the spectrum.

Appendix D

Important techniques used in f0

D.1 Optimal phase differences introduced in the superposition of model harmonics

In order to maintain a clear spectral representation of model peaks, we minimize the sum of the spectral magnitudes by introducing an optimal phase shift to each harmonic partial. Thus, for the hth harmonic partial, it is constructed by:

$$Partial_{F0,h}(t) = a \cdot cos(2\pi \frac{h \cdot F0}{Fe} n + (h-1) \cdot \Phi_{F0}) \tag{D.1}$$

where a represents a constant amplitude and Φ_{F0} represents the optimal phase shift and is calculated as

$$\Phi_{F0} = angle\{FT[w(n) \cdot e^{j2\pi F0/Fe \cdot (n-1)}] \cdot FT[w(n)]\} \tag{D.2}$$

where $w(n)$ is the window function, $FT\{\cdot\}$ denotes the Fourier transform. This result is obtained by minimizing the sum of the absolute squared spectrum for two partials with offset $F0$.

D.2 Eliminating phase slopes in model peaks

Considering a harmonic signal $s(n) = e^{j\omega n + \rho}$ truncated by a rectangluar window $rect(n)$ with length N, the discrete Fourier transform of $s(n) \cdot rect(n)$ is

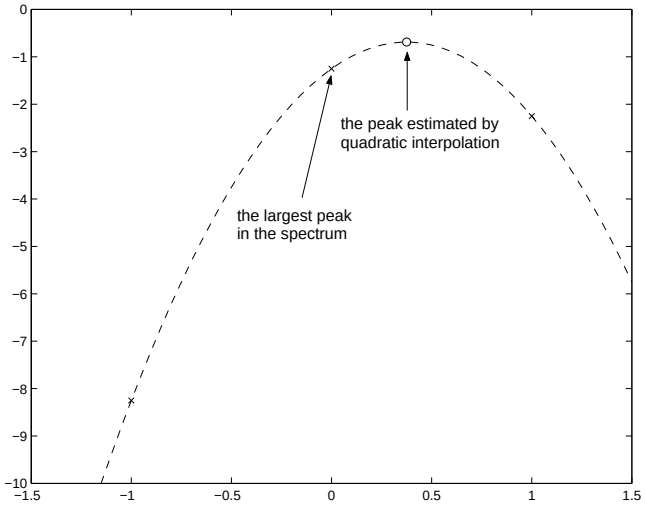

Figure D.1: Quadratic interpolation for the peak position

calculated as

$$\begin{aligned}
&\sum_{n=0}^{N-1} a(n)e^{j\omega n+\rho} \cdot e^{2\pi kn/N} \\
&= e^{j\rho} \cdot \sum_{n=0}^{N-1} e^{j(\omega-2\pi k/N)n} = \sum_{n=0}^{N-1} \frac{1-e^{j(\omega-2\pi k/N)N}}{1-e^{j(\omega-2\pi k/N)}} \\
&= e^{j\rho} \cdot \frac{e^{j(\omega-2\pi k/N)N/2}}{e^{j(\omega-2\pi k/N)/2}} \cdot \frac{sin[(\omega-2\pi k/N)N/2]}{sin[(\omega-2\pi k/N)/2]} \\
&= e^{j\rho} \cdot e^{(\omega-2\pi k/N)(N-1)/2} \cdot \frac{sin[(\omega-2\pi k/N)N/2]}{sin[(\omega-2\pi k/N)/2]}
\end{aligned} \tag{D.3}$$

Thus, we could eliminate the phase variation in model peaks by multiplying the truncated signal with $e^{j(2\pi k/N)(N-1)/2}$.

D.3 Local quadratic approximation for correcting the peak frequency

A real maximum peak of one formant might present between two frequency bins. Thus, we apply a second order polynomial approximation function to obtain a better estimate of the peak frequency(Fig. 4).

D.4 Estimating peak frequencies and frequency slopes using reassignment operators

In [15], F. Auger and P. Flandrin proposed a reassignment method for the time-frequency representation based on finding the center of gravity of the energy distribution around the time-frequency position (t, ω). Given a STFT using the window $h(t)$, the reassignment operators could be expressed as:

$$\hat{t}(t,\omega) = t - \mathcal{R}e\{\frac{STFT_{h_T}(t,\omega) \cdot STFT_h^*(t,\omega)}{|STFT_h(t,\omega)|^2}\} \tag{D.4}$$

$$\hat{\omega}(t,\omega) = \omega - \mathcal{I}m\{\frac{STFT_{h_D}(t,\omega) \cdot STFT_h^*(t,\omega)}{|STFT_h(t,\omega)|^2}\} \tag{D.5}$$

where $h_T(t) = t \cdot h(t)$ and $h_D = \frac{\partial h(t)}{\partial t}$. The frequency slope is thus obtained by

$$\omega'(t,\omega) = \frac{\frac{\partial \hat{\omega}(t,\omega)}{\partial t}}{\frac{\partial \hat{t}(t,\omega)}{\partial t}} \tag{D.6}$$

where

$$\begin{aligned}
\frac{\partial \hat{t}(t,\omega)}{\partial t} &= -\mathcal{R}e\{\frac{STFT_{h_{DT}}(t,\omega) \cdot STFT_h^*(t,\omega)}{|STFT_h(t,\omega)|^2}\} \\
&\quad - \mathcal{R}e\{\frac{STFT_{h_D}(t,\omega) \cdot STFT_h^*(t,\omega)}{|STFT_h(t,\omega)|^2} \cdot \frac{STFT_{h_T}(t,\omega) \cdot STFT_h^*(t,\omega)}{|STFT_h(t,\omega)|^2}\} \\
\frac{\partial \hat{\omega}(t,\omega)}{\partial t} &= \mathcal{I}m\{\frac{STFT_{h_{DD}}(t,\omega) \cdot STFT_h^*(t,\omega)}{|STFT_h(t,\omega)|^2}\} \\
&\quad - \mathcal{I}m\{\frac{STFT_{h_D}(t,\omega) \cdot STFT_h^*(t,\omega)}{|STFT_h(t,\omega)|^4}\}
\end{aligned} \tag{D.7}$$

with $h_{DT}(t) = t \cdot \frac{\partial h(t)}{\partial t}$ and $h_{DD}(t) = \frac{\partial^2 h(t)}{\partial t^2}$. Furthermore, we use this frequency slope to update the peak frequency estimation[21]:

$$\hat{\omega}(t,\omega) = \omega - [t - \hat{t}(t,\omega)] \cdot \omega'(t,\omega) \tag{D.8}$$

Appendix E

Evolutionary algorithms

Evolutionary algorithms are stochastic search methods that mimic the metaphor of natural biological evolution. Evolutionary algorithms operate on a population of potential solutions applying the principle of survival of the fittest to produce better and better approximations to a solution. In each generation, a new set of approximations is created by the process of selecting individuals according to their level of fitness in the problem domain and breeding them together using operators borrowed from natural genetics. This process leads to the evolution of populations of individuals that are better suited to their environment than the individuals that they were created from, just as in natural adaptation. Evolutionary algorithms work on populations of individuals instead of single solutions. In this way the search is performed in a parallel manner.

At the beginning of the computation a number of individuals (the population) are randomly initialized. The objective function is then evaluated for these individuals and the first generation is produced. If the optimization criteria are not met the creation of a new generation starts. Individuals are selected according to their fitness for the production of offspring. Parents are recombined to produce offspring. All offspring will be mutated with a certain probability and then the fitness of the offspring is computed. The offspring are inserted into the population replacing the parents, producing a new generation. This cycle is performed until the optimal criteria are reached.

Such a single population evolutionary algorithm is powerful and performs well on a broad class of problems. However, better results can be obtained by introducing many populations, called subpopulations. Every subpopulation evolves over a few generations isolated (like the single population evolutionary algorithm) before one or more individuals are exchanged between the subpopulations. The multipopulation evolutionary algorithm models the evolution of a species in a way more similar to nature than the single population evolutionary algorithm.

Selection

Selection determines which individuals are chosen for recombination and how many offspring each selected individual produces. The actual selection operates after the first generation. Parents are selected to fit the best of the criteria predefined.

Recombination

Recombination produces new individuals in combining the information contained in the parents. Discrete recombination performs an exchange of variable values between the individuals and is applied in our test.

Mutation

After recombination every offspring undergoes mutation. Offspring variables are mutated by the addition of small perturbations(size of the mutation step), with low probability. In our test, the perturbations based on Gaussian distribution and uniform distribution are assigned with different probabilities.

Reinsertion

If fewer offspring are produced than the size of the original population the offspring have to be reinserted into the old population. Similarly, if not all offspring are to be used in each generation or if more offspring are generated than needed, then a reinsertion scheme must be used to determine which individuals should be inserted into the new population. Usually, the used selection method determines what is applied.

Bibliography

[1] Bregman, A. S.(1990). *Auditory scene analysis*, The MIT Press, 1990.

[2] Plomp, Reinier(1976). *Aspects of tone sensation*, Academic Press, 1976.

[3] Hess, Wolfgang(1983). *Pitch determination of speech signals*, Springer Verlag, 1983.

[4] Martin, Ph.(1982). "Comparison of pitch detection by cepstrum and spectral comb analysis", *ICASSP-82*, pp.180-183, 1982.

[5] Goto, Masataka(2001). "A predominant-F0 estimation method for CD recordings: MAP estimation using EM algorithm for adaptive tone models", *ICASSP 2001 Proseedings*, pp. V-3365-3368, May 2001.

[6] Goto, Masataka(2000). "A robust predominant-F0 estimation method for real-time detection of melody and bass line in CD recordings", *ICASSP 2000 Proseedings*, pp. II-757-760, June 2000.

[7] Walmsley, Paul J., Godsill, Simon J. and Payner, Peter J. W.(1999). "Bayesian graphical models for polyphonic pitch tracking", *Diderot Forum, Vienna*, Dec. 1999.

[8] Davy, Manuel and Godsill, Simon J.(2003). "Bayesian harmonic models for musical signal analysis", *Bayesian Statistics 7*, Oxford University Press, 2003.

[9] Klapuri, Anssi(1999). "Sound onset detection by applying psychoacoustic knowledge", *Proc. ICASSP*, 1999.

[10] Klapuri, Anssi(1998). "Number theoretical means of resolving a mixture of several harmonic sounds", *Proc. of the European Signal Processing Conference EUSIPCO*, 1998.

[11] Martin, Keith D.(1996). "Automatic transcription of simple polyphonic music: robust frond end processing", *The third joint meeting of the Acoustical Societies of America and Japan*, Dec. 1996.

[12] Ellis, Daniel P. W.(1996). "Prediction-driven computational auditory scene analysis", Ph.D. thesis, M.I.T., Cambridge, MA, June 1996.

[13] Krauledat, Matthias(2003). "Fundamental frequency estimation", *Rapport interne de l'equipe Analyse/Synthèse*, 2003.

[14] Cohen, Leon(1995). *Time-frequency analysis*, Prentice Hall, Inc., 1995.

[15] 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*, Vol.43, No. 5, May 1995.

[16] Bishop, Christopher M.(1995). *Neural Network for Pattern Recognition*, Oxford University Press Inc., New York, 1995.

[17] Minka, Thomas P.(1998). "Expectation–Maximization as lower bound maximization", `http://web.media.mit.edu/~ tpminka/papers /learning.html`

[18] Dellaert, Frank(2002). "The expectation maximization algorithm", `http://www.cc.gatech.edu/ dellaert/html/publications.html`

[19] Fletcher, Neville H. and Rossing, Thomas D.(1998). *The physics of musical instruments*, Springer, 1998.

[20] Juan G, Roederer(1994). *The physics and psychophysics of music*, third edition, Springer-Verlag New York Inc., 1995.

[21] Axel Röbel(2002). "Estimating partial frequency and frequency slope using reassignment operator", *Proc. of the International Computer Music Conference (ICMC'02)*, pp. 122-125, Göteborg, 2002.