Statistical Approach to Multipitch Analysis
(統計的手法による多重音解析に関する研究)

Hirokazu Kameoka
亀 岡　弘 和

Contents

Abstract

We deal through this paper with the problem of estimating "information" of each sound source separately from an acoustic signal of compound sound. Here "information" is used in a wide sense to include not only the waveform itself of the separate source signal but also the power spectrum, fundamental frequency (F_0), spectral envelope and other features. Such a technique could be potentially useful for a wide range of applications such as robot auditory sensor, robust speech recognition, automatic transcription of music, waveform encoding for the audio CODEC (compression-decompression) system, a new equalizer system enabling bass and treble controls for separate source, and indexing of music for music retrieval system.

Generally speaking, if the compound signal were separated, then it would be a simple matter to obtain an F_0 estimate from each stream using a single voice F_0 estimation method and, on the other hand, if the F_0s were known in advance, could be very useful information available for separation algorithms. Therefore, source separation and F_0 estimation are essentially a "chicken-and-egg problem", and it is thus perhaps better if one could formulate these two tasks as a joint optimization problem. In Chapter 2, we introduce a method called "Harmonic Clustering", which searches for the optimal spectral masking function and the optimal F_0 estimate for each source by performing the *source separation* step and the F_0 *estimation* step iteratively.

In Chapter 3, we establish a generalized principle of Harmonic Clustering by showing that Harmonic Clustering can be understood as the minimization of the distortion between the power spectrum of the mixed sound and a mixture of spectral cluster models. Based on this fact, it becomes clear that this problem amounts to a maximum likelihood problem with the continuous Poisson distribution as the likelihood function. This Bayesian reformulation enables us not only to impose empirical constraints, which are usually necessary for any underdetermined problems, to the parameters by introducing prior probabilities but also to derive a model selection criterion, that leads to estimating the number of sources. We confirmed through the experiments the effectiveness of the two techniques introduced in this chapter: multiple F_0 estimation and source number estimation.

Human listeners are able to concentrate on listening to a target sound without difficulty even in the situation where many speakers are talking at the same time or many instruments are played together. Recent efforts are being directed toward the attempt to implement this ability by human called the "auditory stream segregation". Such an approach is referred to as the "Computational Auditory Scene Analysis (CASA)". In Chapter 4, we aim at developing a computational algorithm enabling the decomposition of the time-frequency components of the signal of interest into distinct clusters such that each of them is associated with a single auditory stream. To do so, we directly model a spectro-temporal model whose shape can be taken freely within the constraint called "Bregman's grouping cues", and then try to fit the mixture of this model to the observed spectrogram as well as possible. We call this approach "Harmonic-Temporal Clustering". While most of the conventional methods usually perform separately the extraction of the instantaneous features at each discrete time point and the estimation of the whole tracks of these features, the method described in this chapter performs these procedures simultaneously. We confirmed the advantage of the proposed method over conventional methods through experimental evaluations.

Although many efforts have been devoted to both F_0 estimation and spectral envelope estimation intensively in the speech processing area, the problem of determining F_0 and spectral envelope seems to have been tackled independently. If the F_0 were known in advance, then the spectral envelope could be estimated very reliably. On the other hand, if the spectral envelope were known in advance, then we could easily correct subharmonic errors. F_0 estimation and spectral envelope estimation, having such a chicken and egg relationship, should thus be done jointly rather than independently with successive procedures. From this standpoint, we will propose a new speech analyzer that jointly estimates pitch and spectral envelope using a parametric speech source-filter model. We found through the experiments a significant advantage of jointly estimating F_0 and spectral envelope in both F_0 estimation and spectral envelope estimation.

The approaches of the preceding chapters are based on the approximate assumption of additivity of the power spectra (neglecting the terms corresponding to interferences between frequency components), but it becomes usually difficult to infer F_0s when two voices are mixed with close F_0s as far as we are only looking at the power spectrum. In this case not only the harmonic structure but also the phase difference of each signal becomes an important cue for separation. Moreover, having in mind future source separation methods designed for multi-channel signals of multiple sensory input, analysis methods in the complex spectrum

domain including the phase estimation are indispensable. Taking into account the significant effectiveness and the advantage of the approach described in the preceding chapters, we have been motivated to extend it to a complex-spectrum-domain approach without losing its essential characteristics. The main topic of Chapter 6 is the development of a nonlinear optimization algorithm to obtain the maximum likelihood parameter of the superimposed periodic signal model: focusing on the fact that the difficulty of the single tone frequency estimation or the fundamental frequency estimation, which are at the core of the parameter estimation problem for the sinusoidal signal model, comes essentially from the nonlinearity of the model in the frequency parameter, we introduce a new iterative estimation algorithm using a principle called the "auxiliary function method". This idea was inspired by the principle of the EM algorithm. Through simulations, we confirmed that the advantage of the proposed method over the existing gradient descent-based method in the ability to avoid local solutions and the convergence speed. We also confirmed the basic performance of our method through 1ch speech separation experiments on real speech signal.

Abstract in Japanese

本研究は、様々な音が混在する中で目的音の情報（基本周波数やスペクトル包絡など）を分離推定する多重音解析技術を提案するものである。多重音解析技術は実用性が高く、ロボット聴覚、知能的音響センサ、音声認識、音源分離、自動採譜、オーディオコーデックの効率的な符号化機能、楽音ごとにイコライズできる高自由度イコライザ、音楽コンテンツの自動メタデータ化とそれによる多種機能つき音楽検索システムなど、実に広範囲にわたるアプリケーションへの応用が期待される。

多重音中の各音源の基本周波数は、各音源のスペクトルが既知であれば、高い精度で推定できる。一方で、多重音スペクトルは、各音源の基本周波数が既知であれば高い精度で分離できうる。このことから分かるように、多重音スペクトル分離と基本周波数推定の問題はいわゆる「鶏と卵の関係」にある。従って、多重音スペクトルを音源ごとに分解することと各音源の基本周波数は同時最適化問題として解かれるべきであると我々は考えた。そこで、第2章では、音源分離と基本周波数推定を同時最適化問題として定式化し、対象とする混合音のパワースペクトルを音源ごとに対応するようにクラスタ化する分配関数 (バイナリマスク) と、各音源の基本周波数を最適推定する原理を提案する。この最適解探索は音源分離ステップと基本周波数ステップの反復計算により行うことができ、この方法を調波構造化クラスタリングと呼ぶ。

第3章では、調波構造化クラスタリングの原理を一般化したのちに、これをベイズ的枠組で別解釈および再定式化を行う。その結果から、調波構造化クラスタリングはパラメトリックな調波構造モデルの重ね合わせによる観測スペクトルの最適フィッティングであるという解釈ができること、この解釈に基づきさらにはモデル構造選択規準により音源数を推定するための規準が作れることが示される。評価実験により、本章で提案した2つの要素技術である、調波構造化クラスタリングによる多重ピッチ推定法と音源数自動決定法の有効性がいずれも示された。

人間の聴覚機能を計算機で実現しようという試みが活発に進められており、その枠組を総称して計算論的聴覚情景分析 (Computational Auditory Scene Analysis: CASA) と呼ぶ。近年この研究分野における興味の対象は、Bregman が指摘した音脈 (人間がひとまとまりと知覚する音の流れ) が形成されるための条件 (分凝要件) に基づく混合音分離法の実現にある。

第 4 章では、入力された音響信号の時間周波数成分を音脈に対応する時間周波数成分に分解する問題を定式化する。第 4 章で提案する調波時間構造化クラスタリングのアイディアの要点は、Bregman の分凝要件から逸脱しない範囲の自由度をもった時変スペクトルを直接的にモデル化し、これを混合したもので対象の時変スペクトルを説明しようとすることである。各時刻で独立に調波成分を見つけ出すための処理 (周波数方向の群化) と、抽出された調波成分特徴量の時系列を時間方向にスムージングする処理 (時間方向の群化) を多段処理的に行う多くの従来法に対し、調波時間構造化クラスタリングは、これらを協調し合う同時最適化問題として定式化したものに相当し、個々の音源の時間周波数全域に渡ったパワースペクトル構造を一挙に推定できる新しい方法論である。評価実験により、混合音声信号および音楽音響信号の基本周波数推定精度が、それぞれの分野における最先端の従来法を上回ることを確認した。

第 5 章では、従来まで独立な問題として扱われがちであったスペクトル包絡推定と基本周波数推定は本来相補関係にあるべきとの問題意識のもと、これらを同時最適化問題として定式化し、個々の音源のスペクトル包絡推定も同時に行える多重音解析法への応用可能性を示した。また、単一話者音声を対象としたピッチ推定、合成分析、分析合成に関する各評価実験を通して、提案法のようにピッチ周波数とスペクトル包絡を同時推定することがいずれの推定の精度に対しても良い効果をもたらしたことを確認した。

第 5 章までの手法は、パワースペクトルの加法性が近似的に成り立つ (周波数成分間の干渉項を無視できる) という仮定のもとで、観測パワースペクトルから所望の情報を得るためのアプローチであった。しかし、2 音以上の近接する基本周波数の分離推定や、近接周波数成分の分離を高精度に行うためには、調波構造だけでなく各信号の位相差が分離の手がかりになる。さらに、将来的に複数センサ入力の多チャンネル信号を対象とした音源分離手法を視野に入れるのであれば、位相推定を含んだ複素スペクトル領域での解析が不可欠である。第 5 章までの方法論の有効性と上記のような展望を踏まえ、第 6 章では、第 3 章のアプローチをその本質を損なうことなく複素スペクトル領域に拡張できないかということがテーマであり、中心的な議論の対象は、周期信号重畳モデルの最尤パラメータを求めるための非線形最適化アルゴリズムを開発することである。そこで、正弦波重畳モデルのパラメータ推定問題の核である周波数推定ないし基本周波数推定の難しさの本質が、正弦波重畳モデルが周波数パラメータに関して非線形である点にあることに着目し、補助関数を用いた新しい反復推定アルゴリズムを導く。この考え方は、EM アルゴリズムのヒントにして着想したものである。シミュレーション実験により、第 6 章で提案するパラメータ最適化アルゴリズムは勾配法を用いる多数の従来法よりも局所解回避能力と収束速度の面で優れていることを示した。また、実音声を用いた 1 チャネル混合音声分離実験を行い、提案法の基本性能を確認した。

Chapter 1

Introduction

1.1 Background

We deal through this paper with the problem of estimating "information" of each sound source separately from an acoustic signal of concurrent sound sources. The acoustic signal can be the compound sound of several people speaking at the same time, or several musical instruments playing together. Here "information" is used in a wide sense to include not only the waveform itself of the separate source signal but also the power spectrum, fundamental frequency (F_0), spectral envelope and other features. Such a technique could be potentially useful for a wide range of applications such as robot auditory sensor, robust speech recognition, automatic transcription of music, waveform encoding for the audio CODEC (compression-decompression) system, a new equalizer system enabling bass and treble controls for separate source, and indexing of music for music retrieval system. The problem, however, is not so simple to solve. We will henceforth call this kind of problem "multisource analysis". Multisource analysis can be categorized in several types of problem setting depending on the situation one assumes. A situation where there are more sensors than source signals, for instance, is referred to as the overdetermined case, in which the source separation can be performed satisfactorily especially in clean environment by using the well-known "Independent Component Analysis" (see, for example, [46, 47]). A situation where there are less sensors than source signals, on the other hand, is referred to as the underdetermined case. In such a situation, one requires some empirical assumption in addition to the statistical independence of sources. One of the most well-known such assumptions is called the "time-frequency sparseness of speech", which assumes that time-frequency components of sources rarely overlap with each other. This assumption is experimentally supported by the

methods that use a binary mask to extract only the time-frequency components which seems to have propagated from the same spatial direction (or position) [117, 91, 92, 8].

On the contrary, the particular problem of interest in this paper is a multisource analysis where one only has a single sensory input and does not know how many sources in the compound acoustic signal. We will be confronted with such a situation, for example, when we need to detect musical note from monaural CD recordings, or when several different source signals originate from very close position even if we had multiple sensors. The greatest difference from the multisensor case is thus that it is impossible to obtain spatial data of sources.

1.2 Source Separation and F_0 Estimation

The auditory system is able to extract the period despite very different waveforms or spectra of sounds at the ears. Explanations of how this is done have been elaborated since antiquity [27]. Even with a monaural recording, a musically inclined listener can often follow and concentrate on the particular melodic line of each instrument in a polyphonic ensemble. This implies that human can hear several pitches from a single compound waveform. As psychophysical data on this capability are said to be fragmentary (see, for example, [12, 13, 51]), the limits of this capability, and the parameters that determine them, are not well known. This "proof of feasibility" has nevertheless encouraged the search for algorithms for multisource analysis. The task of multisource analysis in essence involves two tasks: source separation and F_0 estimation. If the compound signal representing the mixture were separated into single source signals, then it would be a simple matter to derive an F_0 estimate from each stream using a single voice F_0 estimation algorithm (comprehensive reviews for single voice F_0 estimation methods can be seen in [49, 50]). On the other hand the F_0s, if known in advance, could be very useful information available for separation algorithms. This leads to a "chicken and egg" situation: estimation and segregation are each a prerequisite of the other, the difficulty being to "bootstrap" this process.

Conventional techniques for multisource analysis are usually designed to cope with either of the two tasks, F_0 (multipitch) estimation and source separation. Many publications on methods corresponding to the former type have been proposed [111, 101, 59, 105, 38, 79, 60, 61, 116, 45, 10, 32, 43, 66, 67, 69, 102, 88, 89, 112, 107, 108, 75, 23, 24, 26, 72, 113, 114, 115, 22, 57], which can be found in de Cheveigné's excellent review paper [28].

A learning-based method such like *sparse coding* [110], *non-negative sparse coding* [109, 2, 97], and *non-negative matrix factorization* [94] models the signal or power spectrum as a weighted sum of basis functions and tries to estimate them such that each of them is a waveform structure or a power spectrum structure that seems to recur many times in the whole acoustic signals or spectrogram. This approach enables source separation without estimating F_0s and thus corresponds to the latter type.

However, since source separation and F_0 estimation, as is mentioned beforehand, are in essence a "chicken and egg" problem, it is perhaps better if one could formulate these two tasks as a joint optimization problem. In Chapter 2, a new principle called "Harmonic Clustering" is introduced, which iteratively performs two steps: source separation and F_0 estimation, in which the common objective function is decreased/increased monotonically at each iteration step. This is reformulated in Chapter 3 in a Bayesian framework, which enables further extensions.

1.3 Estimating the Number of Sources

Up to now, no concern was given to finding the number of sources present within a mixture. This is a difficult aspect of multisource analysis. Many studies ignore it and concentrate on the simpler task of producing some fixed number of estimates, regardless of the number of sources.

Some signals are inherently ambiguous, and may be interpreted either as a single voice with low F_0, or as the sum of several voices with higher, harmonically-related F_0s. Tuned to find as many voices as possible (or to favor the shortest possible periods) an algorithm may "dismember" a voice into subsets of partials, tuned to find as few as possible (or the longest possible periods) it may coalesce multiple voices. The voice count is accordingly over- or underestimated.

Iterative estimation methods typically apply a model at each iteration, and assign as much signal power to a voice as fits this model. Iteration continues on the remainder, and stops when the spectrum (or waveform) has been depleted of power. In the presence of noise, it may be difficult to distinguish between residual noise and yet another source.

In the method of [26], cancellation filters are applied successively to remove each periodic voice. The algorithm stops when application of a new filter reduces power by less than a criterion ratio. Klapuri evaluates the "global weight" of the F_0 candidate derived from the

residual after a voice has been suppressed, and stops the search if that weight falls below threshold [67]. In nonnegative deconvolution [90, 95], the number of sources is given by the number of elements of deconvolved matrix with amplitudes greater than some threshold. Wu and colleagues [113, 114, 115] use an HMM to model transitions between states of 0, 1 or 2 voices.

The Bayesian formulation of the Harmonic Clustering enables us further to derive a theoretical framework for estimating the number of sources.

1.4 Temporal and Spectral Continuity

In general, there are two situations in which the multisource analysis becomes extremely difficult to solve: One is when the F_0s of two or more sources coincide at a particular instant of time, and the other is when the partials of one source overlap with those of other sources. Are there any ways to make a reasonable guess for restoring the partial amplitudes of the underlying sources in such situations? A hint for this apparently unsolvable question is the temporal and spectral continuity of source signals.

A common assumption is that voices should change gradually over time. Continuity of F_0 contours is often exploited in so-called "post-processing" algorithms [49] such as median-smoothing, dynamic programming, Kalman filtering [76, 106, 4], hidden Markov models [113, 114, 115], or multiple agents [75, 45] in order to improve the F_0 estimation results obtained via some frame-by-frame F_0 estimation algorithm. In addition to continuity of F_0 tracks, the assumption that partial amplitudes vary smoothly can be used to track voices over instants when F_0s cross. A challenge to the multisource analysis using the continuity-constraints of F_0 and amplitude tracks is called the "Computational Auditory Scene Analysis (CASA)", which will be mentioned more in details in Chapter 4.

An assumption that has been used recently is spectral smoothness, that is, limited variation of partial amplitudes across frequency axis [67, 116, 108, 10, 21, 69]. Speech and many musical instruments usually have smooth spectral envelopes. Irregularity of the compound spectrum then signals the presence of multiple voices, and smoothness allows the contribution of a voice to shared partials to be discounted. For example if two voices are at an octave from each other, partials of even rank are the superposition of both voices. Based on spectral smoothness, the contribution of the voice of the lower F_0 can be inferred from the amplitude of partials of odd rank, and can then be subtracted to reveal the presence of the voice of the

higher F_0. Spectral smoothness has also been used to reduce subharmonic errors [10, 67].

If the F_0s of all the sources within a mixture were known in advance, then the spectral envelope could be inferred very reliably using the spectral smoothness constraint. On the other hand, if the spectral envelope were known in advance, then we could easily correct subharmonic errors as noted above. Here we find another "chicken and egg" situation, which motivates us to formulate a joint estimation method of F_0 and spectral envelope with the spectral smoothness constraint. This will be discussed in Chapter 5.

1.5 Objective of the Thesis

The objective of this paper is to propose a unified methodological framework, in which one can handle (1) source separation, (2) multipitch estimation, (3) estimation of the number of sources, (4) estimation of the continuous temporal trajectories of F_0s and amplitudes, and (5) spectral envelope estimation, at the same time.

Chapter 2

Harmonic Clustering

2.1 Introduction

The auditory system is able to extract the period despite very different waveforms or spectra of sounds at the ears. Explanations of how this is done have been elaborated since antiquity [27]. Even with a monaural recording, a musically inclined listener can often follow and concentrate on the particular melodic line of each instrument in a polyphonic ensemble. This implies that several pitches may be heard from a single compound waveform. As psychophysical data on this capability are said to be fragmentary (see, for example, [12, 13, 51]), the limits of this capability, and the parameters that determine them, are not well known. This "proof of feasibility" has nevertheless encouraged the search for algorithms for multisource analysis. The task of multisource analysis in essence involves two tasks: source separation and F_0 estimation. If the compound signal representing the mixture were separated into single source signals, then it would be a simple matter to derive an F_0 estimate from each stream using a single voice F_0 estimation algorithm (comprehensive reviews for single voice F_0 estimation methods can be seen in [49, 50]). On the other hand the F_0s, if known in advance, could feed some of the separation algorithms. This leads to a "chicken and egg" situation: estimation and segregation are each a prerequisite of the other, the difficulty being to "bootstrap" this process.

Conventional techniques for multisource analysis are usually designed to cope with either of the two tasks, F_0 (multipitch) estimation and source separation. Many publications on methods corresponding to the former type can be found in de Cheveigné's excellent review paper [28]. A learning-based method such like *sparse coding* [110], *non-negative sparse coding* [109, 2, 97], and *non-negative matrix factorization* [94] models the signal or power spectrum

as a weighted sum of basis functions and tries to estimate them such that each of them is a waveform structure or a power spectrum structure that seems to recur many times in the whole acoustic signals or spectrogram. This approach enables source separation without estimating F_0s and thus corresponds to the latter type.

However, since source separation and F_0 estimation, as mentioned beforehand, are in essence a "chicken and egg" problem, it is perhaps better if one could formulate these two tasks as a joint optimization problem. In this chapter, we propose a new principle called "Harmonic Clustering", which iteratively performs two steps: source separation and F_0 estimation, in which the common objective function is decreased/increased monotonically at each iteration step.

2.2 Principle

2.2.1 Binary Masking of Power Spectrum Based on Sparseness

Let us assume here for simplicity that frequency components of a source signal are sparsely distributed so that components rarely overlap with each other. More specifically, it is assumed here that a frequency component at some frequency-bin originates completely from only a single source (see, for example, [117, 91, 92, 8] for the justification for this assumption). Similarly to the Yilmaz's method, we shall consider to estimate an ideal binary mask that extracts only the components that seem to originate from the same source. What differs from Yilmaz's is that we are dealing with an single sensory input and a guide to estimate the ideal binary mask is the harmonic structure, that depends on F_0 of speech.

First we will consider the single-tone case and introduce a very simple yet intuitive idea to estimate the frequency from power spectrum in the next subsection. This idea is extended to the single-voice case in Subsection 2.2.3 and is generalized to the multisource case in Subsection 2.2.4.

2.2.2 Single-Tone Frequency Estimation

According to the Rife's paper [83], the peak frequency of the power spectrum of a single tone is said to be the unbiased maximum likelihood estimator, when noise is assumed to be a Gaussian white noise. Using this result, we introduce a simple yet intuitive objective

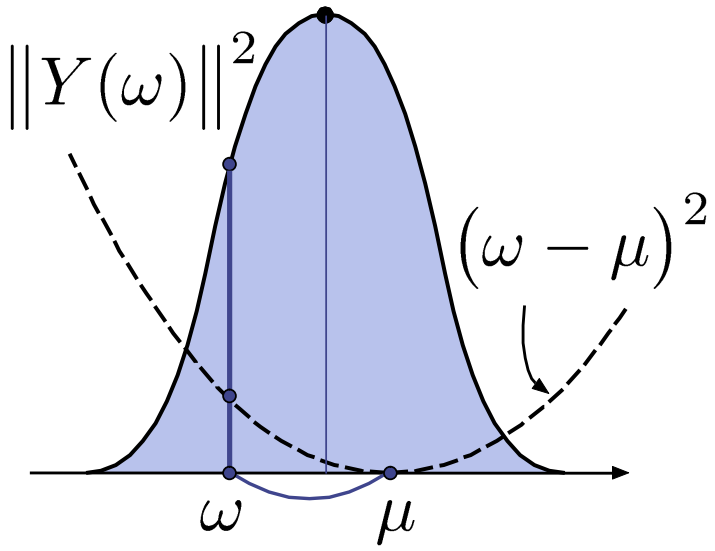


Figure 2.1 The objective function (Eq. (2.1)) is minimized when μ coincides to the peak frequency of $\|Y(\omega)\|^2$

function to obtain the frequency estimate. This concept will be applied also in the following extended versions.

Let $\|Y(\omega)\|^2$ be the observed short-time power spectrum of a single tone signal (complex sinusoid). The shape of this distribution depends on the shape of the window function we choose to use. Assuming the particular case where the peak and the mean frequencies of this distribution coincide (where the distribution is symmetric about the peak), then one can obtain the frequency estimate by finding μ that minimizes

$$\int_{-\infty}^{\infty} (\omega-\mu)^2 \|Y(\omega)\|^2 \mathrm{d}\omega. \tag{2.1}$$

Consequently, the frequency estimate is derived as the mean of the distribution:

$$\mu = \frac{\int_{-\infty}^{\infty} \omega \|Y(\omega)\|^2 \mathrm{d}\omega}{\int_{-\infty}^{\infty} \|Y(\omega)\|^2 \mathrm{d}\omega}. \tag{2.2}$$

2.2.3 Single-Voice F_0 Estimation and Overtone Separation

If one thinks of applying the above method to the single voice case, one may want to separate overtones as if one is dealing with the single tone case problems separately. For this

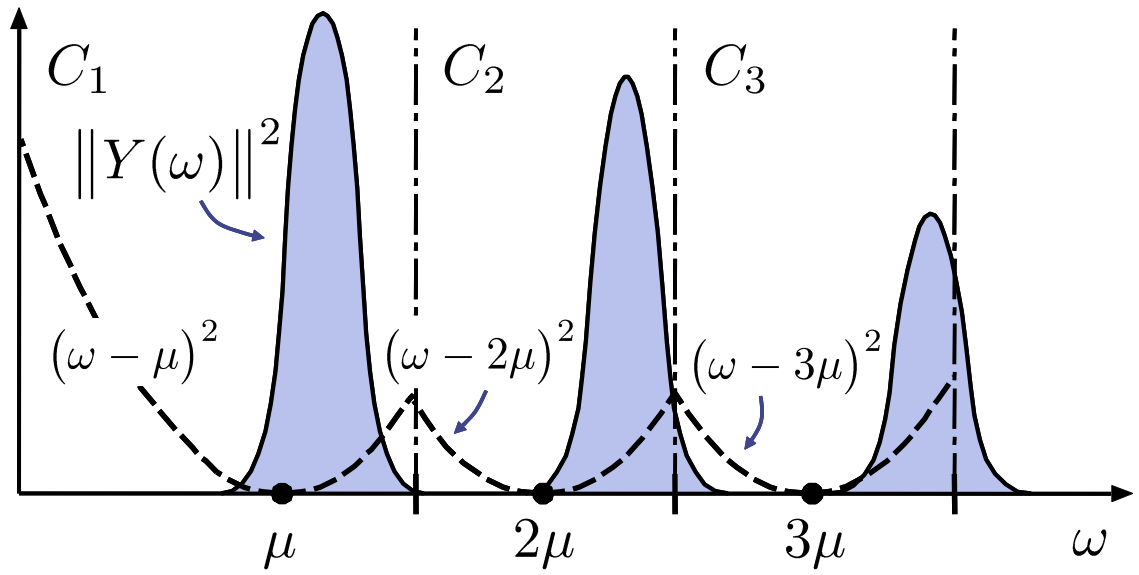


Figure 2.2 The objective function (Eq. (2.12)) can be monotonically decreased by iteratively updating C_n and μ while keeping the other fixed.

purpose, we introduce a binary mask function defined by

$$\mathbf{1}_{C_n}(\omega) = \begin{cases} 1, & \omega \in C_n \\ 0, & \omega \notin C_n \end{cases}, \tag{2.3}$$

where C_n is the set of the frequencies dominated by the n^{th} overtone, which satisfies, for any i and j such that $i \neq j$,

$$C_i \bigcap C_j = \emptyset. \tag{2.4}$$

If we decide not to discard any of the power spectrum portions, then

$$\bigcup_{n=1}^{N} C_n = \mathbb{R}(-\infty, \infty), \tag{2.5}$$

and thus for all $\omega \in \mathbb{R}$,

$$\sum_{n=1}^{N} \mathbf{1}_{C_n}(\omega) = 1, \tag{2.6}$$

because it is proved by the formula:

$$\mathbf{1}_{\bigcup_{i=1}^{I} C_i}(\omega) = \sum_{i=1}^{I} \mathbf{1}_{C_i}(\omega) - \sum_{i,j:i<j} \mathbf{1}_{C_i \bigcap C_j}(\omega) + \sum_{i,j,k:i<j<k} \mathbf{1}_{C_i \bigcap C_j \bigcap C_k}(\omega) - \cdots. \tag{2.7}$$

Using such a masking function, one is able to describe a masked power spectrum portion by

$$\mathbf{1}_{C_n}(\omega)\big\|Y(\omega)\big\|^2, \tag{2.8}$$

that correspond to the n^{th} overtone. Therefore, we can apply the same method described in Subsection 2.2.2 and

$$\int_{-\infty}^{\infty} \big(\omega - \mu_n\big)^2 \mathbf{1}_{C_n}(\omega) \big\|Y(\omega)\big\|^2 \mathrm{d}\omega \tag{2.9}$$

corresponds to the cost function for the frequency estimate of the n^{th} overtone. As we want to make the cost function as small as possible not only for the n^{th} overtone but for all the components at the same time, we should write as follows the objective function in the single voice case:

$$\sum_{n=1}^{N} \int_{-\infty}^{\infty} \big(\omega - \mu_n\big)^2 \mathbf{1}_{C_n}(\omega) \big\|Y(\omega)\big\|^2 \mathrm{d}\omega. \tag{2.10}$$

If we further assume that the overtone frequencies are integer multiplies of the fundamental frequency μ such that

$$\mu_n = n\mu, \tag{2.11}$$

then the objective function can be written further as

$$\sum_{n=1}^{N} \int_{-\infty}^{\infty} \big(\omega - n\mu\big)^2 \mathbf{1}_{C_n}(\omega) \big\|Y(\omega)\big\|^2 \mathrm{d}\omega. \tag{2.12}$$

This objective function can be monotonically decreased by iteratively updating C_n and μ while keeping the other fixed. In each iteration, C_n and μ should be updated to

$$C_n = \Big\{ \ \omega \ : \ n = \underset{n'}{\operatorname{argmin}} \ (\omega - n'\mu)^2 \Big\}, \tag{2.13}$$

$$\mu = \frac{\displaystyle\sum_{n=1}^{N} n \int_{-\infty}^{\infty} \omega \ \mathbf{1}_{C_n}(\omega) \big\|Y(\omega)\big\|^2 \mathrm{d}\omega}{\displaystyle\sum_{n=1}^{N} n^2 \int_{-\infty}^{\infty} \mathbf{1}_{C_n}(\omega) \big\|Y(\omega)\big\|^2 \mathrm{d}\omega}. \tag{2.14}$$

The update of C_n separates the observed power spectrum $\|Y(\omega)\|^2$ into clusters corresponding to the overtones using the F_0 estimated hypothetically at the previous step and the update of μ reestimates F_0 using these spectral clusters.

2.2.4 Multipitch Estimation and Source Separation

The method derived above is easily extendable to the multisource case. Let the binary mask function, used to extract the n^{th} partial component of the k^{th} source, be defined by

$$\mathbf{1}_{C_{k,n}}(\omega) = \begin{cases} 1, & \omega \in C_{k,n} \\ 0, & \omega \notin C_{k,n} \end{cases}, \tag{2.15}$$

where $C_{k,n}$ is the set of the frequencies dominated by the n^{th} overtone of the k^{th} source. It is assumed here again that

$$\sum_{k=1}^{K}\sum_{n=1}^{N}\mathbf{1}_{C_{k,n}}(\omega)=1. \tag{2.16}$$

Using such a masking function, one is able to describe a masked power spectrum portion by

$$\mathbf{1}_{C_{k,n}}(\omega)\big\|Y(\omega)\big\|^2, \tag{2.17}$$

that correspond to the n^{th} overtone of the k^{th} source. Therefore, if we denote by μ_k the F_0 estimate of the k^{th} source, then

$$\int_{-\infty}^{\infty}\big(\omega-n\mu_k\big)^2\mathbf{1}_{C_{k,n}}(\omega)\big\|Y(\omega)\big\|^2\mathrm{d}\omega \tag{2.18}$$

corresponds to the cost function for the frequency estimate of the n^{th} overtone of the k^{th} source. As we want to make the cost function as small as possible not only for this component but for all the components at the same time, we should write as follows the objective function:

$$\sum_{k=1}^{K}\sum_{n=1}^{N}\int_{-\infty}^{\infty}\big(\omega-n\mu_k\big)^2\mathbf{1}_{C_{k,n}}(\omega)\big\|Y(\omega)\big\|^2\mathrm{d}\omega. \tag{2.19}$$

This objective function can be monotonically decreased in a similar way by iteratively updating $C_{k,n}$ and μ_k while keeping the other fixed. In each iteration, $C_{k,n}$ and μ_k should be updated to

$$C_{k,n}=\Big\{\ \omega\ :\ (k,n)=\underset{k',n'}{\operatorname{argmin}}\ (\omega-n'\mu_{k'})^2\Big\}, \tag{2.20}$$

$$\mu_k=\frac{\displaystyle\sum_{n=1}^{N}n\int_{-\infty}^{\infty}\omega\ \mathbf{1}_{C_{k,n}}(\omega)\big\|Y(\omega)\big\|^2\mathrm{d}\omega}{\displaystyle\sum_{n=1}^{N}n^2\int_{-\infty}^{\infty}\mathbf{1}_{C_{k,n}}(\omega)\big\|Y(\omega)\big\|^2\mathrm{d}\omega}. \tag{2.21}$$

The update of $C_{k,n}$ separates the observed power spectrum $\|Y(\omega)\|^2$ into clusters each of which corresponds to an overtone of one particular source using the F_0s estimated hypothetically at the previous step and the update of μ_ks reestimates F_0s using these spectral clusters.

This iterative algorithm therefore consists of the source separation step and the multipitch estimation step, leading us to solve a joint optimization problem of source separation and multipitch estimation. We call this method "Harmonic Clustering". We reformulate this idea in Chapter 3 and try to explain it from the Bayesian point of view, which enables various extensions.

Chapter 3

Bayesian Harmonic Clustering

3.1 Introduction

In this chapter, we aim at extending the idea introduced in Chapter 2. The Harmonic Clustering is extended to a principle based on the estimation of the optimal fuzzy masking function for the clustering source by source of the power spectrum of the mixed sound of interest. Whether each of the spectral clusters has a harmonic structure or not is considered to be the criterion for this optimization problem. More specifically, we will consider a decomposition of the power spectrum of the mixed sound in which every spectral clusters has a harmonic structure as the optimal solution. We will show that this optimization problem is equivalent to the problem of minimizing the distortion between the power spectrum of the mixed sound and a mixture of spectral cluster models used as the clustering criterion. Meanwhile, from the viewpoint of statistical estimation, the distortion minimization procedure is none other than the regression analysis. It follows from this that the method constitutes in maximizing a likelihood function. Thus looking at the problem from the perspective of statistical estimation, the empirical constraints which are necessary in any undetermined problem can now be introduced, based on Bayes theorem, in the form of prior distributions. Moreover, as many empirical constraints which at first looked irrelevant can now be expressed with the same measure (that is, probability), the problem becomes more organized, and the perspective of a formulation and the intuitive meaning of the problem appear more clearly. Furthermore, through model selection, estimation of the optimal number of clusters, *i.e.* the number of sources, in the sense of posterior distribution is also performed.

As we explained in the preceding paragraph, this "extended" Harmonic Clustering can be understood as the minimization of the distortion between the power spectrum of the mixed

sound and a mixture of spectral cluster models, or as the optimal decomposition (clustering) of the power spectrum using spectral cluster models. Consequently, after deriving in the following section the specific form of a spectral cluster model from the ideal case of periodic signal models, we formulate in section Section 3.3 the problem separately from these two points of view and show that they eventually both lead to the same algorithm. Then, in sections Section 3.4 and Section 3.5, we show that optimal estimation under empirical constraints can be performed through Maximal *A Posteriori* estimation, and that a criterion for the source number estimation can be obtained from the model selection criterion.

3.2 Spectral Cluster Model

3.2.1 Definition of Fourier Transform Pair

Denoting by $Y(\omega)$ the Fourier transform of $y(t)$, the Fourier transform pair is defined by

$$\mathscr{F}\big[y(t)\big] = Y(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y(t) e^{-j\omega t} \mathrm{d}t \tag{3.1}$$

$$\mathscr{F}^{-1}\big[Y(\omega)\big] = y(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} Y(\omega) e^{j\omega t} \mathrm{d}\omega. \tag{3.2}$$

3.2.2 Definition of Analytic Signal

We define the analytic signal of a real signal $x(t)$ by

$$y(t) = x(t) + jz(t), \tag{3.3}$$

where $z(t)$ is the Hilbert transform of $x(t)$, defined as

$$z(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{x(\tau)}{t-\tau} \mathrm{d}\tau. \tag{3.4}$$

3.2.3 Gabor Transform Output of Periodic Signal Model

Assuming that all source signals are perfectly periodic in a short time range, we will consider as the spectral cluster model the output of the Gabor transform (STFT) of a periodic signal model around $t = 0$. Consider here as the k^{th} source signal model the analytic signal representation of a periodic signal given by

$$f_k(t) \triangleq \sum_{n=1}^{N} \widetilde{a}_{k,n} e^{j\left(n\mu_k t + \varphi_{k,n}\right)}, \quad t \in (-\infty, \infty), \tag{3.5}$$

where μ_k is the fundamental frequency, $\varphi_{k,n}$ the starting phase and $\widetilde{a}_{k,n}$ the amplitude of the n^{th} partial, respectively. Denoting by $w(t)$ a window function, let

$$g_k(t) \triangleq w(t) f_k(t) \tag{3.6}$$

be the short-time signal enhanced by $w(t)$ around $t = 0$. As the window function $w(t) \geq 0$ can be chosen arbitrary, we choose to use a Gaussian window. This type of STFT is called the Gabor transform. The Fourier transform of the left- and right-hand sides of Eq. (3.6) is, by the convolution theorem, given by

$$\begin{aligned} G_k(\omega) &= \frac{1}{\sqrt{2\pi}} W(\omega) * F_k(\omega) && (3.7)\\ &= \frac{1}{\sqrt{2\pi}} W(\omega) * \left(\sqrt{2\pi} \sum_{n=1}^{N} \widetilde{a}_{k,n} e^{j\varphi_{k,n}} \delta\left(\omega - n\mu_k\right) \right) && (3.8)\\ &= \sum_{n=1}^{N} \widetilde{a}_{k,n} e^{j\varphi_{k,n}} W(\omega - n\mu_k), && (3.9) \end{aligned}$$

where $F_k(\omega) \triangleq \mathscr{F}\big[f_k(t)\big]$ and $W(\omega) \triangleq \mathscr{F}\big[w(t)\big]$. As $w(t)$ is a Gaussian window, its Fourier transform is again a Gaussian-type function such that

$$W(\omega) = \exp\left(-\frac{\omega^2}{4\sigma^2} \right). \tag{3.10}$$

Hence, Eq. (3.9) can be written as

$$G_k(\omega) = \sum_{n=1}^{N} \widetilde{a}_{k,n} e^{j\varphi_{k,n}} \exp\left(-\frac{\big(\omega - n\mu_k\big)^2}{4\sigma^2} \right). \tag{3.11}$$

The power spectrum of Eq. (3.11) can be written as

$$\begin{aligned} \big\|G_k(\omega)\big\|^2 &= \left\| \sum_{n=1}^{N} \widetilde{a}_{k,n} e^{j\varphi_{k,n}} \exp\left(-\frac{\big(\omega - n\mu_k\big)^2}{4\sigma^2} \right) \right\|^2 \\ &= \sum_{n=1}^{N} \left\| \widetilde{a}_{k,n} e^{j\varphi_{k,n}} \exp\left(-\frac{\big(\omega - n\mu_k\big)^2}{4\sigma^2} \right) \right\|^2 \\ &\quad + \sum_{n \neq n'} \widetilde{a}_{k,n} \widetilde{a}_{k',n'} e^{j\left(\varphi_{k,n} - \varphi_{k',n'}\right)} \exp\left(-\frac{\big(\omega - n\mu_k\big)^2}{4\sigma^2} \right) \exp\left(-\frac{\big(\omega - n'\mu_{k'}\big)^2}{4\sigma^2} \right). \end{aligned} \tag{3.12}$$

If we now assume that the time-frequency components are sparsely distributed so that the partials rarely overlap, the second term could be negligibly smaller than the first term in

the above equation. This assumption justifies the additivity of power spectra and the power spectrum of the k^{th} source signal model is then expressed as a Gaussian mixture model:

$$\left\|G_k(\omega)\right\|^2 \approx \sum_{n=1}^{N} \widetilde{a}_{k,n}{}^2 \exp\left(- \frac{\left(\omega - n\mu_k\right)^2}{2\sigma^2} \right), \tag{3.13}$$

whose maxima are centered over prospective harmonics $\omega = n\mu_k$. Putting $a_{k,n} \triangleq \sqrt{2\pi}\sigma\widetilde{a}_{k,n}{}^2$, one finally obtains

$$\left\|G_k(\omega)\right\|^2 \approx \sum_{n=1}^{N} \frac{a_{k,n}}{\sqrt{2\pi}\sigma} \exp\left(- \frac{\left(\omega - n\mu_k\right)^2}{2\sigma^2} \right). \tag{3.14}$$

3.2.4 Constant Q Filterbank Output of Periodic Signal Model

Similarly, one can derive as well the constant Q filterbank output of a periodic signal model. Let the wavelet basis function defined by

$$\psi_{\alpha,t}(u) \triangleq \frac{1}{\sqrt{2\pi}\alpha}\psi\left(\frac{u-t}{\alpha}\right), \tag{3.15}$$

where α is the scale parameter, t the shift parameter and $\psi(u)$ an arbitrary analyzing wavelet that has the center frequency of 1 and satisfies the admissible condition. $\psi_{\alpha,t}(u)$ is used to measure the component of period α at time t. Now letting

$$f_k(u) \triangleq \sum_{n=1}^{N} \widetilde{a}_{k,n} e^{j\left(n\mu_k u + \varphi_{k,n}\right)}, \quad u \in (-\infty, \infty) \tag{3.16}$$

be the k^{th} source signal model, its continuous wavelet transform is defined by

$$W_k\left(\log\tfrac{1}{\alpha}, t\right) \triangleq \Big\langle f_k(u), \psi_{\alpha,t}(u) \Big\rangle_{u\in(-\infty,\infty)} \tag{3.17}$$

$$= \Big\langle F_k(\omega), \Psi_{\alpha,t}(\omega) \Big\rangle_{\omega\in(-\infty,\infty)}, \tag{3.18}$$

where $F_k(\omega) \triangleq \mathscr{F}[f_k(u)]$ and $\Psi_{\alpha,t}(\omega) \triangleq \mathscr{F}[\psi_{\alpha,t}(u)]$ The equality in the second line follows from the Parseval's theorem. Defining by $\Psi(\omega) \triangleq \mathscr{F}[\psi(u)]$, then the Fourier transform of Eq. (3.15) can be written as

$$\Psi_{\alpha,t}(\omega) = \Psi(\alpha\omega)e^{-j\omega t}, \tag{3.19}$$

and from Eq. (3.18), one obtains

$$W_k\left(\log\tfrac{1}{\alpha}, t\right) = \int_{-\infty}^{\infty} F_k(\omega)\Psi^*(\alpha\omega)e^{j\omega t}\mathrm{d}\omega. \tag{3.20}$$

One immediately realizes that Eq. (3.20) amounts to the inverse Fourier transform of $F_k(\omega)\Psi(\alpha\omega)$. $W_k(\log\frac{1}{\alpha},t)$ could thus be interpreted as an output signal from the subband filter with center frequency of $1/\alpha$ and with frequency response $\Psi(\alpha\omega)$ with the input $f_k(t)$. The Fourier transform of the k^{th} source signal model is given as

$$F_k(\omega)=\sqrt{2\pi}\sum_{n=1}^{N}\widetilde{a}_{k,n}e^{j\varphi_{k,n}}\delta\big(\omega-n\mu_k\big). \tag{3.21}$$

By substituting this result into Eq. (3.20), one obtains

$$W_k\left(\log\tfrac{1}{\alpha},t\right)=\sum_{n=1}^{N}\widetilde{a}_{k,n}e^{j\varphi_{k,n}}\Psi^*(an\mu_k)e^{jn\mu_k t}. \tag{3.22}$$

By changing the variable $x=\log\frac{1}{\alpha}$ and by putting $\Omega_k\triangleq\log\mu_k$, W_k can be expressed in the time-logfrequency domain:

$$W_k(x,t)=\sum_{n=1}^{N}\widetilde{a}_{k,n}\Psi^*\Big(ne^{-x+\Omega_k}\Big)e^{j\left(\varphi_{k,n}+ne^{\Omega_k}t\right)}. \tag{3.23}$$

We will henceforth simply call Ω_k the pitch frequency. As the frequency characteristic $\Psi(\omega)$ of the analyzing wavelet can be chosen arbitrarily, we use here the following unimodal real function whose maximum is taken at $\omega=1$ (see Fig. 3.1):

$$\Psi(\omega)=\begin{cases}\exp\left(-\dfrac{\big(\log\omega\big)^2}{4\sigma^2}\right) & (\omega>0)\\ 0 & (\omega\leqq 0)\end{cases}. \tag{3.24}$$

Eq. (3.23) is then given as

$$W_k(x,t)=\sum_{n=1}^{N}\widetilde{a}_{k,n}\exp\left(-\frac{\big(x-\Omega_k-\log n\big)^2}{4\sigma^2}\right)e^{j\left(\varphi_{k,n}+ne^{\Omega_k}t\right)}, \tag{3.25}$$

and the resulting power spectrum of Eq. (3.22) can be written as

$$\begin{aligned}\big\|W_k(x,t)\big\|^2&=\left\|\sum_{n=1}^{N}\widetilde{a}_{k,n}\exp\left(-\frac{\big(x-\Omega_k-\log n\big)^2}{4\sigma^2}\right)e^{j\left(\varphi_{k,n}+ne^{\Omega_k}t\right)}\right\|^2\\&=\sum_{n=1}^{N}\left\|\widetilde{a}_{k,n}\exp\left(-\frac{\big(x-\Omega_k-\log n\big)^2}{4\sigma^2}\right)e^{j\left(\varphi_{k,n}+ne^{\Omega_k}t\right)}\right\|^2\\&\quad+\sum_{n\neq n'}\widetilde{a}_{k,n}\widetilde{a}_{k',n'}\exp\left(-\frac{\big(x-\Omega_k-\log n\big)^2}{4\sigma^2}\right)\\&\qquad\exp\left(-\frac{\big(x-\Omega_{k'}-\log n'\big)^2}{4\sigma^2}\right)e^{j\left(ne^{\Omega_k}t+n'e^{\Omega_{k'}}t+\varphi_{k,n}+\varphi_{k',n'}\right)}.\end{aligned} \tag{3.26}$$

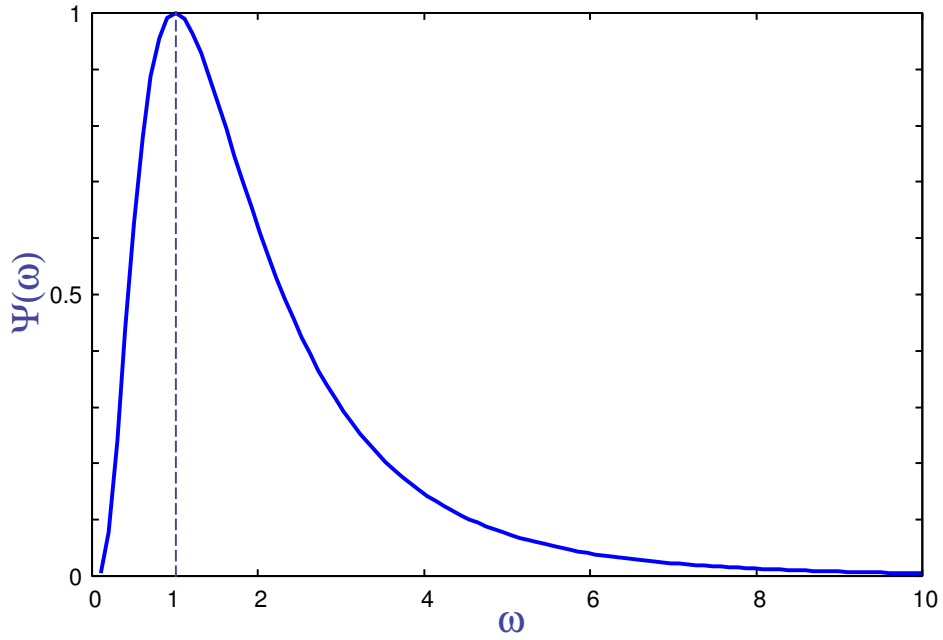


Figure 3.1 *Frequency response* $\Psi(\omega)$ *given by Eq. (3.24) for* $\sigma = \frac{1}{2}$.

If we assume here again that the time-frequency components are sparsely distributed so that the partials rarely overlap, the second term could be negligibly smaller than the first term in the above equation. Putting $a_{k,n} \triangleq \sqrt{2\pi}\sigma\|\widetilde{a}_{k,n}\|^2$ for simplicity of notation, one obtains a Gaussian mixture model whose maxima are centered over prospective harmonics $x = \Omega_k(t) + \log n$:

$$\left\|W(x,t)\right\|^2 \approx \sum_{k=1}^{K}\sum_{n=1}^{N} \frac{a_{k,n}}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(x - \Omega_k - \log n\right)^2}{2\sigma^2} \right), \tag{3.27}$$

whose graphical representation can be seen in Fig. 4.1. Let us denote simply by $\|W(x)\|^2$ the power spectrum $\|W(x,0)\|^2$ and consider it as the spectral cluster model. We are now able to assess how clearly a harmonic structure appears in a spectral cluster by measuring the distance between the cluster and this cluster model.

3.3 Principle

3.3.1 Optimal Separation of Power Spectrum

We will henceforth suppose the situation where we obtain the observed spectrum by constant Q analysis. We formulate the problem of the decomposition of the observed power

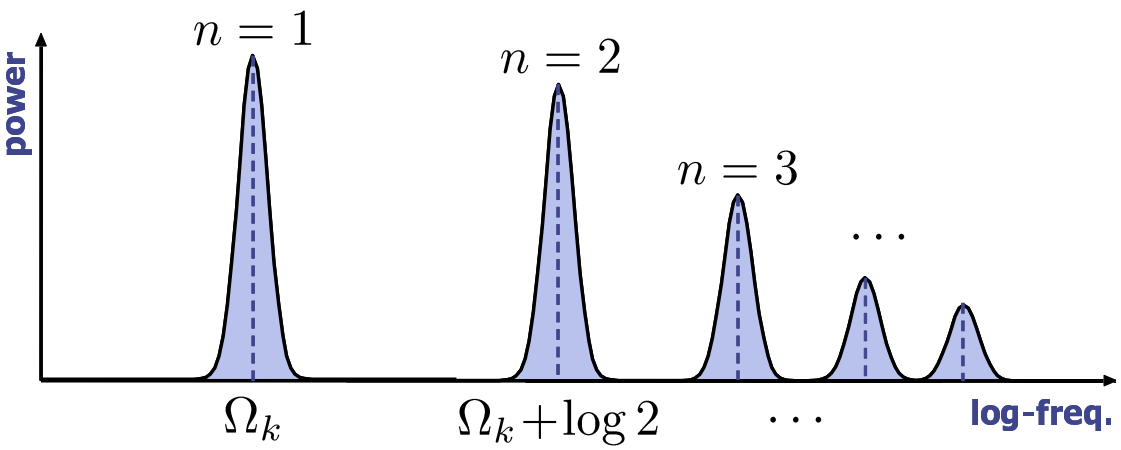


Figure 3.2 *Graphical representation of Eq. (3.27).*

spectrum into distinct clusters, which is said to be 'optimal' when all the clusters are harmonically structured. Let $\boldsymbol{\Theta}$ refers to $\{\Omega_k, \{a_{k,n}\}_{n=1}^{N}\}_{k=1}^{K}$.

We define by $\|Y(x)\|^2$ the power spectrum of the signal of interest obtained by the constant Q analysis. Let us introduce a spectral masking function $m_k(x)$ that extracts the components associated with the k^{th} source from $\|Y(x)\|^2$. For $x \in \mathbb{R}$, $m_k(x)$ indicates the percentage of the portion of $\|Y(x)\|^2$ shared to the k^{th} source, such that satisfies

$$\sum_{k=1}^{K} m_k(x) = 1 \tag{3.28}$$

$$0 < m_k(x) < 1, \quad k \in \{1, \cdots, K\}. \tag{3.29}$$

Assuming again additivity of power spectra, a portion of the observed power spectrum is thus given arbitrarily by

$$m_k(x)\big\|Y(x)\big\|^2, \quad x \in (-\infty, \infty), \tag{3.30}$$

which we call a "spectral cluster". As we expect the spectral cluster to be associated with a single harmonic structure, we need to introduce a measure function that specifies how clearly a harmonic structure appears in this spectral cluster. One possible measure function may be the I divergence [30] (The reason of this choice will be made clear in Subsection 3.3.2.) between $m_k(x)\|Y(x)\|^2$ and the spectral cluster model $\|W_k(x)\|^2$ we derived in Subsection 3.2.4:

$$\int_{-\infty}^{\infty} \left(m_k(x)\big\|Y(x)\big\|^2 \log \frac{m_k(x)\big\|Y(x)\big\|^2}{\|W_k(x)\|^2} - \left(m_k(x)\big\|Y(x)\big\|^2 - \|W_k(x)\|^2 \right) \right) \mathrm{d}x. \tag{3.31}$$

The more clearly the harmonic structure appears in $m_k(x)\|Y(x)\|^2$, the smaller this value may be. The optimal clustering achieved by minimizing their sum with respect to $m_k(x)$

and $\boldsymbol{\Theta}$:

$$\sum_{k=1}^{K}\int_{-\infty}^{\infty}\left(m_k(x)\big\|Y(x)\big\|^2\log\frac{m_k(x)\big\|Y(x)\big\|^2}{\|W_k(x)\|^2}-\Big(m_k(x)\big\|Y(x)\big\|^2-\|W_k(x)\|^2\Big)\right)\mathrm{d}x, \quad (3.32)$$

tries to make all separate clusters to be harmonically structured.

In the same way, let us introduce a spectral masking function $m_{k,n}(x)$ that extracts the components associated with the n^{th} partial of the k^{th} source from $\|Y(x)\|^2$. For $x\in\mathbb{R}$, $m_{k,n}(x)$ indicates the percentage of the portion of $\|Y(x)\|^2$ shared to the n^{th} partial of the k^{th} source, such that satisfies

$$\sum_{k=1}^{K}\sum_{n=1}^{N}m_{k,n}(x)=1 \quad (3.33)$$

$$0<m_{k,n}(x)<1,\quad k\in\{1,\cdots,K\},\ n\in\{1,\cdots,N\} \quad (3.34)$$

a portion of the observed power spectrum is thus given arbitrarily by

$$m_{k,n}(x)\big\|Y(x)\big\|^2,\quad x\in(-\infty,\infty) \quad (3.35)$$

which we call a "spectral cluster". In the same way, the optimal clustering can be achieved by minimizing

$$\Phi(\boldsymbol{\Theta},m)=\sum_{k=1}^{K}\sum_{n=1}^{N}\int_{-\infty}^{\infty}\left(m_{k,n}(x)\big\|Y(x)\big\|^2\log\frac{m_{k,n}(x)\big\|Y(x)\big\|^2}{\mathcal{W}_{k,n}(x)}-\Big(m_{k,n}(x)\big\|Y(x)\big\|^2-\mathcal{W}_{k,n}(x)\Big)\right)\mathrm{d}x. \quad (3.36)$$

with respect to $\boldsymbol{\Theta}$ and $m_{k,n}(x)$. To do so, we shall find it most convenient to minimize this objective function recursively with respect to $m_{k,n}(x)$ and $\boldsymbol{\Theta}$ while keeping the other fixed. As both steps necessarily decreases the objective function, which is bounded by below, the convergence of this recursive algorithm is thus guaranteed.

We shall first derive the update equation for the spectral masking function $m_{k,n}(x)$ that minimizes $\Phi(\boldsymbol{\Theta},m)$ with fixed $\boldsymbol{\Theta}$. Adding to the objective function the Lagrange multiplier term that ensures Eq. (3.33):

$$-\int_{-\infty}^{\infty}\lambda(x)\left(\sum_{k=1}^{K}\sum_{n=1}^{N}m_{k,n}(x)-1\right)\mathrm{d}x, \quad (3.37)$$

its partial derivative with respect to $m_{k,n}(x)$ is given as

$$\frac{\partial\Phi(\boldsymbol{\Theta},m)}{\partial m}=\big\|Y(x)\big\|^2\left(\log\frac{\mathcal{W}_{k,n}(x)}{m_{k,n}(x)}-1\right)-\lambda(x). \quad (3.38)$$

Setting this to 0, one obtains

$$m_{k,n}(x) = \mathcal{W}_{k,n}(x) \exp\left(\frac{\left\|Y(x)\right\|^2}{\lambda(x)} - 1\right). \tag{3.39}$$

From Eq. (3.33), the Lagrange multiplier $\lambda(x)$ is given as

$$\sum_{k=1}^{K}\sum_{n=1}^{N} \mathcal{W}_{k,n}(x) \exp\left(\frac{\left\|Y(x)\right\|^2}{\lambda(x)} - 1\right) = 1, \tag{3.40}$$

which yields

$$\widehat{m}_{k,n}(x) = \frac{\mathcal{W}_{k,n}(x)}{\sum_k \sum_n \mathcal{W}_{k,n}(x)}. \tag{3.41}$$

Substituting this result into Eq. (4.31), we obtain

$$\Phi(\boldsymbol{\Theta}, \widehat{m}) = \int_{-\infty}^{\infty} \left(\left\|Y(x)\right\|^2 \log \frac{\left\|Y(x)\right\|^2}{\sum_k \sum_n \mathcal{W}_{k,n}(x)} - \left(\left\|Y(x)\right\|^2 - \sum_k \sum_n \mathcal{W}_{k,n}(x) \right) \right) \mathrm{d}x, \tag{3.42}$$

from which we see that what we are trying to minimize w.r.t $\boldsymbol{\Theta}$ is the I divergence between the whole observed power spectrum and the sum of all the spectral cluster models.

3.3.2 Minimization of Distortion Measure

Optimally fitting a parametric function with respect to observed values corresponds, from the viewpoint of statistical estimation, to regression analysis. That is, if we consider that the observations $\|Y(x_i)\|^2$ at the discrete points x_i are generated from the regression model $\|W(x)\|^2$ with a randomly oscillating noise, one can come back naturally to a maximum likelihood estimation problem. Denoting by

$$P\big(\|Y(x_i)\|^2 \big| \boldsymbol{\Theta}\big) \tag{3.43}$$

the output probability of observation $\|Y(x_i)\|^2$ from the regression model $\|W(x)\|^2$ with parameter $\boldsymbol{\Theta}$ (in other words, the likelihood of the parameter $\boldsymbol{\Theta}$ of the regression model with respect to the observation $\|Y(x_i)\|^2$), our goal is to maximize the joint probability that all the observations $\boldsymbol{Y} = (\|Y(x_1)\|^2, \cdots, \|Y(x_I)\|^2)^T$ were generated independently by the regression model,

$$P(\boldsymbol{Y}|\boldsymbol{\Theta}) = \prod_{i=1}^{I} P\big(\|Y(x_i)\|^2 \big| \boldsymbol{\Theta}\big), \tag{3.44}$$

or its logarithm (hereafter mentioned as log-likelihood)

$$\log P(\boldsymbol{Y}|\boldsymbol{\Theta}) = \sum_{i=1}^{I} \log P\big(\|Y(x_i)\|^2\big|\boldsymbol{\Theta}\big). \tag{3.45}$$

For example, if we now consider the relation

$$\big\|Y(x_i)\big\|^2 = \big\|W(x_i)\big\|^2 + \epsilon_i, \tag{3.46}$$

it is often assumed that ϵ_i is a Gaussian white noise, *i.e.* $\epsilon_i \sim \mathcal{N}(0, \nu^2)$, and in this case expression (3.43) is defined as

$$P\big(\|Y(x_i)\|^2\big|\boldsymbol{\Theta}\big) = \frac{1}{\sqrt{2\pi}\nu} \exp\left(-\frac{\Big(\big\|Y(x_i)\big\|^2 - \big\|W(x_i)\big\|^2\Big)^2}{2\nu^2}\right). \tag{3.47}$$

Substituting it into (3.45), the quantity to maximize becomes

$$\sum_{i=1}^{I}\left(\log\frac{1}{\sqrt{2\pi}\nu} - \frac{\Big(\big\|Y(x_i)\big\|^2 - \big\|W(x_i)\big\|^2\Big)^2}{2\nu^2}\right), \tag{3.48}$$

and we thus see that this is equivalent to the least mean square estimation problem between $\|Y(x)\|^2$ and $\|W(x)\|^2$. However, despite the fact that $\|Y(x)\|^2$ is a power spectrum distribution density, the above likelihood function $P(\|Y(x_i)\|^2|\boldsymbol{\Theta})$ is non-zero even when $\|W(x_i)\|^2 < 0$. Of course, $P\big(\|Y(x_i)\|^2\big|\boldsymbol{\Theta}\big)$ does not need to be a Gaussian distribution, and for other distribution shapes the essential interpretation as a regression analysis problem is not lost. Here, it is desirable that $P(\|Y(x_i)\|^2|\boldsymbol{\Theta})$ is only defined for $\|W(x_i)\|^2 \geqq 0$, and the Poisson distribution is a representative example of such a probability density function. Poisson distribution is usually defined as a probability density function of random variables on non-negative integers, but it can be extended to a probability density function of random variables on all non-negative real numbers. Distinguishing it from the usual Poisson distribution, we will call it continuous Poisson distribution. The continuous Poisson distribution of $\|Y(x_i)\|^2$ with parameter $\|W(x)\|^2$ is given by

$$P\big(\|Y(x_i)\|^2\big|\boldsymbol{\Theta}\big) = \frac{e^{-\left\|W(x_i)\right\|^2}\Big(\big\|W(x_i)\big\|^2\Big)^{\left\|Y(x_i)\right\|^2}}{\Gamma\Big(\big\|Y(x_i)\big\|^2 + 1\Big)}, \tag{3.49}$$

where $\Gamma(\cdot)$ is the Gamma function

$$\Gamma(z) \triangleq \int_0^\infty e^{-t}t^{z-1}\mathrm{d}t, \tag{3.50}$$

and the likelihood is defined as 0 when $\|W(x)\|^2$ is negative. The shape of the distribution of the likelihood of $\mathbf{\Theta}$ with respect to $\|W(x_i)\|^2$ is shown in Fig. 3.3.2. Substituting this expression into (3.45), we obtain the log-likelihood to maximize:

$$L(\mathbf{\Theta}) \triangleq \sum_{i=1}^{I} \log P\big(\|Y(x_i)\|^2 \big| \mathbf{\Theta}\big) \tag{3.51}$$

$$= \sum_{i=1}^{I} \left(\big\|Y(x_i)\big\|^2 \log \big\|W(x_i)\big\|^2 - \big\|W(x_i)\big\|^2 - \log \Gamma\Big(\big\|Y(x_i)\big\|^2 + 1\Big)\right). \tag{3.52}$$

As shown in Fig. 3.3.2, the above distribution is a unimodal distribution reaching its maximum only when $\|Y(x)\|^2 = \|W(x)\|^2$, which implies as expected that this maximum likelihood problem amounts to a model fitting one. In the same way as we have shown that when the likelihood function is considered to be a Gaussian distribution the maximum likelihood problem becomes equivalent to a least-mean square estimation, the maximum likelihood problem under the above continuous Poisson distribution type likelihood function is equivalent to the minimization with respect to $\big\|W(x)\big\|^2$ of a distortion measure between distributions called I-divergence:

$$\sum_{i=1}^{I} \left(\big\|Y(x_i)\big\|^2 \log \frac{\big\|Y(x_i)\big\|^2}{\big\|W(x_i)\big\|^2} - \Big(\big\|Y(x_i)\big\|^2 - \big\|W(x_i)\big\|^2\Big)\right). \tag{3.53}$$

This is clear if we compare this expression to Eq. (3.52). As shown in Fig. 3.3.2, the distortion measure inside the parentheses in Eq. (3.53) is a non-symmetrical measure giving more penalty to positive errors, and thus emphasizes the goodness of fitting between spectral peaks. In that regard, it is similar to the Itakura-Saito distance [54] derived in Linear Predictive Coding (LPC).

We have explained the model fitting from a statistical estimation point of view. From the above perspective, we can redefine the problem as a Maximum *A Posteriori* estimation problem by introducing very naturally a prior distribution $P(\mathbf{\Theta})$, which we will explain in more details later. In the next subsection, we show how one can derive an efficient iterative estimation algorithm.

3.3.3 Iterative Maximum Likelihood Estimation

Goto [45], by considering hypothetically the frequencies as observation data and the normalized power spectrum as the probability distribution of the observation data, uses the EM

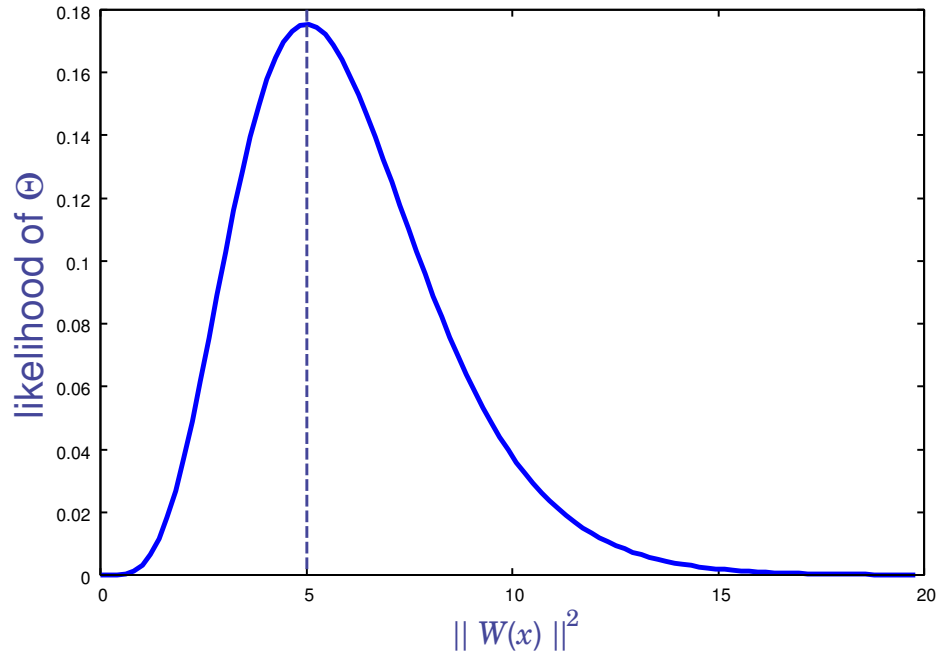


Figure 3.3 *Graphical representation of the likelihood function (3.49) for* $\|Y(x)\|^2 = 5$.

algorithm to maximize the probability that the whole observation data have been generated by a statistical model represented by a GMM. However, from a statistical signal processing viewpoint, it is not obvious whether the assumption that the frequencies behave stochastically is appropriate or not. As the power spectrum $\|Y(x)\|^2$ is actually not a probability distribution and the mixed sound model $\|W(x)\|^2$ is also not a statistical model, formulas from the probability theory (Bayes theorem, marginalization operations, etc.) cannot be applied in a rigorous manner to their distributions, and the fact that the EM algorithm derived using Bayes' rule could be used to perform approximation between them is thus definitely non-trivial. The goal of this subsection is to derive an iterative estimation algorithm formally equivalent to the EM algorithm without making use of Bayes' rule. This derivation justifies of course our method, but eventually also supports simultaneously the validity of Goto's method.

The goal of the problem (maximum likelihood estimation) is now

$$\widehat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}}\, L(\boldsymbol{\Theta}). \tag{3.54}$$

Looking back at Eq. (3.14), $\big\|W(x)\|^2$ can be written as a sum over k and n of terms of the form

$$\mathcal{W}_{k,n}(x) \triangleq \frac{a_{k,n}}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\big(x - \Omega_k - \log n\big)^2}{\sigma^2} \right),$$

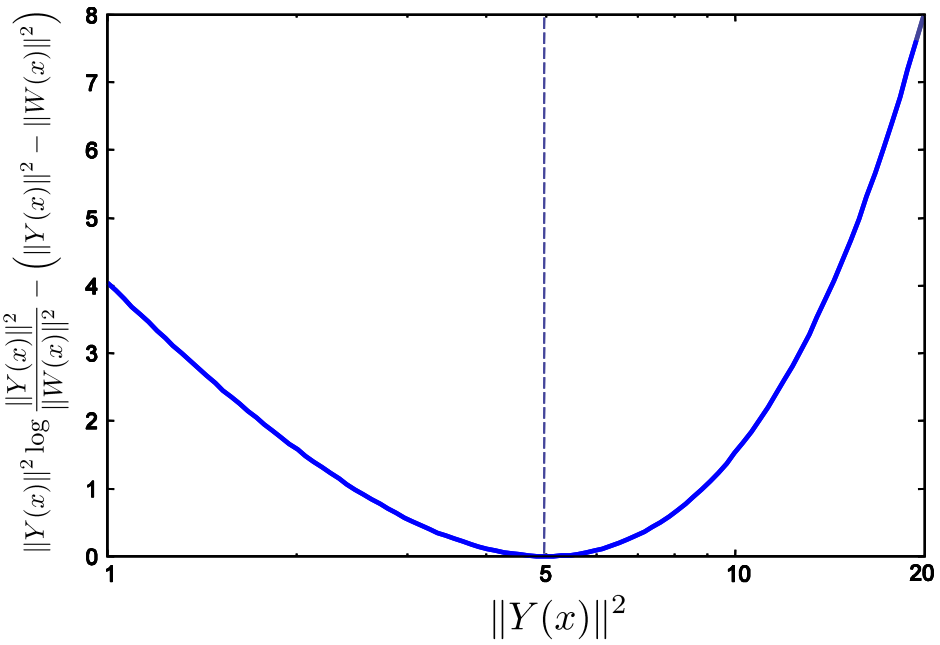


Figure 3.4 *The distortion measure inside the parentheses in (3.53) for* $\|W(x)\|^2 = 5$.

and one can thus write $L(\mathbf{\Theta})$ as

$$L(\mathbf{\Theta}) = \sum_{i=1}^{I} \left(\left\|Y(x_i)\right\|^2 \log \sum_{k=1}^{K} \sum_{n=1}^{N} \mathcal{W}_{k,n}(x_i) - \sum_{k=1}^{K} \sum_{n=1}^{N} \mathcal{W}_{k,n}(x_i) - \log \Gamma\left(\left\|Y(x_i)\right\|^2 + 1\right) \right). \tag{3.55}$$

Approximating the second term in the parentheses above by a Gaussian integral, we get

$$\sum_{i=1}^{I} \mathcal{W}_{k,n}(x_i) \approx \int_{-\infty}^{\infty} \mathcal{W}_{k,n}(x) \mathrm{d}x = a_{k,n}, \tag{3.56}$$

and thus $\sum_i \sum_{k,n} \mathcal{W}_{k,n}(x_i) \approx \sum_{k,n} a_{k,n}$. However, one cannot obtain analytically $\mathbf{\Theta}$ maximizing the above $L(\mathbf{\Theta})$. The specific reason for this is that $L(\mathbf{\Theta})$ has a nonlinear term expressed as the logarithm of a sum of several exponential terms.

If we now notice that the logarithm function is convex, introducing arbitrary weight functions $m_{k,n}(x)$ such that $\forall x$,

$$\sum_{k=1}^{K} \sum_{n=1}^{N} m_{k,n}(x) = 1, \quad \forall k, n : 0 < m_{k,n}(x) < 1, \tag{3.57}$$

we obtain the following inequality:

$$\begin{aligned} L(\boldsymbol{\Theta}) &= \sum_{i=1}^{I} \left(\left\| Y(x_i) \right\|^2 \log \sum_{k=1}^{K} \sum_{n=1}^{N} m_{k,n}(x_i) \frac{\mathcal{W}_{k,n}(x_i)}{m_{k,n}(x_i)} - \log \Gamma\Big(\left\| Y(x_i) \right\|^2 + 1 \Big) \right) - \sum_{k=1}^{K} \sum_{n=1}^{N} a_{k,n} \\ &\geqq \sum_{i=1}^{I} \left(\sum_{k=1}^{K} \sum_{n=1}^{N} \left\| Y(x_i) \right\|^2 m_{k,n}(x_i) \log \frac{\mathcal{W}_{k,n}(x)}{m_{k,n}(x_i)} - \log \Gamma\Big(\left\| Y(x_i) \right\|^2 + 1 \Big) \right) - \sum_{k=1}^{K} \sum_{n=1}^{N} a_{k,n}. \end{aligned} \tag{3.58}$$

Let us denote the right-hand side of this inequality by $L^{-}[\boldsymbol{\Theta}, m]$:

$$\begin{aligned} &L^{-}[\boldsymbol{\Theta}, m] \\ &\triangleq \sum_{i=1}^{I} \left(\sum_{k=1}^{K} \sum_{n=1}^{N} \left\| Y(x_i) \right\|^2 m_{k,n}(x_i) \log \frac{\mathcal{W}_{k,n}(x_i)}{m_{k,n}(x_i)} - \log \Gamma\Big(\left\| Y(x_i) \right\|^2 + 1 \Big) \right) - \sum_{k=1}^{K} \sum_{n=1}^{N} a_{k,n}. \end{aligned} \tag{3.59}$$

What must be noticed in the above inequality is not only that a lower bound function (right-hand side) has been obtained for $L(\boldsymbol{\Theta})$, but that in this lower bound function $L^{-}(\boldsymbol{\Theta}, m)$ the exponential inside $\mathcal{W}_{k,n}(x)$ has disappeared and become a second-order function in Ω_k, thus suggesting that it should be possible to obtain analytically Ω_k maximizing $L^{-}(\boldsymbol{\Theta}, m)$. Using this fact, we develop hereafter a method to increase $L(\boldsymbol{\Theta})$ indirectly using $L^{-}(\boldsymbol{\Theta}, m)$.

$L^{-}(\boldsymbol{\Theta}, m)$ contains a new variable $m_{k,n}(x)$ which did not appear in $L(\boldsymbol{\Theta})$. For any fixed $\boldsymbol{\Theta}$, if we maximize the lower bound function with respect to $m_{k,n}(x)$, equality is reached in the inequality, with $L^{-}(\boldsymbol{\Theta}, m)$ always staying smaller than $L(\boldsymbol{\Theta})$. The latter is a direct consequence of the inequality while the former can be verified by looking for $m_{k,n}(x)$ maximizing $L^{-}(\boldsymbol{\Theta}, m)$. Let us first differentiate with respect to $m_{k,n}(x_i)$ the lower bound function to which the Lagrange multiplier term

$$-\sum_{i=1}^{I} \lambda_i \left(\sum_{k=1}^{K} \sum_{n=1}^{N} m_{k,n}(x_i) - 1 \right) \tag{3.60}$$

has been added. We obtain

$$\frac{\partial L^{-}}{\partial m} = \left\| Y(x_i) \right\|^2 \left(\log \frac{\mathcal{W}_{k,n}(x_i)}{m_{k,n}(x_i)} - 1 \right) - \lambda_i, \tag{3.61}$$

and putting this to 0, we get

$$m_{k,n}(x_i) = \mathcal{W}_{k,n}(x_i) \exp\left(\frac{\left\| Y(x_i) \right\|^2}{\lambda_i} - 1 \right). \tag{3.62}$$

According to condition (3.57),

$$\sum_{k=1}^{K}\sum_{n=1}^{N}\mathcal{W}_{k,n}(x_i)\exp\left(\frac{\left\|Y(x_i)\right\|^2}{\lambda_i}-1\right)=1, \tag{3.63}$$

and the Lagrange multipliers λ_i can thus be obtained. We eventually get

$$m_{k,n}(x_i)=\frac{\mathcal{W}_{k,n}(x_i)}{\sum\limits_{k}\sum\limits_{n}\mathcal{W}_{k,n}(x_i)}. \tag{3.64}$$

Introducing this result into Eq. (3.59), we can verify that indeed $L^{-}(\boldsymbol{\Theta},m)=L(\boldsymbol{\Theta})$.

From this state of equality, if we now increase $L^{-}(\boldsymbol{\Theta},m)$ with respect to $\boldsymbol{\Theta}$, automatically $L(\boldsymbol{\Theta})$ shall also increase. This is due to the fact that the convex inequality ensures that $L(\boldsymbol{\Theta})$ is necessarily larger than the increased $L^{-}(\boldsymbol{\Theta},m)$. From the above, we can see that performing alternately the maximization of $L^{-}(\boldsymbol{\Theta},m)$ with respect to $m_{k,n}(x)$ and an increase of $L^{-}(\boldsymbol{\Theta},m)$ with respect to $\boldsymbol{\Theta}$, $L(\boldsymbol{\Theta})$ will monotonically increase. The parameter estimation algorithm is thus composed of the two following steps:

Step 0 Set the initial parameters $\boldsymbol{\Theta}_0$, put $\ell=1$.

Step 1 $m^{(\ell)}=\underset{m}{\operatorname{argmax}}\,L^{-}(\boldsymbol{\Theta}^{(\ell-1)},m)$.

Step 2 Set $\boldsymbol{\Theta}^{(\ell)}$ as $\boldsymbol{\Theta}$ such that $L^{-}(\boldsymbol{\Theta},m^{(\ell)})\geqq L^{-}(\boldsymbol{\Theta}^{(\ell-1)},m^{(\ell)})$, put $\ell\leftarrow\ell+1$ and go back to Step 1.

As $L(\boldsymbol{\Theta})$ is bounded above, from the preceding discussion, we can see that the convergence of the iterative estimation algorithm is guaranteed.

A point which should particularly be noticed here is that the iterative estimation of the pitch frequencies Ω_k through the EM algorithm, which could not be obtained in the methods of Chazan et al. [22] and Jinachitra et al. [57], can now be performed. We shall give the details about the update equations of the model parameter set $\boldsymbol{\Theta}$ later, but let us verify here first that the update equation for Ω_k can be obtained analytically. Putting to 0 the partial derivative of $L^{-}(\boldsymbol{\Theta},m)$ with respect to Ω_k

$$\frac{\partial L^{-}}{\partial\Omega_k}=\sum_{i=1}^{I}\sum_{n=1}^{N}\left\|Y(x_i)\right\|^2 m_{k,n}(x_i)\frac{x_i-\Omega_k-\log n}{\sigma^2} \tag{3.65}$$

the update equation for Ω_k

$$\widehat{\Omega}_k=\frac{\sum\limits_{i=1}^{I}\sum\limits_{n=1}^{N}\left\|Y(x_i)\right\|^2 m_{k,n}(x_i)\big(x_i-\log n\big)}{\sum\limits_{i=1}^{I}\sum\limits_{n=1}^{N}\left\|X(x_i)\right\|^2 m_{k,n}(x_i)} \tag{3.66}$$

can be obtained analytically.

The iterative computation presented above eventually follows formally the same procedure as the EM algorithm, but as we do not assume that $\|Y(x)\|^2$ and $\|W(x)\|^2$ are probability distributions, its derivation method is slightly different from the original EM algorithm [33]. In that sense, the above derivation gives another interpretation of the EM algorithm.

3.4 Bayesian Harmonic Clustering

3.4.1 Maximum *A Posteriori* (MAP) Estimation

Based on the above preparatory work, from the viewpoint of statistical estimation as in Subsection 3.3.2, the empirical constraints which are necessary in underdetermined problems can be smoothly introduced through Bayes theorem, and many problems can be dealt with. Moreover, as many empirical constraints which at first looked irrelevant can now be expressed with the same measure (*i.e.*, probability), the problem becomes more organized, and the perspective of a formulation and the intuitive meaning of the problem appear more clearly.

First, from the Bayes theorem, the posterior probability of $\boldsymbol{\Theta}$ is given by

$$P(\boldsymbol{\Theta}|\boldsymbol{Y}) = \frac{P(\boldsymbol{Y}|\boldsymbol{\Theta})P(\boldsymbol{\Theta})}{P(\boldsymbol{Y})}. \tag{3.67}$$

It is then through the prior probability $P(\boldsymbol{\Theta})$ that the relation to the empirical constraints appears. Let us consider here the maximization of the posterior probability $P(\boldsymbol{\Theta}|\boldsymbol{Y})$:

$$\begin{aligned}
\operatorname*{argmax}_{\boldsymbol{\Theta}} P(\boldsymbol{\Theta}|\boldsymbol{Y}) &= \operatorname*{argmax}_{\boldsymbol{\Theta}} P(\boldsymbol{Y}|\boldsymbol{\Theta})P(\boldsymbol{\Theta}) && (3.68)\\
&= \operatorname*{argmax}_{\boldsymbol{\Theta}} \Big(\log P(\boldsymbol{Y}|\boldsymbol{\Theta}) + \log P(\boldsymbol{\Theta}) \Big) && (3.69)\\
&= \operatorname*{argmax}_{\boldsymbol{\Theta}} \Big(L(\boldsymbol{\Theta}) + \log P(\boldsymbol{\Theta}) \Big). && (3.70)
\end{aligned}$$

This is the Maximum *A Posteriori* estimation of the model parameters $\boldsymbol{\Theta}$. As can be seen in Eq. (3.70), the objective function in this case is only the objective function $L(\boldsymbol{\Theta})$ used in the discussion of Subsection 3.3.3, to which $\log P(\boldsymbol{\Theta})$ has been added. As $\log P(\boldsymbol{\Theta})$ does not depend on $m_{k,n}(x_i)$, the update equation of $m_{k,n}(x_i)$ stays the same as (3.64). If we can obtain update equations for $\boldsymbol{\Theta}$ from $L^-(\boldsymbol{\Theta}, m) + \log P(\boldsymbol{\Theta})$, then in the same was as in Subsection 3.3.3, we will be able to derive an iterative algorithm monotonically increasing $L(\boldsymbol{\Theta}) + \log P(\boldsymbol{\Theta})$.

3.4.2 Smoothness of Spectral Envelope

In speech and music, the empirical constraint that "the spectral envelope is smooth" is relatively largely accepted. The necessary condition such that the spectral envelope is smooth is that the values of $a_{k,n}$ and $a_{k,n-1}$ should be sufficiently close. Therefore, one strategy is to define the prior distribution such that the probability should get larger as the values of $a_{k,n}$ and $a_{k,n-1}$ become closer.

For simplicity, let us first suppose that in $\mathbf{\Theta}$, $\{\Omega_k\}$ and $\{a_{k,n}\}$ are independent, and furthermore that the $\{a_{k,n}\}$ are independent across sources. One can then separate the variables as follows:

$$P(\mathbf{\Theta}) = P(\Omega_1, \cdots, \Omega_K) \prod_k P(a_{k,1}, \cdots, a_{k,N}). \tag{3.71}$$

We can decompose furthermore $P(a_{k,1}, \cdots, a_{k,N})$ in

$$\begin{aligned} P(a_{k,1}, \cdots, a_{k,N}) &= P(a_{k,1}) P(a_{k,2}, \cdots, a_{k,N} | a_{k,1}) \\ &= P(a_{k,1}) P(a_{k,2}|a_{k,1}) P(a_{k,3}, \cdots, a_{k,N} | a_{k,1}, a_{k,2}) \\ &= \cdots \\ &= P(a_{k,1}) P(a_{k,2}|a_{k,1}) P(a_{k,3}|a_{k,1}, a_{k,2}) \cdots P(a_{k,N}|a_{k,1}, \cdots, a_{k,N}). \end{aligned} \tag{3.72}$$

If we now suppose that the power $a_{k,n}$ of the n-th harmonic component only depends on the power $a_{k,n-1}$ of the neighboring component, $P(a_{k,1}, a_{k,2} \cdots, a_{k,N})$ can be expressed as a Markov chain probability:

$$P(a_{k,1}, \cdots, a_{k,N}) \approx P(a_{k,1}) \prod_{n=2}^{N} P(a_{k,n}|a_{k,n-1}). \tag{3.73}$$

The probability $P(a_{k,n}|a_{k,n-1})$ should become larger as the powers of the neighboring harmonic components are closer. Moreover, as $a_{k,n}$ is a power (and thus non-negative), we would like to consider a probability distribution for which the probability density function is only defined for $a_{k,n} \geqq 0$. For example, let us assume it follows the Gamma distribution:

$$P(a_{k,n}|a_{k,n-1}) = \frac{(\gamma-1)^\gamma}{\Gamma(\gamma)} \frac{{a_{k,n}}^{\gamma-1} e^{-\frac{(\gamma-1)a_{k,n}}{a_{k,n-1}}}}{{a_{k,n-1}}^\gamma}. \tag{3.74}$$

This distribution's probability density function is only defined for $a_{k,n} \geqq 0$, and is a unimodal distribution which takes its maximum at the parameter $a_{k,n-1}$. $\gamma > 0$ is called shape parameter, and as the peak of the distribution becomes sharper as γ becomes larger, it can

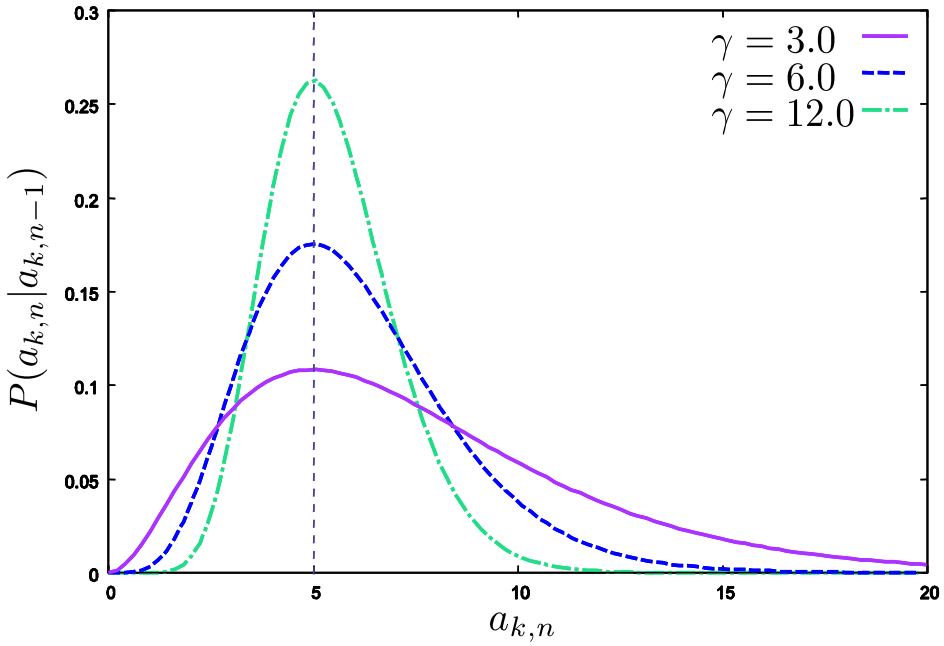


Figure 3.5 *Illustration of* $P(a_{k,n}|a_{k,n-1})$ *when* $a_{k,n-1} = 5$ *(shape parameter* $\gamma = 3, 6, 12$*).*

be considered as a constant, that is used for adjusting the effect of the prior distribution. An illustration of the Gamma distribution is shown in Fig. 3.4.2.

If we assume $P(\Omega_1, \cdots, \Omega_K)$ and $P(a_{k,1})$ follow uniform distributions (all values can be taken evenly), from the above we obtain that $\log P(\boldsymbol{\Theta})$ can be written specifically as

$$\log P(\boldsymbol{\Theta}) = \eta + K(N-1)\log\frac{(\gamma-1)^\gamma}{\Gamma(\gamma)} - \sum_k \left(\sum_{n=2}^{N} \frac{(\gamma-1)a_{k,n}}{a_{k,n-1}} + \sum_{n=2}^{N-1} \log a_{k,n} + \gamma \log a_{k,1} - (\gamma-1)\log a_{k,N} \right), \quad (3.75)$$

where $\eta = \log P(\Omega_1, \cdots, \Omega_K) + \sum_k \log P(a_{k,1}) = \text{const.}$

3.4.3 Update Equations for the Model Parameters

We can now finally derive the update equations of the Step 2 of Subsection 3.3.3. As explained earlier, we want to obtain $\boldsymbol{\Theta}$ increasing or maximizing $L^-(\boldsymbol{\theta}, m) + \log P(\boldsymbol{\Theta})$.

As $\log P(\boldsymbol{\Theta})$ does not depend on Ω_k, the update equation of μ_k is already given as presented in (3.66). The update equation of the $a_{k,n}$ is performed, for each k, by using sequentially from $n = 1$ to $n = N$ the following update equations (Coordinate Descent method), ensuring that $L^-(\boldsymbol{\theta}, m) + \log P(\boldsymbol{\Theta})$ does not decrease. Hereafter, let us write

$$\Phi_{k,n} = \sum_{i=1}^{I} \|X(x_i)\|^2 m_{k,n}(x_i). \quad (3.76)$$

The update equation for $a_{k,n}$ can then be obtained as follows. We first put to 0 the partial derivative of $L^-(\boldsymbol{\theta}, m) + \log P(\boldsymbol{\Theta})$ with respect to $a_{k,1}$,

$$\frac{1}{a_{k,1}}\Phi_{k,1} - 1 + \frac{(\gamma-1)a_{k,2}}{{a_{k,1}}^2} - \frac{\gamma}{a_{k,1}},$$

and obtain

$$\widehat{a}_{k,1} = \frac{\Phi_{k,1}-\gamma}{2} + \left(\frac{(\Phi_{k,1}-\gamma)^2}{4} + (\gamma-1)a_{k,2}\right)^{\frac{1}{2}}, \tag{3.77}$$

where $a_{k,2}$ in the above update equation is the value updated one step before. Then putting to 0 the partial derivative of $L^-(\boldsymbol{\theta}, m) + \log P(\boldsymbol{\Theta})$ with respect to $a_{k,n}(n = 2, \cdots, N-1)$,

$$\frac{1}{a_{k,n}}\Phi_{k,n} - 1 - \frac{\gamma-1}{a_{k,n-1}} + \frac{(\gamma-1)a_{k,n+1}}{{a_{k,n}}^2} - \frac{1}{a_{k,n}},$$

we obtain

$$\widehat{a}_{k,n} = \frac{a_{k,n-1}}{a_{k,n-1}+\gamma-1}\left(\frac{\Phi_{k,n}-1}{2} + \left(\frac{(\Phi_{k,n}-1)^2}{4} + \frac{(\gamma-1)(a_{k,n-1}+\gamma-1)a_{k,n+1}}{a_{k,n-1}}\right)^{\frac{1}{2}}\right), \tag{3.78}$$

where $a_{k,n-1}$ is the latest updated value and $a_{k,n+1}$ is the value updated one step before. Finally, putting to 0 the partial derivative with respect to $a_{k,N}$,

$$\frac{\partial L^-}{\partial a_{k,N}} = \frac{1}{a_{k,N}}\Phi_{k,N} - 1 - \frac{\gamma-1}{a_{k,N-1}} + \frac{\gamma-1}{a_{k,N}},$$

we obtain

$$\widehat{a}_{k,N} = \frac{a_{k,N-1}\big(\Phi_{k,N}+\gamma-1\big)}{a_{k,N-1}+\gamma-1}, \tag{3.79}$$

where $a_{k,N-1}$ is the latest updated value.

3.5 A Criterion for Source Number Estimation

3.5.1 Model Selection using Bayesian Information Criterion

Another important characteristic of Bayesian statistical inference is that a model structure selection criterion can be derived. Model structure specifies the model function class and the number of free parameters, but as here the function class is fixed, it indicates the number of free parameters. Model selection criterion is a criterion to determine comparatively which kind of model structure shall have a model which would be likely to generate a given observation data. Up to now, the discussion was done under the assumption that the number of sources K and the number of harmonic components N of the mixed sound model were

already known, but in general the number of sources mixed in the input mixed sound signal is unknown. Therefore, if one could derive a model selection criterion, it would lead a new criterion to estimate the number of sources K and the number of harmonic components N.

Meanwhile, as compared to the estimation of the number of sources K, estimating the number of harmonic components is not engineeringly such an important problem, hereafter for the sake of simplicity we shall assume that N is an experimentally fixed constant. Then, through comparison of all the mixed sound models for K varying from $K = 1$ to $K = \widetilde{K}$, where $\widetilde{K}$ is the maximum source number, and selection of the best model structure, the number of sources can be estimated.

We first express the model structure index as $M^{(K)}$, where the superscript refers to the number of sources, which is related to the number of free parameters. Similarly, denoting by $\boldsymbol{\Theta}^{(K)}$ the number of model parameters of a mixed sound model for which the number sources is K, the problem considered here is to find the model structure $M^{(K)}$ that maximizes the posterior probability of $M^{(K)}$,

$$P\big(M^{(K)}|\boldsymbol{Y}\big) = \frac{P\big(\boldsymbol{Y}|M^{(K)}\big)P\big(M^{(K)}\big)}{P\big(\boldsymbol{Y}\big)}, \tag{3.80}$$

where $\boldsymbol{Y}$ refers to the set of observations $\|Y(x_1)\|^2, \cdots, \|Y(x_I)\|^2$. Assuming that the prior probability $P(M^{(K)})$ of the model structure is a uniform distribution, the problem amounts to performing the maximum likelihood estimation of the model structure:

$$\widehat{M}^{(K)} = \underset{M^{(K)}}{\operatorname{argmax}}\, P\big(\boldsymbol{Y}|M^{(K)}\big). \tag{3.81}$$

In Subsection 3.3.2, as we assumed implicitly that K was fixed, the model structure index was actually omitted in the right-hand side of Eq. (3.44). If we now consider that K is an unknown variable, $P(\boldsymbol{Y}|\boldsymbol{\Theta})$ should be written more exactly $P(\boldsymbol{Y}|\boldsymbol{\Theta}^{(K)}, M^{(K)})$. Then, as

$$P\big(\boldsymbol{Y}, \boldsymbol{\Theta}^{(K)}|M^{(K)}\big) = P\big(\boldsymbol{Y}|\boldsymbol{\Theta}^{(K)}, M^{(K)}\big)P\big(\boldsymbol{\Theta}^{(K)}|M^{(K)}\big), \tag{3.82}$$

if we marginalize both sides with respect to $\boldsymbol{\Theta}^{(K)}$, from

$$P\big(\boldsymbol{Y}|M^{(K)}\big) = \int P\big(\boldsymbol{Y}|\boldsymbol{\Theta}^{(K)}, M^{(K)}\big)P\big(\boldsymbol{\Theta}^{(K)}|M^{(K)}\big)\mathrm{d}\boldsymbol{\Theta}^{(K)} \tag{3.83}$$

$$= \int \exp\big\{L(\boldsymbol{\Theta}^{(K)})\big\}P\big(\boldsymbol{\Theta}^{(K)}|M^{(K)}\big)\mathrm{d}\boldsymbol{\Theta}^{(K)}, \tag{3.84}$$

where, $L(\boldsymbol{\Theta}^{(K)}) = \log P(\boldsymbol{Y}|\boldsymbol{\Theta}^{(K)}, M^{(K)})$, we can obtain the desired model selection criterion. This is actually none other than the denominator of the right-hand side of Eq. (3.67). The

criterion for estimation of the model structure is thus the "marginal probability of the observation data", which tends to be disregarded in the context of maximum likelihood estimation and Maximum *A Posteriori* estimation of model parameters.

The next point that we have to discuss is how to obtain the marginal distribution of the above equation. One could think of computing numerically the integral with respect to $\boldsymbol{\Theta}^{(K)}$, but this would require considerable computational cost and is thus not realistic. Here we can use the well-known Bayesian Information Criterion (BIC)[7, 93], model evaluation criterion derived by approximating the marginal probability of the above equation. Its principle is based on the assumption that when the data number I (here referring to the number of discrete frequency points) is sufficiently large, as the integrand in the above equation concentrates in the vicinity of the value of the maximal likelihood estimator (or of the Maximum *A Posteriori* estimator) $\widehat{\boldsymbol{\Theta}}$, the integration value depends on the behavior in the neighborhood of $\widehat{\boldsymbol{\Theta}}$, and $L(\boldsymbol{\Theta}^{(K)})$ and $\log P(\boldsymbol{\Theta}^{(K)}|M^{(K)})$ can be approximated around $\widehat{\boldsymbol{\Theta}}^{(K)}$ respectively by their 2nd order and 0th order Taylor expansion [93]. This corresponds to approximating the posterior distribution of $\boldsymbol{\Theta}$ by a multidimensional Gaussian distribution centered on the value $\widehat{\boldsymbol{\Theta}}$ of the maximum likelihood estimator (or of the Maximum *A Posteriori* estimator), and in case the maximum likelihood estimator is asymptotically normal, this approximation is justified. The marginalization operation of (3.84) can then be easily performed. The above question can thus be approximated in the following way:

$$P\big(\boldsymbol{Y}|M^{(K)}\big) \approx \exp\Big\{L\big(\widehat{\boldsymbol{\Theta}}^{(K)}\big)\Big\}\, P\big(\widehat{\boldsymbol{\Theta}}^{(K)}|M^{(K)}\big)\big(2\pi\big)^{D^{(K)}/2} I^{-D^{(K)}/2} \Big|J\big(\widehat{\boldsymbol{\Theta}}^{(K)}\big)\Big|^{-1/2}, \quad (3.85)$$

where I is the number of elements of the observation time series $\boldsymbol{Y}$, $D^{(K)}$ is the number of free parameters in the mixed sound model $\|\mathcal{Y}(x)\|^2$ when the number of sources is K, and $J(\widehat{\boldsymbol{\Theta}}^{(K)})$ is the Fisher information matrix. Taking the logarithm of this equation, multiplying by -2 and approximating further, we obtain the BIC:

$$\begin{aligned} &-2\log P\big(\boldsymbol{Y}|M^{(K)}\big) \\ &\quad\approx -2L\big(\widehat{\boldsymbol{\Theta}}^{(K)}\big) + D^{(K)}\log I + \log\big|J\big(\widehat{\boldsymbol{\Theta}}^{(K)}\big)\big| - D^{(K)}\log\big(2\pi\big) - 2\log P\big(\widehat{\boldsymbol{\Theta}}^{(K)}|M^{(K)}\big) \\ &\quad\approx -2L\big(\widehat{\boldsymbol{\Theta}}^{(K)}\big) + D^{(K)}\log I. \end{aligned} \quad (3.86)$$

For more details on the above derivation, we shall refer to [7, 93].

3.5.2 Model Selection Algorithm

In this subsection, we present the global structure of the Bayesian Harmonic Clustering algorithm, including the Maximum *A Posteriori* estimation of the parameters and the model selection of the spectral cluster models.

1. Set the initial value $\widetilde{K}$ of the number of sources K.
2. Estimate the Maximum *A Posteriori* parameter $\widehat{\boldsymbol{\Theta}}^{(K)}$ using the iterative algorithm presented in Subsection 3.3.3.
   (a) Set the initial parameters $\boldsymbol{\Theta}$ $\widetilde{K}$ of the spectral cluster models. Detect the top $\widetilde{K}$ peaks of the power spectrum $\|Y(x)\|^2$ and use them as initial parameters Ω_k.
   (b) Update $m_{k,n}(x)$ through Eq. (3.64).
   (c) Update Ω_k through Eq. (3.66).
   (d) After updating $a_{k,n}$ through Eq. (3.77), Eq. (3.78) and Eq. (3.79), return to (b).

   After convergence, proceed to 3.
3. Compute BIC(K) through Eq. (3.86). If $K \neq 1$, proceed to 4. If $K = 1$, proceed to 5.
4. Find the spectral cluster model with smallest power,
$$\check{k} = \operatorname*{argmin}_k \sum_{n=1}^{N} a_{k,n}, \tag{3.87}$$
   and eliminate it. Set $K = K - 1$ and return to 2.
5. Find K minimizing BIC. For this model structure, look for the Maximum *A Posteriori* estimation Parameter
$$\widehat{K} = \operatorname*{argmin}_K \mathrm{BIC}(K) \Rightarrow \widehat{\boldsymbol{\Theta}}^{(\widehat{K})} \tag{3.88}$$
   as the final solution.

3.6 Experimental Evaluation

3.6.1 Condition

Considering music is the typical example of multipitch audio signal, the proposed method was tested on a framewise musical note estimation task using 8 pieces of real music performance data excerpted from RWC music database (the list of the experimental data is shown

Table 3.1 *Experimental conditions*

frequency analysis	Sampling rate	16 kHz
	frame shift	32 msec
	mother wavelet	Gabor function
	frequency resolution	12.0 cent
	frequency range	60–3000 Hz
proposed algorithm	initial # of harmonic kernels	10
	# of partials	8
	σ	3.0×10^{-3}
	$\bar{r}_n$	$0.6547 \times n^{-2}$
	$\bar{d}$	3.0
	ρ_n	$0.01 \times \frac{1}{n}$
PreFEst-core[45]	pitch resolution	20 cent
	# of partials	8
	# of harmonic models	200
	standard deviation of Gaussian	3.0
	$\bar{r}_n$	$0.6547 \times n^{-2}$
	$\bar{d}$	3.0

in table 3.2). Time series of power spectrum was analyzed using Gabor wavelet transform with frame shift of 16ms for input digital signals of 16kHz sampling rate. The lower bound of the frequency range and the frequency resolution were 60Hz and 12cent, respectively. The experimental conditions are shown in detail in table 3.1.

The purpose of this experiment is to clarify the effect of using BIC, and the multipitch estimation accuracy of the Bayesian Harmonic Clustering. As the first task, we compare the performance of the source number estimation method using BIC with that of a simple intensity thresholding for F_0 candidate truncation. As the second task, we chose as a comparison *'PreFEst-core'[45]. Since PreFEst-core is actually designed to be an extraction of the most

*Note that we have only implemented the 'PreFEst-core', *i.e.*, a framewise pitch likelihood estimation, for the evaluation and not included the 'PreFEst-back-end', *i.e.*, multi-agent based pitch tracking algorithm.

dominant F_0 trajectory from multipitch signals and does not include a specific procedure for source number determination, we decided to include the same intensity thresholding for decision making under the same condition to make a proper comparison. The specific way of intensity thresholding we have implemented is to regard the harmonic kernels, or the tone models as referred in [45], as silence, whose integral, *i.e.*, $w_k \sum_{\forall n} r_{k,n}$ is smaller than a particular value.

Let us refer to these three types of methods as following:

- *proposed A*: Bayesian Harmonic Clustering and minimum BIC model selection for source number estimation,
- *proposed B*: Bayesian Harmonic Clustering and intensity thresholding for F_0 candidate truncation.
- *conventional*: PreFEst-core[45] and intensity thresholding for F_0 candidate truncation.

We expect that the effectiveness of the source number estimation method using BIC can be confirmed through comparison between *proposed A* & *B*, and as well the effectiveness of the Bayesian Harmonic Clustering-based multipitch estimation estimation through comparison between *proposed B* & *conventional*.

3.6.2 Results

A typical example of the F_0 estimates obtained by *proposed A* together with the corresponding handcrafted reference MIDI data is demonstrated in Fig. 4.6.

The average accuracy rates over all experimental data of the *proposed A* and the rest of the two methods with different thresholds are shown in Fig. 3.7. One sees from the result that *proposed A* obviously outperforms *proposed B*, and as well *proposed B* significantly outperforms *conventional*. Therefore, both elements in our proposed method, *i.e.*, applying information criterion to source number estimation and Bayesian Harmonic Clustering-based multipitch estimation was proved to be effective.

For more detail, see table 3.4 showing accuracy rate for each experimental data. From the results of *proposed B* and *conventional*, the proper threshold that gives the best accuracy rate, tends to depend highly on test data, obviously because if the relative power level differs among several data, a proper threshold for a particular data is not always proper also for

Table 3.2 *List of the experimental data excerpted from the RWC music database [44]*

Symbol	Title (Genre)	Composer/Player	Instruments	# of frames
data(1)	Crescent Serenade (Jazz)	S. Yamamoto	Guitar	4427
data(2)	For Two (Jazz)	H. Chubachi	Guitar	6555
data(3)	Jive (Jazz)	M. Nakamura	Piano	5179
data(4)	Lounge Away (Jazz)	S. Yamamoto	Guitar	9583
data(5)	For Two (Jazz)	M. Nakamura	Piano	9091
data(6)	Jive (Jazz)	H. Chubachi	Guitar	3690
data(7)	Three Gimnopedies no. 1 (Classic)	E. Satie	Piano	6571
data(8)	Nocturne no.2, op.9-2(Classic)	F. F. Chopin	Piano	7258

others. When considering a practical use, it is, however, inconvenient to tune thresholds carefully every time we test on different data. It should be emphasized that the *proposed A* works reliably even without such exhausting tuning.

3.7 Summary of Chapter 3

In this chapter, we have proposed the principle of Harmonic Clustering estimating the optimal spectral masking functions clustering source by source the power spectrum of the mixed sound signal of interest. We have shown that Harmonic Clustering can be understood as the minimization of the distortion between the power spectrum of the mixed sound and a mixture of spectral cluster models, or as the optimal decomposition (clustering) of the power spectrum using spectral cluster models, and we presented the formulation of the problem in these two points of view. Moreover, starting from the fact that the minimization of the distortion measure can be understood as a maximum likelihood problem with the continuous Poisson distribution as likelihood function, we showed that, by introducing prior distributions, optimal estimation under empirical constraints can performed through Maximum *A Posteriori* estimation. Furthermore, we showed that a criterion for source number selection could simultaneously be obtained through model selection criterion. Experimental evaluations proved the effectiveness of the two elemental techniques introduced in this

Table 3.3 *Results obtained by PreFEst-core [45]. Columns (A)∼(J) and (K)∼(R) show the accuracies with different thresholds: (A)*2.0×10^8*, (B)*2.5×10^8*, (C)*5.0×10^8*, (D)*7.5×10^8*, (E)*10×10^8*, (F)*$15\times^8$*, (G)*17.5×10^8*, (H)*20×10^8*, (I)*25×10^8*, (J)*27.5×10^8*, (K)*7.5×10^9*, (L)*1.0×10^{10}*, (M)*2.0×10^{10}*, (N)*3.0×10^{10}*, (O)*4.0×10^{10}*, (P)*5.0×10^{10}*, (Q)*6.0×10^{10}*, (R)*7.0×10^{10}.

	Accuracy(%)									
	conventional [45]									
	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)
data(1)	56.6	62.49	75.9	81.6	83.3	**84.6**	83.0	81.5	78.4	75.8
data(2)	68.7	**69.6**	66.3	59.0	53.7	36.3	32.4	30.3	26.8	26.5
data(3)	-20.8	-7.3	31.7	47.8	56.9	65.1	69.5	71.9	**75.5**	71.8
data(4)	55.1	56.8	60.7	63.3	63.1	63.6	**64.1**	62.3	60.6	60.2
data(5)	50.7	53.2	**61.0**	60.0	58.8	59.3	57.6	58.0	57.5	49.7
data(6)	-7.2	6.6	37.9	51.1	57.7	65.9	65.6	**66.7**	66.3	65.7
data(7)	51.6	54.1	**62.7**	52.4	47.0	45.9	42.7	41.1	42.2	42.7
data(8)	20.8	22.9	36.6	**42.5**	38.5	39.1	38.8	37.7	32.7	30.6
Average	39.1	43.3	55.2	57.1	56.5	55.9	55.0	54.4	53.0	50.7

chapter, multipitch estimation and automatic source number estimation based on Harmonic Clustering.

We discussed here multipitch estimation for short-time frames of mixed signals. In the next chapter, assuming the continuity in the time direction of the F_0 and of the power, we extend this method to a global spectral structure estimation method on the whole time-frequency domain.

Table 3.4 *Results obtained by the proposed method (proposed A and proposed B). Columns (A)∼(J) and (K)∼(R) show the accuracies with different thresholds: (A)*2.0×10^8*, (B)*2.5×10^8*, (C)*5.0×10^8*, (D)*7.5×10^8*, (E)*10×10^8*, (F)*$15\times^8$*, (G)*17.5×10^8*, (H)*20×10^8*, (I)*25×10^8*, (J)*27.5×10^8*, (K)*7.5×10^9*, (L)*1.0×10^{10}*, (M)*2.0×10^{10}*, (N)*3.0×10^{10}*, (O)*4.0×10^{10}*, (P)*5.0×10^{10}*, (Q)*6.0×10^{10}*, (R)*7.0×10^{10}.

	Accuracy(%)								
	proposed B								*proposed A*
	(K)	(L)	(M)	(N)	(O)	(P)	(Q)	(R)	
data(1)	42.4	72.1	76.8	79.4	85.9	**87.2**	86.8	82.7	**76.3**
data(2)	76.4	85.3	86.3	**86.4**	69.7	65.6	59.9	57.9	**84.8**
data(3)	37.3	52.4	57.5	61.0	69.0	70.2	70.5	**71.6**	**72.6**
data(4)	64.5	66.3	66.5	67.0	69.0	**69.7**	69.1	67.8	**76.7**
data(5)	62.6	65.3	66.3	**66.9**	66.8	64.1	63.3	62.7	**72.1**
data(6)	27.1	54.4	61.8	66.3	76.7	78.6	80.8	**82.0**	**57.4**
data(7)	64.5	74.4	77.7	**79.2**	76.6	75.1	70.9	69.9	**76.5**
data(8)	63.7	76.5	78.2	**78.7**	74.9	66.4	56.6	50.6	**75.5**
Average	58.4	69.2	71.6	**73.0**	72.4	70.6	67.9	66.2	**74.9**

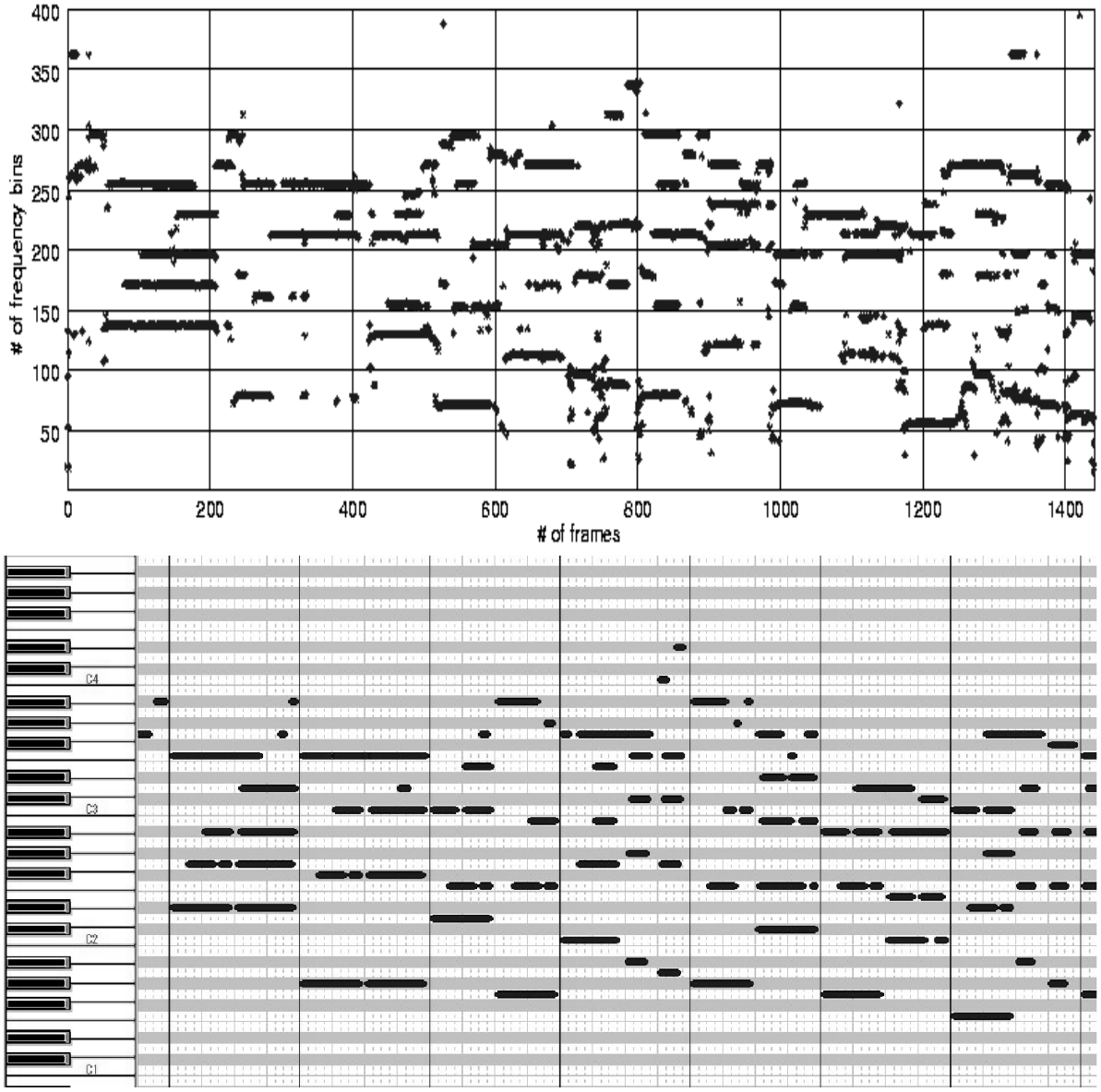


Figure 3.6 *A multipitch estimation result(top) by the proposed method and the hand-labeled MIDI reference data displayed in piano-roll form (bottom).*

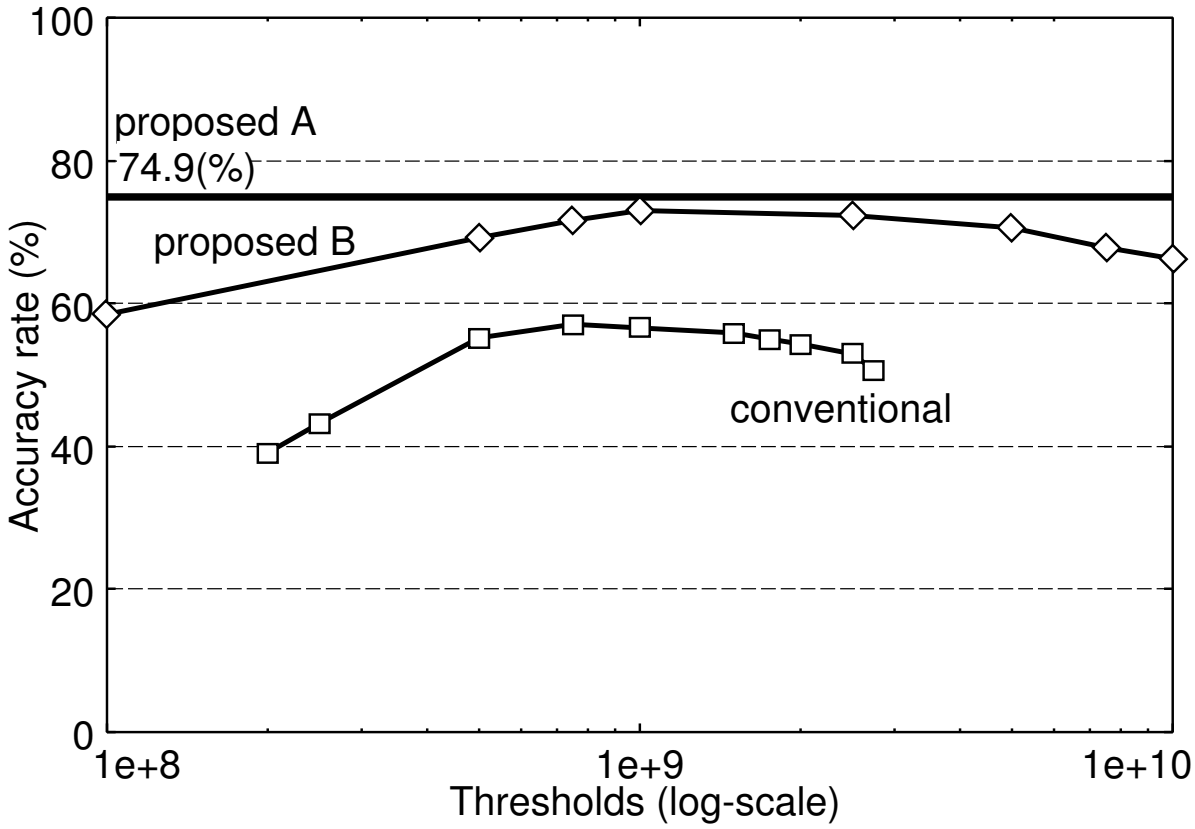


Figure 3.7 *Average accuracy rates over all test data of 'proposed A' (Bayesian Harmonic Clustering multipitch estimation & minimum BIC model selection), 'proposed B' (Bayesian Harmonic Clustering multipitch estimation & thresholding) and 'conventional' with different thresholds.*

Chapter 4

Harmonic-Temporal Clustering

4.1 Introduction

Human listeners are able to concentrate on listening to a target sound without difficulty even in the situation where many speakers are talking at the same time. This fact has persuaded many scientists that the auditory system of human has a significant ability to recognize the external environment actively. This nature is referred to as the "auditory scene analysis (ASA)" and has been attracting interest since Bregman's book was published [16]. In [16], Bregman has shown through experiments the psychological evidences concerning the ability of the auditory system, such that:

1. Acoustic signal is "segregated" into spectrogram-like pieces, which is called the "auditory elements".

2. Auditory elements that originate from the same source are likely to be "grouped" to form the "auditory stream".

3. The grouping cues are said to be related to:

   (a) harmonicity,

   (b) common onset and offset,

   (c) coherent amplitude and frequency modulation,

   (d) continuity of amplitude and frequency,

   (e) proximity of time-frequency components,

   (f) common spatial location.

Recent efforts are being directed toward the attempt to implement this ability of the auditory system. Such a framework is called the "Computational Auditory Scene Analysis (CASA)".

The main focus of today's CASA research is to develop a source separation method based upon the grouping cues suggested by Bregman. More specifically, the main purpose is to extract useful features (for example, F_0) or to restore the target signal of interest by performing the segregation process and grouping process through a computational algorithm.

Cooke [29], Brown *et al.* [18], Ellis [37], Fishbach [40], Nakatani *et al.* [75] developed source separation methods utilizing the grouping cues. As most of these methods use artificial-intelligence-based or rule-based approaches, they enable the introduction of various constraints in a top-down manner, but the algorithms tend to have many thresholding steps, that often make systems too complicated to handle. Nishi *et al.* [76], Unoki *et al.* [106], Abe *et al.* [3, 4], Wu *et al.* [115] tried to formulate the CASA problem as an optimization problem using the grouping cues as mathematically formalized constraints. Kashino *et al.* [60] presented a CASA algorithm designed specifically for an automatic transcription use. Goto's PreFEst [45] is in some sense a CASA method.

In most of these conventional methods, they usually implement the grouping process in the following way: first extract instantaneous feature at each discrete time point and then estimate the whole tracks of those features by exploiting hidden Markov model (HMM), multiple agents, or some dynamical system such as Kalman filtering. The first half of this procedure is for finding the set of frequency components that seem to originate from the same source using only the "harmonicity" constraint. This step corresponds to the grouping process in the frequency direction. The second half, on the other hand, is for interpolating over incorrect values of the features possibly taken at the previous step using the rest of the cues. This step corresponds to the grouping process in the time direction.

From the engineering point of view, however, one cannot necessarily conclude that this is the optimal way of performing the grouping process. It is quite obvious that the more accurate the grouping process in the frequency direction, the more reliable the result of that in the time direction. On the other hand, we hope to know, if possible, the features at preceding and succeeding time points to estimate a high precision result of the feature extraction at the current time assuming they change gradually over time. Therefore, these two processes should be done essentially in a cooperative way and not independently in succession for even more reliable results. This belief has led us to formulate a unified estimation framework for the two dimensional structure of time-frequency power spectra, in contrast to

the conventional strategy. We will call this method "Harmonic-Temporal Clustering".

4.2 Abstract and Organization of Chapter 4

we aim at developing a computational algorithm enabling the decomposition of the time-frequency components of the signal of interest into distinct clusters such that each of them is associated with a single auditory stream. To do so, we directly model a spectro-temporal model whose shape can be taken freely within the Bregman's constraint, and then try to fit the mixture of this model to the observed spectrogram as well as possible.

As constant Q filterbank is known to be a good model for the auditory periphery system, we will first derive in Subsection 4.3.1 the constant Q filterbank output of a pseudoperiodic signal model and then give a specific form for the spectro-temporal structure that is associated with the auditory stream in the succeeding subsections. In Subsection 4.4, we present the optimization algorithm, that performs segregation of the observed spectrogram and the parameter estimation of the auditory stream model at the same time.

4.3 Spectro-Temporal Cluster Model

4.3.1 Constant Q Filterbank Output of Pseudoperiodic Signal

Consider as the k^{th} source signal model the analytic signal representation of a pseudo-periodic signal given by

$$f_k(u) = \sum_{n=1}^{N} \widetilde{w}_{k,n}(u) e^{j\left(n\theta_k(u)+\varphi_{k,n}\right)}, \quad u \in (-\infty, \infty), \tag{4.1}$$

where u is the time, $n\theta_k(u) + \varphi_{k,n}$ is the instantaneous phase of the n^{th} partial and $\widetilde{w}_{k,n}(u)$ the instantaneous amplitude. This signal model implies that it satisfies the 'harmonicity', out of the Bregman's grouping cues. We will first derive its constant Q filterbank output. Let us define the wavelet basis function by

$$\psi_{\alpha,t}(u) \triangleq \frac{1}{\sqrt{2\pi}\alpha}\psi\left(\frac{u-t}{\alpha}\right), \tag{4.2}$$

where α is the scale parameter such that $\alpha > 0$, t the shift parameter and $\psi(u)$ an arbitrary analyzing wavelet that has the center frequency of 1 and satisfies the admissible condition.

$\psi_{\alpha,t}(u)$ is used to measure the component of period α at time t. The continuous wavelet transform of $f_k(u)$ is then defined by

$$W_k\big(\log \tfrac{1}{a}, b\big) \triangleq \Big\langle f_k(u), \psi_{\alpha,t}(u) \Big\rangle_{u\in\mathbb{R}} \tag{4.3}$$

$$= \int_{-\infty}^{\infty} \sum_{n=1}^{N} \widetilde{w}_{k,n}(u) e^{j\left(n\theta_k(u)+\varphi_{k,n}\right)} \psi^*_{\alpha,t}(u) \mathrm{d}u. \tag{4.4}$$

As the dominant part of $\psi^*_{\alpha,t}(u)$ is generally localized only around time t, the result of the integral in Eq. (4.4) depends heavily on the portion of $\theta_k(t)$ and $\widetilde{w}_{k,n}(t)$ near t. Taking into account that the instantaneous phase $\theta_k(t)$ and the instantaneous amplitude $\widetilde{w}_{k,n}(t)$ of the signal of interest often change gradually over time, approximating $\theta_k(t)$ and $\widetilde{w}_{k,n}(t)$ by zero and first order Taylor series expansions around time t:

$$\widetilde{w}_{k,n}(u) = \widetilde{w}_{k,n}(t) + \left.\frac{\mathrm{d}\widetilde{w}_{k,n}(u)}{\mathrm{d}u}\right|_{u=t} \big(u-t\big) + \frac{1}{2} \left.\frac{\mathrm{d}^2\widetilde{w}_{k,n}(u)}{\mathrm{d}u^2}\right|_{u=t} \big(u-t\big)^2 + \cdots$$

$$\approx \widetilde{w}_{k,n}(t), \tag{4.5}$$

$$\theta_k(u) = \theta_k(t) + \left.\frac{\mathrm{d}\theta_k(u)}{\mathrm{d}u}\right|_{u=t} \big(u-t\big) + \frac{1}{2} \left.\frac{\mathrm{d}^2\theta_k(u)}{\mathrm{d}u^2}\right|_{u=t} \big(u-t\big)^2 + \cdots$$

$$\approx \theta_k(t) + \theta'_k(t)\big(u-t\big), \tag{4.6}$$

may not affect significantly the result of Eq. (4.4). As the instantaneous frequency is defined as the first order derivative of the instantaneous phase, $\theta'_k(u)$ is the instantaneous F_0 frequency (a F_0 trajectory function) of the k^{th} source, which we will henceforth denote by $\mu_k(u)$. From these approximations, Eq. (4.5) and Eq. (4.6), Eq. (4.4) can be written as

$$W_k\big(\log \tfrac{1}{\alpha}, t\big) \approx \sum_{n=1}^{N} \widetilde{w}_{k,n}(t) e^{j\left(n\theta_k(t)+\varphi_{k,n}\right)} \int_{-\infty}^{\infty} e^{jn\mu_k(t)(u-t)} \psi^*_{\alpha,t}(u) \mathrm{d}u. \tag{4.7}$$

Using the Parseval's theorem, the integral part is given explicitly as

$$\int_{-\infty}^{\infty} e^{jn\mu_k(t)(u-t)} \psi^*_{\alpha,t}(u) \mathrm{d}u = \left\langle e^{jn\mu_k(t)(u-t)}, \frac{1}{\sqrt{2\pi}\alpha} \psi\left(\frac{u-t}{\alpha}\right) \right\rangle_{u\in\mathbb{R}(-\infty,\infty)} \tag{4.8}$$

$$= \left\langle e^{jn\mu_k(t)u}, \frac{1}{\sqrt{2\pi}\alpha} \psi\left(\frac{t}{\alpha}\right) \right\rangle_{u\in\mathbb{R}(-\infty,\infty)} \tag{4.9}$$

$$= \left\langle \sqrt{2\pi}\delta\big(\omega - n\mu_k(t)\big), \frac{1}{\sqrt{2\pi}} \Psi(\alpha\omega) \right\rangle_{\omega\in\mathbb{R}(-\infty,\infty)} \tag{4.10}$$

$$= \Psi^*\big(\alpha n \mu_k(t)\big), \tag{4.11}$$

which yields

$$W_k\big(\log \tfrac{1}{\alpha}, t\big) \approx \sum_{n=1}^{N} \widetilde{w}_{k,n}(t) \Psi^*\big(an\mu_k(t)\big) e^{j\left(n\theta_k(t)+\varphi_{k,n}\right)}. \tag{4.12}$$

By changing the variable $x = \log \frac{1}{\alpha}$ and by putting $\Omega_k(t) \triangleq \log \mu_k(t)$, W_k can be expressed in the time-logfrequency domain:

$$W_k(x,t) = \sum_{n=1}^{N} \widetilde{w}_{k,n}(t) \Psi^* \Big(n e^{-x+\Omega_k(t)} \Big) e^{j\left(n\theta_k(t)+\varphi_{k,n}\right)}. \tag{4.13}$$

As the frequency characteristic $\Psi(\omega)$ of the analyzing wavelet can be chosen arbitrarily, we use here again the following unimodal real function whose maximum is taken at $\omega = 1$ (see Fig. 3.1):

$$\Psi(\omega) = \Psi^*(\omega) = \begin{cases} \exp\left(-\frac{\left(\log \omega\right)^2}{4\sigma^2} \right) & (\omega > 0) \\ 0 & (\omega \leqq 0) \end{cases}. \tag{4.14}$$

Eq. (4.13) is then given as

$$W_k(x,t) = \sum_{n=1}^{N} \widetilde{w}_{k,n}(t) \exp\left(-\frac{\left(x - \Omega_k(t) - \log n\right)^2}{4\sigma^2} \right) e^{j\left(n\theta_k(t)+\varphi_{k,n}\right)}, \tag{4.15}$$

and the resulting power spectrum of Eq. (4.15) can be written as

$$\begin{aligned} \left\|W_k(x,t)\right\|^2 = \sum_{n=1}^{N} \left\| \widetilde{w}_{k,n}(t) \exp\left(-\frac{\left(x - \Omega_k(t) - \log n\right)^2}{4\sigma^2} \right) e^{j\left(n\theta_k(t)+\varphi_{k,n}\right)} \right\|^2 \\ + \sum_{n \neq n'} \widetilde{w}_{k,n}(t) \widetilde{w}_{k,n'}(t) \exp\left(-\frac{\left(x - \Omega_k(t) - \log n\right)^2}{4\sigma^2} \right) \\ \exp\left(-\frac{\left(x - \Omega_k(t) - \log n'\right)^2}{4\sigma^2} \right) e^{j(n\theta_k(t)+n'\theta_k(t)+\varphi_{k,n}+\varphi_{k,n'})}. \end{aligned} \tag{4.16}$$

If we now assume that the time-frequency components are sparsely distributed so that the partials rarely overlap, the second term could be negligibly smaller than the first term in the above equation. This assumption justifies the additivity of power spectra and the power spectrum of the k^{th} source signal model is then expressed as a Gaussian mixture model whose maxima are centered over prospective harmonics $x = \Omega_k(t) + \log n$. Putting $w_{k,n}(t) \triangleq \sqrt{2\pi}\sigma \|\widetilde{w}_{k,n}(t)\|^2$ (instantaneous power), one obtains

$$\left\|W_k(x,t)\right\|^2 \approx \sum_{n=1}^{N} \frac{w_{k,n}(t)}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(x - \Omega_k(t) - \log n\right)^2}{2\sigma^2} \right). \tag{4.17}$$

The graphical representation of its cutting plane at time t can be seen in Fig. 4.1.

4.3.2 Nonparametric and Parametric Modeling

There may be two possible ways to enforce 'continuity' constraints on the temporal trajectories of the instantaneous power of each partial and the instantaneous F_0 frequency. One

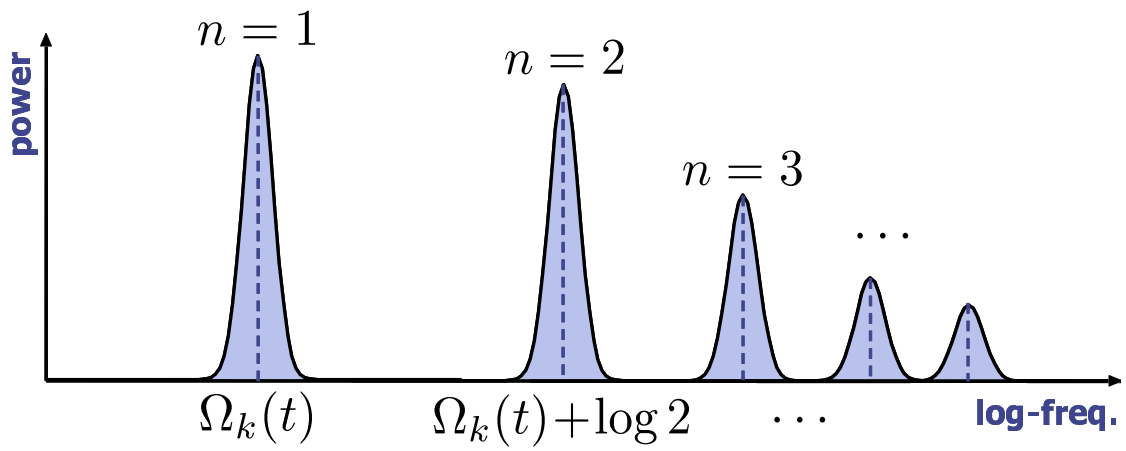


Figure 4.1 *Graphical representation of Eq. (4.17).*

is to adopt particular classes of parametric function for $w_{k,n}(t)$ and $\Omega_k(t)$. In this case, the smoothness of the functions can often be controlled by the degree-of-freedom of the models we choose to apply. Second is to consider both $w_{k,n}(t)$ and $\Omega_k(t)$ as nonparametric functions and to try to estimate them directly. In this case, the smoothness of the functions can be controlled by a gradient penalizing term added to the cost function. This kind of penalizer is often called a 'regularization term' in the image processing area. In the Bayesian point of view, essentially the same role is played by the prior distribution. Details will be presented in Subsection 4.4.1. In order to distinguish between these two ways of modeling $w_{k,n}(t)$ and $\Omega_k(t)$, we will call the spectro-temporal source model in the former way the "parametric spectro-temporal model", and that in the latter way the "nonparametric spectro-temporal model".

Formulation of the parameter estimation algorithm depends on the choice of the parametric model or the nonparametric model. After we show a thinkable class of parametric function for $w_{k,n}(t)$ and $\Omega_k(t)$ in Subsection 4.3.3, we formulate the optimal clustering algorithms, "nonparametric HTC" in Subsection 4.4.1 and "parametric HTC" in Subsection 4.4.2.

4.3.3 Parametric Spectro-Temporal Cluster Model

Assuming the 'common onset' and the 'common amplitude' of the partial components for the source model, the instantaneous power should be of a variable separable form of the partial index n and the time t:

$$w_{k,n}(t) = \widetilde{v}_{k,n} u_k(t). \tag{4.18}$$

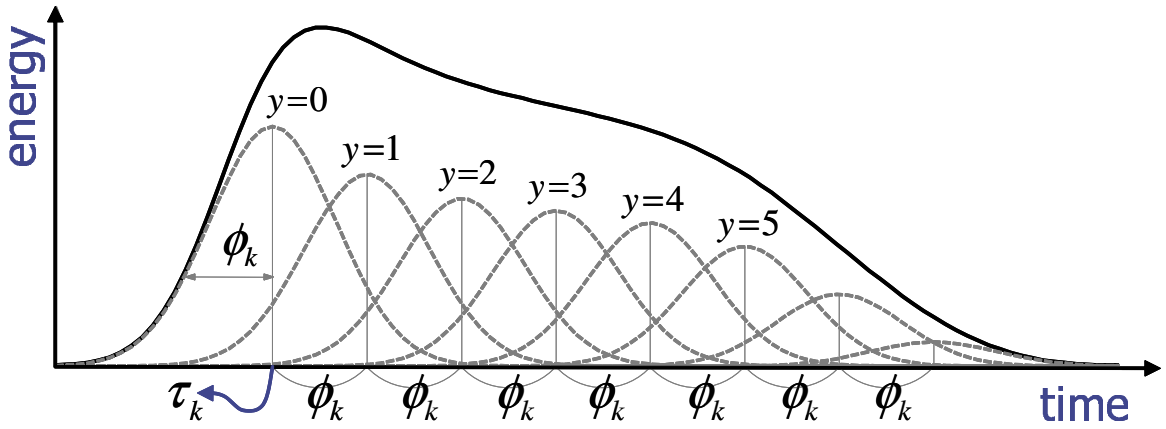


Figure 4.2 *Graphical representation of Eq. (4.21).*

Letting $u_k(t)$ satisfy

$$\forall k, \quad \int_{-\infty}^{\infty} u_k(t)\mathrm{d}t = 1, \tag{4.19}$$

then the parameter $\widetilde{v}_{k,n}$ corresponds to the total energy of the n^{th} partial of the k^{th} source such that $\widetilde{v}_{k,n} = \int_{-\infty}^{\infty} w_{k,n}(t)\mathrm{d}t$. Let further be $\widetilde{v}_{k,n} \triangleq w_k v_{k,n}$, hence

$$w_{k,n}(t) = w_k v_{k,n} u_k(t), \tag{4.20}$$

and let $v_{k,n}$ satisfy $\sum_n v_{k,n} = 1$ for convenience. The normalized common power envelope $u_k(t)$ should be a smooth and non-negative function that has a time spread from minus to plus infinity, which can be modeled by a following type of constrained Gaussian mixture model (see Fig. 4.2):

$$u_k(t) = \sum_{y=0}^{Y-1} \frac{u_{k,y}}{\sqrt{2\pi}\phi_k} \exp\left(-\frac{(t-\tau_k-\kappa y\phi_k)^2}{2{\phi_k}^2}\right). \tag{4.21}$$

τ_k is the center of the forefront Gaussian, that could be considered as an onset time estimate, $u_{k,y}$ the weight parameter for each kernel, that allows the function to have variable shapes. To satisfy Eq. (4.19), $u_{k,y}$ must only be normalized to unity:

$$\forall k, \quad \sum_y u_{k,y} = 1. \tag{4.22}$$

The particularity of this function is that the centers of the Gaussian function kernels are spaced by a distance proportional to the common diffusion parameter ϕ_k with a proportionality coefficient κ, which we henceforth set to 1 (see Fig. Fig. 4.2). This tying ensures the smoothness of the curve by preventing adjacent kernels from being separated from each

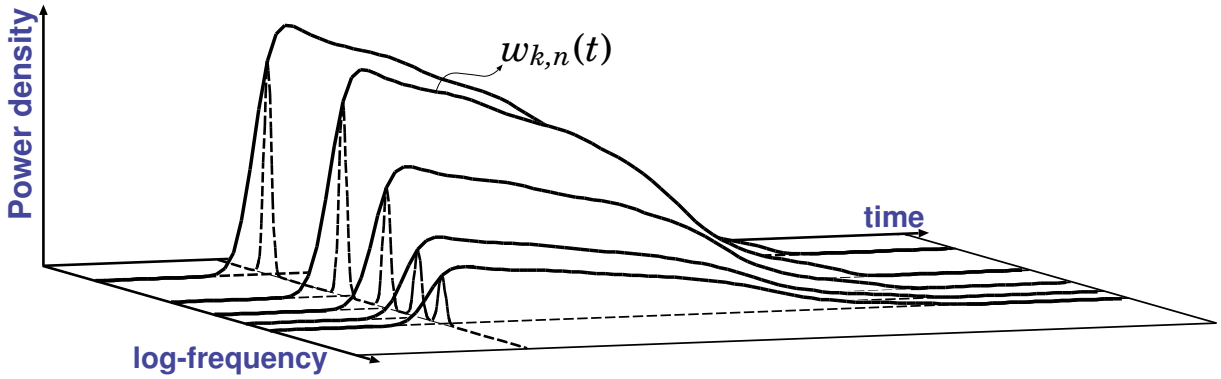


Figure 4.3 *The spectro-temporal model associated with an audio stream.*

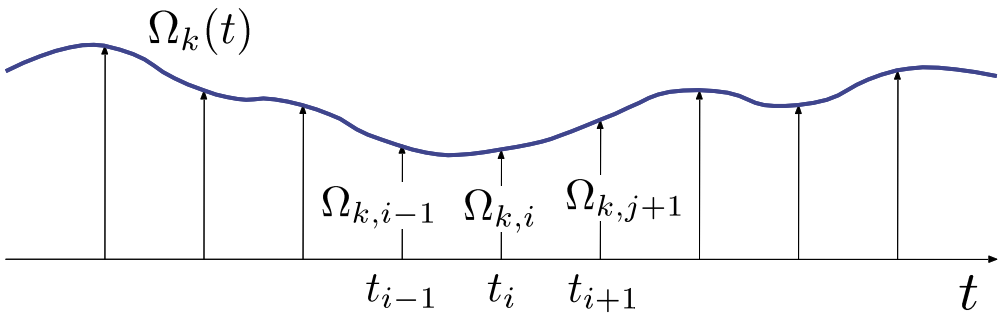


Figure 4.4 *Cubic spline F_0 track function (Eq. (4.25))*

other. ϕ_k also works as a parameter to make a linear stretch of $u_k(t)$ in the time direction allowing to express various durations of sources. Moreover, by forbidding switches in the position of the kernels, it reduces the singularity of the system, improving the optimization perspectives. Substituting Eq. (4.20) and Eq. (4.21) into Eq. (4.17), one obtains

$$\sum_{n=1}^{N}\sum_{y=0}^{Y-1}\frac{w_k v_{k,n} u_{k,y}}{2\pi\sigma\phi_k} e^{-\frac{(x-\Omega_k(t)-\log n)^2}{2\sigma^2}-\frac{(t-\tau_k-y\phi_k)^2}{2\phi_k^2}}. \tag{4.23}$$

Its graphical representation can be seen in Fig. 4.23.

We choose two types of models for the F_0 trajectory function $\Omega_k(t)$, a polynomial of time t:

$$\Omega_k(t) \triangleq \Omega_{k,0} + \Omega_{k,1}t + \Omega_{k,2}t^2 + \Omega_{k,3}t^3 + \cdots, \tag{4.24}$$

and a cubic spline function (see Fig. 4.4):

$$\Omega_k(t) \triangleq \frac{1}{t_{i+1}-t_i}\Big(\Omega_{k,i}(t_{i+1}-t) + \Omega_{k,i+1}(t-t_i) - \frac{1}{6}(t-t_i)(t_{i+1}-t)\big[(t_{i+2}-t)\Omega''_{k,i} + (t-t_{i-1})\Omega''_{k,i+1}\big]\Big), \quad t\in[t_i,t_{i+1}). \tag{4.25}$$

In the cubic spline F_0 contour function, the analysis interval is divided into subintervals $[t_i, t_{i+1})$ which are assumed of equal length. The parameters of the spline contour model are then the values $\Omega_{k,i}$ of the F_0 at each bounding point t_i. The values $\Omega''_{k,i}$ of the second derivative at those points are given by the expression $\boldsymbol{\Omega}'' = \boldsymbol{M}\boldsymbol{\Omega}$ for a certain matrix $\boldsymbol{M}$ which can be explicitly computed offline if we consider $t_1, \cdots, t_I$ are constant parameters, under the hypothesis that the first-order derivative is 0 at the bounds of the analysis interval.

If we are able to estimate $\{\Omega_{k,i}\}_{i=1}^{I}, \{v_{k,n}\}_{n=1}^{N}, \{u_{k,y}\}_{y=0}^{Y-1}, w_k, \tau_k, \phi_k$, then the k^{th} source signal can be reconstructed by Eq. (4.1) whose starting phase is chosen arbitrarily. The parameter estimation algorithm will be discussed in Subsection 4.4.

4.4 Optimal Clustering Algorithm

We consider here the problem of decomposing the observed time-frequency spectrum into distinct clusters that correspond to the auditory stream. Two ways of solution to this problem is presented in the following subsections.

4.4.1 Nonparametric HTC

We will consider here the nonparametric case. Let $\boldsymbol{\Theta}$ refers to $\{\Omega_k(t), \{w_{k,n}(t)\}_{n=1}^{N}\}_{k=1}^{K}$.

We define by $\|Y(x,t)\|^2$ the time-logfrequency power spectrum of the signal of interest obtained by the constant Q analysis. Let us introduce a masking function $m_{k,n}(x,t)$ that extracts the spectro-temporal components associated with the n^{th} partial of the k^{th} source from $\|Y(x,t)\|^2$. For $(x,t) \in \mathbb{R}^2$, $m_{k,n}(x,t)$ indicates the percentage of the portion of $\|Y(x,t)\|^2$ shared to the n^{th} partial of the k^{th} source, such that satisfies

$$\sum_{k=1}^{K}\sum_{n=1}^{N} m_{k,n}(x,t) = 1 \tag{4.26}$$

$$0 < m_{k,n}(x,t) < 1, \quad k \in \{1, \cdots, K\},\ n \in \{1, \cdots, N\}. \tag{4.27}$$

A portion of the observed power spectrum is thus given arbitrarily by

$$m_{k,n}(x,t)\big\|Y(x,t)\big\|^2, \quad (x,t) \in \mathbb{R}^2, \tag{4.28}$$

which we call a "spectral cluster". As we expect the spectral cluster to be associated with the auditory stream, we need to introduce a measure function that specifies how well the

spectral cluster fits all the Bregman's grouping cues. One possible measure function may be the I divergence between $m_{k,n}(x,t)\|Y(x,t)\|^2$ and the spectro-temporal model we derived in Section 3.2:

$$\mathcal{W}_{k,n}(x) \triangleq \frac{w_{k,n}(t)}{\sqrt{2\pi}\sigma} \exp\left(- \frac{\big(x - \Omega_k(t) - \log n\big)^2}{2\sigma^2} \right), \tag{4.29}$$

which is written as

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Bigg(m_{k,n}(x,t)\big\|Y(x,t)\big\|^2 \log \frac{m_{k,n}(x,t)\big\|Y(x,t)\big\|^2}{\mathcal{W}_{k,n}(x,t)} - \Big(m_{k,n}(x,t)\big\|Y(x,t)\big\|^2 - \mathcal{W}_{k,n}(x,t) \Big) \Bigg) \mathrm{d}x\mathrm{d}t. \tag{4.30}$$

The optimal clustering can thus be achieved by minimizing their sum:

$$\Phi(\boldsymbol{\Theta}, m) = \sum_{k=1}^{K}\sum_{n=1}^{N} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Bigg(m_{k,n}(x,t)\big\|Y(x,t)\big\|^2 \log \frac{m_{k,n}(x,t)\big\|Y(x,t)\big\|^2}{\mathcal{W}_{k,n}(x,t)} - \Big(m_{k,n}(x,t)\big\|Y(x,t)\big\|^2 - \mathcal{W}_{k,n}(x,t) \Big) \Bigg) \mathrm{d}x\mathrm{d}t \tag{4.31}$$

with respect to $\boldsymbol{\Theta}$ and $m_{k,n}(x,t)$. To do so, we shall find it most convenient to minimize this objective function recursively with respect to $m_{k,n}(x,t)$ and $\boldsymbol{\Theta}$ while keeping the other fixed. The minimization with respect to $m_{k,n}(x,t)$ decomposes the observed power spectrum using the auditory stream models estimated hypothetically at the previous step and the minimization with respect to $\boldsymbol{\Theta}$, on the other hand, updates the auditory stream models to a more convincing one using these separate clusters. Both steps necessarily decreases the objective function, which is bounded below, and the convergence of this recursive algorithm is thus guaranteed.

The update equation for the spectral masking function $m_{k,n}(x,t)$ that minimizes $\Phi(\boldsymbol{\Theta}, m)$ when $\boldsymbol{\Theta}$ is fixed is obtained analytically as

$$\widehat{m}_{k,n}(x,t) = \frac{\mathcal{W}_{k,n}(x,t)}{\sum_{k}\sum_{n} \mathcal{W}_{k,n}(x,t)}. \tag{4.32}$$

Substituting this result into Eq. (4.31), we obtain

$$\Phi(\boldsymbol{\Theta}, \widehat{m}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Bigg(\big\|Y(x,t)\big\|^2 \log \frac{\big\|Y(x,t)\big\|^2}{\sum_{k}\sum_{n} \mathcal{W}_{k,n}(x,t)} - \Big(\big\|Y(x,t)\big\|^2 - \sum_{k}\sum_{n} \mathcal{W}_{k,n}(x,t) \Big) \Bigg) \mathrm{d}x\mathrm{d}t, \tag{4.33}$$

from which we realize that what we are trying to minimize w.r.t $\mathbf{\Theta}$ is the I divergence between the whole observed power spectrum and the mixture of all the spectro-temporal source models. From the statistical point of view, this minimization is understood as a maximum likelihood (regression analysis) where its log-likelihood is given explicitly by

$$\begin{aligned} L(\mathbf{\Theta}) &\triangleq \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \log P\big(\|Y(x)\|^2\big|\mathbf{\Theta}\big)\mathrm{d}x\mathrm{d}t \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\Big(\big\|Y(x,t)\big\|^2 \log\big\|W(x,t)\big\|^2 - \big\|W(x,t)\big\|^2 - \log\Gamma\Big(\big\|Y(x,t)\big\|^2+1\Big)\Big)\mathrm{d}x\mathrm{d}t. \end{aligned} \tag{4.34}$$

See Chapter 3 for more detailed discussion.

Now we shall derive the update equation for $\mathbf{\Theta}$ that minimizes $\Phi(\mathbf{\Theta}, m)$ when $m_{k,n}(x,t)$ is fixed. The optimal $\Omega_k(t)$ and $w_{k,n}(t)$ that minimizes the functional $\Phi(\mathbf{\Theta}, m)$ can be obtained by the variational method. The variation of $\Phi(\mathbf{\Theta}, m)$ with respect to $\Omega_k(t)$ and $w_{k,n}(t)$ given as

$$\delta\Phi(\mathbf{\Theta}, m) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\left(\frac{\partial\Phi(\mathbf{\Theta}, m)}{\partial\Omega_k}\right)\delta\Omega_k + \left(\frac{\partial\Phi(\mathbf{\Theta}, m)}{\partial w_{k,n}}\right)\delta w_{k,n}\right)\mathrm{d}x\mathrm{d}t, \tag{4.35}$$

is identically 0 if $\frac{\partial\Phi(\mathbf{\Theta},m)}{\partial\Omega_k} = 0$ and $\frac{\partial\Phi(\mathbf{\Theta},m)}{\partial w_{k,n}} = 0$. Hence, setting

$$\frac{\partial\Phi(\mathbf{\Theta}, m)}{\partial\Omega_k} = \sum_{n=1}^{N}\int_{-\infty}^{\infty} m_{k,n}(x,t)\big\|Y(x,t)\big\|^2\frac{-\big(x-\Omega_k(t)-\log n\big)}{\sigma^2}\mathrm{d}x, \tag{4.36}$$

to 0, one obtains

$$\Omega_k(t) = \frac{\displaystyle\sum_{n=1}^{N}\int_{-\infty}^{\infty} m_{k,n}(x,t)\big\|Y(x,t)\big\|^2\big(x-\log n\big)\mathrm{d}x}{\displaystyle\sum_{n=1}^{N}\int_{-\infty}^{\infty} m_{k,n}(x,t)\big\|Y(x,t)\big\|^2\mathrm{d}x}. \tag{4.37}$$

Eq. (4.37) implies a frame-by-frame F_0 parameter update. Similarly, setting

$$\frac{\partial\Phi(\mathbf{\Theta}, m)}{\partial w_{k,n}} = 1 - \int_{-\infty}^{\infty} m_{k,n}(x,t)\big\|Y(x,t)\big\|^2\frac{1}{w_{k,n}(t)}\mathrm{d}x, \tag{4.38}$$

to 0, one obtains

$$w_{k,n}(t) = \int_{-\infty}^{\infty} m_{k,n}(x,t)\big\|Y(x,t)\big\|^2\mathrm{d}x. \tag{4.39}$$

This also implies a frame-by-frame partial power parameter update. Therefore, the HTC method essentially amounts to the BHC, if $\Omega_k(t)$ and $w_{k,n}(t)$ are both represented in a nonparametric way and if no constraints are assumed on their smoothness.

Next we shall introduce a penalizing term into the objective function $\Phi(\mathbf{\Theta}, m)$ in order to enforce the smoothness constraints on $\Omega_k(t)$ and $w_{k,n}(t)$. In a Bayesian point of view, this penalizing term corresponds to the prior distribution term $\log P(\mathbf{\Theta})$ when thinking of maximizing the posterior probability $L(\mathbf{\Theta}) + \log P(\mathbf{\Theta})$. Following the same way adopted in the regularization theory, which is often used in the image processing area, we can use the square integral of the first order partial derivative of $\Omega_k(t)$:

$$\int_{-\infty}^{\infty} \left(\frac{\partial \Omega_k(t)}{\partial t} \right)^2 \mathrm{d}t. \tag{4.40}$$

The smoother $\Omega_k(t)$ is, the smaller this value. Hence, when thinking of minimizing

$$\Phi(\mathbf{\Theta}, m) + \eta \sum_{k=1}^{K} \int_{-\infty}^{\infty} \left(\frac{\partial \Omega_k(t)}{\partial t} \right)^2 \mathrm{d}t, \tag{4.41}$$

one must try to make as small as possible not only $\Phi(\mathbf{\Theta}, m)$ but also the second term as well. η is a constant parameter that should be chosen experimentally to control the effect of the two terms. The larger this value, the flatter the contour tends to be estimated.

Let us consider the discrete-time case where the first order derivative of $\Omega_k(t)$ is approximated by the difference between the values taken at adjacent time points. Hence, Eq. (4.41) is written as

$$\begin{aligned} &\sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{i=1}^{I} \int_{-\infty}^{\infty} \Delta t \Bigg(m_{k,n}(x, t_i) \big\| Y(x, t_i) \big\|^2 \log \frac{m_{k,n}(x, t_i) \big\| Y(x, t_i) \big\|^2}{\mathcal{W}_{k,n}(x, t_i)} \\ &- \Big(m_{k,n}(x, t_i) \big\| Y(x, t_i) \big\|^2 - \mathcal{W}_{k,n}(x, t_i) \Big) \Bigg) \mathrm{d}x + \eta \sum_{k=1}^{K} \sum_{i=2}^{I} \Delta t \left(\frac{\Omega_k(t_{i-1}) - \Omega_k(t_i)}{\Delta t} \right)^2. \end{aligned} \tag{4.42}$$

Similarly, we shall include a penalizing term also for $w_{k,n}(t)$. The update equations for $\Omega_k(t)$ and $w_{k,n}(t)$ can then be derived using Eq. (4.42) and the maximum posterior estimation algorithm is thus formalized. The rest of the formulation shall be omitted.

4.4.2 Parametric HTC

We will consider here, on the other hand, the parametric case where $w_{k,n}(t)$ and $\Omega_k(t)$ are represented by the parametric models shown in Subsection 4.3.3. Let $\mathbf{\Theta}$ refers to $\{\{\Omega_{k,i}\}_{i=1}^{I}, \{v_{k,n}\}_{n=1}^{N}, \{u_{k,y}\}_{y=0}^{Y-1}, w_k, \tau_k, \phi_k\}_{k=1}^{K}$.

We define by $\|Y(x,t)\|^2$ the time-logfrequency power spectrum of the signal of interest obtained by the constant Q analysis. Let us introduce a masking function $m_{k,n,y}(x,t)$ that

extracts the spectro-temporal components associated with the y^{th} temporal element of the n^{th} partial of the k^{th} source from $\|Y(x,t)\|^2$. For $(x,t) \in \mathbb{R}^2$, $m_{k,n,y}(x,t)$ indicates the percentage of the portion of $\|Y(x,t)\|^2$ shared to the y^{th} temporal element of the n^{th} partial of the k^{th} source, such that satisfies

$$\sum_{k=1}^{K}\sum_{n=1}^{N}\sum_{y=0}^{Y-1} m_{k,n}(x,t) = 1 \tag{4.43}$$

$$0 < m_{k,n,y}(x,t) < 1, \quad k \in \{1,\cdots,K\},\ n \in \{1,\cdots,N\},\ y \in \{0,\cdots,Y-1\}. \tag{4.44}$$

Binary mask technique is often used in the research area of CASA and multichannel blind source separation to separate sources by allocating all the component in each time-frequency bin to a single source. On the contrary, the spectral masking function $m_{k,n,y}(x,t)$ is similar in some sense to this technique but could be understood as a masking function that has a fuzzy membership to every source. A portion of the observed power spectrum is thus given arbitrarily by

$$m_{k,n,y}(x,t)\big\|Y(x,t)\big\|^2, \quad (x,t) \in \mathbb{R}^2 \tag{4.45}$$

which we call again as a "spectral cluster". As we expect the spectral cluster to be associated with the auditory stream, we need to introduce a measure function that specifies how well the spectral cluster fits all the Bregman's grouping cues. We shall use again the I divergence between $m_{k,n,y}(x,t)\|Y(x,t)\|^2$ and the spectro-temporal model we derived in Section 3.2:

$$\mathcal{W}_{k,n,y}(x) \triangleq \frac{w_k v_{k,n} u_{k,y}}{2\pi\sigma\phi_k} e^{-\frac{(x-\Omega_k(b)-\log n)^2}{2\sigma^2} - \frac{(b-\tau_k-y\phi_k)^2}{2\phi_k^2}}, \tag{4.46}$$

which is written as

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\Bigg(m_{k,n,y}(x,t)\big\|Y(x,t)\big\|^2 \log\frac{m_{k,n,y}(x,t)\big\|Y(x,t)\big\|^2}{\mathcal{W}_{k,n,y}(x,t)} - \Big(m_{k,n,y}(x,t)\big\|Y(x,t)\big\|^2 - \mathcal{W}_{k,n}(x,t)\Big)\Bigg)\mathrm{d}x\mathrm{d}t. \tag{4.47}$$

The optimal clustering can thus be achieved by minimizing their sum:

$$\Phi(\boldsymbol{\Theta}, m) = \sum_{k=1}^{K}\sum_{n=1}^{N}\sum_{y=0}^{Y-1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\Bigg(m_{k,n,y}(x,t)\big\|Y(x,t)\big\|^2 \log\frac{m_{k,n,y}(x,t)\big\|Y(x,t)\big\|^2}{\mathcal{W}_{k,n,y}(x,t)} - \Big(m_{k,n,y}(x,t)\big\|Y(x,t)\big\|^2 - \mathcal{W}_{k,n}(x,t)\Big)\Bigg)\mathrm{d}x\mathrm{d}t \tag{4.48}$$

with respect to $\boldsymbol{\Theta}$ and $m_{k,n}(x,t)$. To do so, we shall find it most convenient to minimize this objective function recursively with respect to $m_{k,n}(x,t)$ and $\boldsymbol{\Theta}$ while keeping the other

fixed. The minimization with respect to $m_{k,n}(x,t)$ decomposes the observed power spectrum using the auditory stream models estimated hypothetically at the previous step and the minimization with respect to $\mathbf{\Theta}$, on the other hand, updates the auditory stream models to a more convincing one using these separate clusters. Both steps necessarily decreases the objective function, which is bounded below, and the convergence of this recursive algorithm is thus guaranteed.

The update equation for the spectral masking function $m_{k,n}(x,t)$ such that minimizes $\Phi(\mathbf{\Theta}, m)$ when $\mathbf{\Theta}$ is fixed is obtained analytically as

$$\widehat{m}_{k,n,y}(x,t) = \frac{\mathcal{W}_{k,n,y}(x,t)}{\sum_k \sum_n \sum_y \mathcal{W}_{k,n,y}(x,t)}. \tag{4.49}$$

Substituting this result into Eq. (4.48), one obtains again the I divergence between the whole observed power spectrum and the mixture of all the spectro-temporal source models:

$$\Phi(\mathbf{\Theta}, \widehat{m}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left(\left\| Y(x,t) \right\|^2 \log \frac{\left\| Y(x,t) \right\|^2}{\sum_k \sum_n \sum_y \mathcal{W}_{k,n,y}(x,t)} - \left(\left\| Y(x,t) \right\|^2 - \sum_k \sum_n \sum_y \mathcal{W}_{k,n,y}(x,t) \right) \right) \mathrm{d}x\mathrm{d}t. \tag{4.50}$$

As mentioned beforehand, this clustering objective $\Phi(\mathbf{\Theta}, \widehat{m})$ can be monotonically decreased by the following 2-step iteration:

Step 0 Set initially $\mathbf{\Theta}_0$ and put $\ell = 1$.

Step 1 Update the spectral masking function: $\boldsymbol{m}^{(\ell)} = \operatorname*{argmin}_m \Phi(\mathbf{\Theta}^{(\ell-1)}, m)$

Step 2 Update $\mathbf{\Theta}$ to $\mathbf{\Theta}^{(\ell)}$ such that $\Phi(\mathbf{\Theta}, m^{(\ell)}) \leqq \Phi(\mathbf{\Theta}^{(\ell-1)}, m^{(\ell)})$ and set $\ell \leftarrow \ell + 1$ and then return to Step 1.

Setting to zero the partial derivative of Eq. (4.50), the update equation of each parameter

at Step 2 of the ℓ-th iteration is derived analytically as follows:

$$w_k^{(\ell)} = \sum_{n=1}^{N}\sum_{y=0}^{Y-1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} m_{k,n,y}^{(\ell)}(x,t)\big\|Y(x,t)\big\|^2 \mathrm{d}x\mathrm{d}t, \tag{4.51}$$

$$\Omega_{k,0}^{(\ell)} = \frac{1}{w_k^{(\ell)}}\sum_{n=1}^{N}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} m_{k,n,y}^{(\ell)}(x,t)\big\|Y(x,t)\big\|^2 (x-\log n)\mathrm{d}x\mathrm{d}t, \tag{4.52}$$

$$\tau_k^{(\ell)} = \frac{1}{w_k^{(\ell)}}\sum_{n=1}^{N}\sum_{y=0}^{Y-1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} m_{k,n,y}^{(\ell)}(x,t)\big\|Y(x,t)\big\|^2 (t-y\phi_k^{(\ell-1)})\mathrm{d}x\mathrm{d}t, \tag{4.53}$$

$$v_{k,n}^{(\ell)} = \frac{1}{w_k^{(\ell)}}\sum_{y=0}^{Y-1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} m_{k,n,y}^{(\ell)}(x,t)\big\|Y(x,t)\big\|^2 \mathrm{d}x\mathrm{d}t, \tag{4.54}$$

$$u_{k,y}^{(\ell)} = \frac{1}{w_k^{(\ell)}}\sum_{n=1}^{N}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} m_{k,n,y}^{(\ell)}(x,t)\big\|Y(x,t)\big\|^2, \tag{4.55}$$

$$\phi_k^{(\ell)} = \frac{1}{2w_k^{(\ell)}}\left(\left(\alpha_k^{(\ell)2} + 4\beta_k^{(\ell)}w_k^{(\ell)}\right)^{\frac{1}{2}} - \alpha_k^{(\ell)}\right), \tag{4.56}$$

$$\begin{cases} \alpha_k^{(\ell)} \triangleq \displaystyle\sum_{n=1}^{N}\sum_{y=0}^{Y-1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} m_{k,n,y}^{(\ell)}(x,t)\big\|Y(x,t)\big\|^2 y(t-\tau_k^{(\ell)})\mathrm{d}x\mathrm{d}t \\ \beta_k^{(\ell)} \triangleq \displaystyle\sum_{n=1}^{N}\sum_{y=0}^{Y-1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} m_{k,n,y}^{(\ell)}(x,t)\big\|Y(x,t)\big\|^2 (t-\tau_k^{(\ell)})^2\mathrm{d}x\mathrm{d}t \end{cases}.$$

We showed in the above only the update equation for $\Omega_{k,0}$, which is the coefficient of the 0^{th} order term in the polynomial-type F_0 trajectory function given by Eq. (4.24). Note that the update equations for the coefficients of the other terms can be derived analytically as well. On the other hand, the update equation for each term in the cubic-spline-type F_0 trajectory function given by Eq. (4.25) is derived as follow:

$$\Omega_{k,i}^{(\ell)} = \frac{\displaystyle\sum_{n=1}^{N}\sum_{y=0}^{Y-1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(x-\widetilde{\Omega}_{k,n,i}^{(\ell)}(t;\boldsymbol{\Omega}_{k,i}^{(\ell)})\right)\frac{\partial\Omega_k(t)}{\partial\Omega_{k,i}} m_{k,n,y}^{(\ell)}(x,t)\big\|Y(x,t)\big\|^2 \mathrm{d}x\mathrm{d}t}{\displaystyle\sum_{n=1}^{N}\sum_{y=0}^{Y-1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\frac{\partial\Omega_k(t)}{\partial\Omega_{k,i}}\right)^2 m_{k,n,y}^{(\ell)}(x,t)\big\|Y(x,t)\big\|^2 \mathrm{d}x\mathrm{d}t}. \tag{4.57}$$

where $\boldsymbol{\Omega}_{k,i}^{(\ell)} = (\Omega_{k,1}^{(\ell)},\ldots,\Omega_{k,i-1}^{(\ell)},\Omega_{k,i}^{(\ell-1)},\Omega_{k,i+1}^{(\ell-1)},\ \ldots,\Omega_{k,I}^{(\ell-1)})$ and $\widetilde{\Omega}_{k,n,i}^{(\ell)}(t;\boldsymbol{\Omega}_{k,i}^{(\ell)}) = \Omega_k(t;\boldsymbol{\Omega}_{k,i}^{(\ell)}) - \frac{\partial\Omega_k(t)}{\partial\Omega_{k,i}}\Omega_{k,i}^{(\ell)} + \log n$ does not depend on $\Omega_{k,i}$ and $\frac{\partial\Omega_k(t)}{\partial\Omega_{k,i}}$ only depends on t and the fixed matrix M.

The so far constant σ, which depends on which value we set in the front-end constant Q analysis, can be regarded as a free variable σ_k for each k and its update equation can be derived analytically. The ML estimate of σ_k itself is not what we really want to obtain as its true value is already known, but by updating σ_k in parallel to the other parameters,

we expect that it could avoid other variables to be trapped in local optima. As we know empirically that in parameter learning of GMM, the update of the variance parameter of each Gaussian component often helps other parameters getting out of local optima, this is the reason why we treat σ_k as free parameters here. The update equation for σ_k is given as

$$\sigma_k^{(\ell)} = \left(\frac{1}{w_k^{(\ell)}} \sum_{n=1}^{N} \sum_{y=0}^{Y-1} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} m_{k,n,y}^{(\ell)}(x,t) \big\| Y(x,t) \big\|^2 \left(x - \Omega_k^{(\ell)}(t) - \log n \right)^2 \mathrm{d}x\mathrm{d}t \right)^{\frac{1}{2}}. \quad (4.58)$$

4.5 Bayesian HTC

4.5.1 Reformulation

We used in Subsection 4.4.2 the I-divergence [30] to measure the "distortion" between the two distributions:

$$\mathcal{I}(\boldsymbol{\Theta}) \triangleq \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left(\|Y(x,t)\|^2 \log \frac{\|Y(x,t)\|^2}{W(x,t;\boldsymbol{\Theta})} - \Big(\|Y(x,t)\|^2 - W(x,t;\boldsymbol{\Theta}) \Big) \right) \mathrm{d}x\mathrm{d}t, \quad (4.59)$$

where

$$W(x,t;\boldsymbol{\Theta}) = \sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{y=0}^{Y-1} \mathcal{W}_{k,n,y}(x,t), \quad (4.60)$$

is the sum of all the source models spread in the time-frequency plane, and we were looking for $\boldsymbol{\Theta}_{\text{opt}} = \operatorname{argmin}_{\boldsymbol{\Theta}} \mathcal{I}(\Theta)$. Keeping only the terms depending on $\boldsymbol{\Theta}$ and reversing the sign of this expression, one defines the following function to maximize w.r.t. $\boldsymbol{\Theta}$:

$$\mathcal{J}(\boldsymbol{\Theta}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Big(\|Y(x,t)\|^2 \log W(x,t;\boldsymbol{\Theta}) - W(x,t;\boldsymbol{\Theta}) \Big) \mathrm{d}x\mathrm{d}t. \quad (4.61)$$

Using this function $\mathcal{J}$, one can derive the likelihood of the parameter $\boldsymbol{\Theta}$:

$$P(Y|\boldsymbol{\Theta}) \triangleq e^{\mathcal{J}(\boldsymbol{\Theta}) - \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \log \Gamma \left(1 + \|Y(x,t)\|^2 \right) \mathrm{d}x\mathrm{d}t}, \quad (4.62)$$

where $\Gamma(\cdot)$ is the Gamma function and the second part of the exponent ensures that we obtain a probability measure. One can indeed see this probability as the joint probability of all the variables $\|Y(x,t)\|^2$ independently following Poisson-like distributions of parameter $W(x,t)$. This way of presenting the problem enables us to interpret it as a Maximum *A Posteriori* (MAP) estimation problem and to introduce prior functions on the parameters as

follows, using Bayes theorem:

$$\begin{aligned}\widehat{\boldsymbol{\Theta}}_{\mathrm{MAP}} &= \operatorname*{argmax}_{\boldsymbol{\Theta}} P(\boldsymbol{\Theta}|Y) \\ &= \operatorname*{argmax}_{\boldsymbol{\Theta}} \Big(\log P(Y|\boldsymbol{\Theta}) + \log P(\boldsymbol{\Theta}) \Big) \\ &= \operatorname*{argmax}_{\boldsymbol{\Theta}} \Big(\mathcal{J}(\boldsymbol{\Theta}) + \log P(\boldsymbol{\Theta}) \Big). \end{aligned} \tag{4.63}$$

Our goal is now equivalent to the maximization with respect to $\boldsymbol{\Theta}$ of $\mathcal{J}(\boldsymbol{\Theta}) + \log P(\boldsymbol{\Theta})$. The problem is that in the term $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \|Y(x,t)\|^2 \log \sum_{k,n,y} W_{k,n,y}(x,t)\,\mathrm{d}x\mathrm{d}t$, there is a sum inside the logarithm, and that is why we cannot obtain an analytical solution. But if we introduce non-negative membership degrees $m_{k,n,y}(x,t)$ summing to 1 for each (x,t), one can write, using the concavity of the logarithm:

$$\log \sum_{k,n,y} \mathcal{W}_{k,n,y}(x,t;\boldsymbol{\Theta}) = \log \sum_{k,n,y} m_{k,n,y}(x,t) \frac{\mathcal{W}_{k,n,y}(x,t)}{m_{k,n,y}(x,t)} \tag{4.64}$$

$$= \log \left\langle \frac{\mathcal{W}_{k,n,y}(x,t)}{m_{k,n,y}(x,t)} \right\rangle_m \tag{4.65}$$

$$\geq \left\langle \log \frac{\mathcal{W}_{k,n,y}(x,t)}{m_{k,n,y}(x,t)} \right\rangle_m = \sum_{k,n,y} m_{k,n,y}(x,t) \log \frac{\mathcal{W}_{k,n,y}(x,t)}{m_{k,n,y}(x,t)}, \tag{4.66}$$

where $\langle\cdot\rangle_m$ denotes the convex combination with coefficients m. Moreover, the inequality (4.64) becomes an equality for

$$\widehat{m}_{k,n,y}(x,t) = \frac{\mathcal{W}_{k,n,y}(x,t)}{\sum_k \sum_n \sum_y \mathcal{W}_{k,n,y}(x,t)}. \tag{4.67}$$

We can thus iteratively maximize the likelihood by alternately updating $\boldsymbol{\Theta}$ and the membership degrees m, which act as auxiliary parameters, while keeping the other fixed:

(E-step) Update the spectral masking function: $\boldsymbol{m}^{(\ell)} = \operatorname*{argmax}_m \mathcal{J}^+(\boldsymbol{\Theta}^{(\ell-1)}, m)$.

(M-step) Update $\boldsymbol{\Theta}$ to $\boldsymbol{\Theta}^{(\ell)}$ such that $\mathcal{J}^+(\boldsymbol{\Theta}, m^{(\ell)}) + \log P(\boldsymbol{\Theta}) \geq \mathcal{J}^+(\boldsymbol{\Theta}^{(\ell-1)}, m^{(\ell)}) + \log P(\boldsymbol{\Theta}^{\ell-1})$ and set $\ell \leftarrow \ell + 1$ and then return to E-step.

with

$$\mathcal{J}^+(\boldsymbol{\Theta}, m) \triangleq \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Big(\sum_{k,n,y} m_{kny}(x,t) \|Y(x,t)\|^2 \log \frac{\mathcal{W}_{k,n,y}(x,t)}{m_{k,n,y}(x,t)} - W(x,t;\boldsymbol{\Theta}) \Big) \mathrm{d}x\mathrm{d}t. \tag{4.68}$$

One must notice that this iterative procedure is called the EM algorithm [33]. For all m, we indeed have from (4.64) that

$$\mathcal{J}(\boldsymbol{\Theta}) + \log P(\boldsymbol{\Theta}) \geq \mathcal{J}^+(\boldsymbol{\Theta}, m) + \log P(\boldsymbol{\Theta}), \tag{4.69}$$

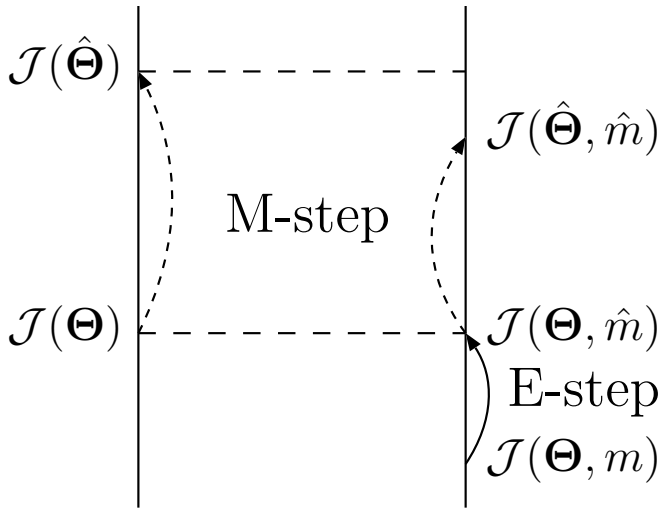


Figure 4.5 *Optimization through the EM algorithm. During the E-step, the auxiliary parameter m is updated to wide$\hat{m}$ so that $\mathcal{J}(\mathbf{\Theta}) = \mathcal{J}^{+}(\mathbf{\Theta}, \text{wide}\hat{m})$. Then, during the M-step, $\mathcal{J}^{+}(\mathbf{\Theta}, \widehat{m})$ is optimized w.r.t. $\mathbf{\Theta}$, ensuring that $\mathcal{J}(\widehat{\mathbf{\Theta}}) \geq \mathcal{J}^{+}(\widehat{\mathbf{\Theta}}, \widehat{m}) \geq \mathcal{J}^{+}(\mathbf{\Theta}, \widehat{m}) = \mathcal{J}(\mathbf{\Theta})$. The local maximization of $\mathcal{J}(\mathbf{\Theta})$ can thus be performed through the maximization of the auxiliary function $\mathcal{J}^{+}(\mathbf{\Theta}, m)$ alternately w.r.t. m and $\mathbf{\Theta}$.*

and $\mathcal{J}^{+}(\mathbf{\Theta}, \hat{m})$ can be used as an auxiliary function to maximize, enabling us to obtain analytical update equations. The optimization process is illustrated in Fig. 4.5.1. The E-step is straightforward and is dealt with in exactly the same way as in Chapter 3.

4.5.2 Prior Distribution

As seen in Subsection 4.5.1, the optimization of our model can be naturally extended to a Maximum *A Posteriori* (MAP) estimation by introducing prior distributions $P(\mathbf{\Theta})$ on the parameters, which work as penalty functions that try to keep the parameters within a specified range. The parameters which are the best compromise with empirical constraints are then obtained through equation Eq. (4.63).

By introducing such a prior distribution on v_{kn}, it becomes possible to prevent half-pitch errors, as the resulting source model would usually have a harmonic structure with zero power for all the odd order harmonics, which is abnormal for usual speech and instruments. A prior distribution on $u_{k,y}$, on the other hand, helps to avoid overfitting many source models

to the observed power envelope of a single source, as the resulting individual source models in this case would often have abnormal power envelopes. We apply the Dirichlet distribution, which is explicitly given by:

$$p(\boldsymbol{v}_k) \triangleq \frac{\Gamma\Big(\sum_n (d_v \bar{v}_n + 1)\Big)}{\prod_n \Gamma(d_v \bar{v}_n + 1)} \prod_n {v_{k,n}}^{d_v \bar{v}_n}, \tag{4.70}$$

$$p(\boldsymbol{u}_k) \triangleq \frac{\Gamma\Big(\sum_y (d_u \bar{u}_y + 1)\Big)}{\prod_y \Gamma(d_u \bar{u}_y + 1)} \prod_y {u_{k,y}}^{d_u \bar{u}_y}, \tag{4.71}$$

where $\bar{v}_n$ and $\bar{u}_y$ is the most preferred 'expected' values of $v_{k,n}$ and $u_{k,y}$ such that $\sum_n \bar{v}_n = 1$ and $\sum_y \bar{u}_y = 1$, d_v and d_u the weighting constants of the priors and $\Gamma(\cdot)$ the Gamma function. The maximum values for $P(\boldsymbol{v}_k)$ and $P(\boldsymbol{u}_k)$ are taken respectively when $v_{k,n} = \bar{v}_n$ for all n and $uk, y = \bar{u}_y$ for all y. When d_v and d_u are zero, $P(\boldsymbol{v}_k)$ and $P(\boldsymbol{u}_k)$ become uniform distributions. The choice of this particular distribution allows us to give an analytical form of the update equations of $v_{k,n}$ and $u_{k,y}$:

$$v_{k,n}^{(\ell)} = \frac{1}{d_v + w_k^{(\ell)}} \left(d_v \bar{v}_n + \sum_{y=0}^{Y-1} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} m_{k,n,y}^{(\ell)}(x,t) \big\| Y(x,t) \big\|^2 \mathrm{d}x\mathrm{d}t \right), \tag{4.72}$$

$$u_{k,y}^{(\ell)} = \frac{1}{d_u + w_k^{(\ell)}} \left(d_u \bar{u}_y + \sum_{n=1}^{N} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} m_{k,n,y}^{(\ell)}(x,t) \big\| Y(x,t) \big\|^2 \mathrm{d}x\mathrm{d}t \right). \tag{4.73}$$

Although the spline model can be used as is, one can also introduce in the same way a prior distribution on the parameters z_j of the spline F_0 contour, in order to avoid an overfitting problem with the spline function. Indeed, spline functions have a tendency to take large variations, which is not natural for the F_0 contour of a speech utterance. Moreover, the F_0 contour might also be hard to obtain on voiced parts with relatively lower power, or poor harmonicity. The neighboring voiced portions with higher power help the estimation over these intervals by providing a good prior distribution.

To build this prior distribution, we assume that the z_j form a Markov chain, such that

$$\prod_{k=1}^{K} P(\Omega_{k,0}, \ldots, \Omega_{k,n}) = \prod_{k=1}^{K} P(\Omega_{k,0}) \prod_{j=1}^{I-1} P(\Omega_{k,j} | \Omega_{k,j-1}), \tag{4.74}$$

and assume furthermore that $\Omega_{k,0}$ follows a uniform distribution and that, conditionally to $\Omega_{k,j-1}$, $\Omega_{k,j}$ follows a Gaussian distribution of center $\Omega_{k,j-1}$ and variance σ_s^2 corresponding to the weighting constant of the prior distribution:

$$P(\Omega_{k,j} | \Omega_{k,j-1}) = \frac{1}{\sqrt{2\pi}\sigma_s} e^{-\frac{(\Omega_{k,j} - \Omega_{k,j-1})^2}{2\sigma_s^2}}. \tag{4.75}$$

In the derivative with respect to $\Omega_{k,j}$ used above to obtain (Eq. (4.57)) add up two new terms $\frac{\partial \log P(\Omega_{k,j}|\Omega_{k,j-1})}{\partial \Omega_{k,j}} + \frac{\partial \log P(\Omega_{k,j+1}|\Omega_{k,j})}{\partial \Omega_{k,j}}$, and the update equation Eq. (4.57) then becomes

$$\Omega_{k,j}^{(\ell)} = \frac{\frac{2}{\sigma_s^2} \cdot \frac{\Omega_{k,j-1}^{(\ell)} + \Omega_{k,j+1}^{(\ell-1)}}{2} + A_j^{(\ell)}}{\frac{2}{\sigma_s^2} + B_j^{(\ell)}}, \tag{4.76}$$

where $A_j^{(\ell)}$ and $B_j^{(\ell)}$ are respectively the numerator and denominator of the right term of equation Eq. (4.57). The update equation for the boundary points is derived similarly.

The update equations for the rest of the parameters are given as is shown in Subsection 4.4.2.

4.6 Experimental Evaluation

A perceptual unit as defined in ASA does not necessarily coincide with a single physical event, but we may be able to show by investigating in an engineering way through experimental evaluations the performance of our algorithm in a particular case how deeply Bregman's grouping cues are related to a physical phenomenon. In this subsection, to show the effectiveness of the Harmonic Temporal Clustering (hereafter HTC), we perform F_0 estimation experiments on various kinds of acoustic signals and evaluate its performance.

4.6.1 Note Estimation from Acoustic Signals of Music

We first evaluated accuracies of note estimation using real-performed music acoustic signals excerpted from RWC music database [44]. The experimental data used for the evaluation can be seen in Table 4.1. The Power spectrum time series was analyzed by the wavelet transform (constant Q analysis) using Gabor-wavelet basis functions with a time resolution of 16 ms for the lowest frequency subband on an input signal digitalized at a 16 kHz sampling rate. To speed up the computation time, we set the time resolution across all the subbands equally to 16ms. The lower bound of the frequency range and the frequency resolution were 60 Hz and 12 cents, respectively. The initial parameters of $(\mu_{k0}, \tau_k | k = 1, \cdots, K)$ for the HTC source models were automatically determined by picking the 60 largest peaks in the observed spectrogram of 400 consecutive frames (6.4s). After the parameters converged, the

Table 4.1 *List of the experimental data excerpted from RWC music database [44].*

Symbol	Title (Genre)	Composer/Player	Instruments	Ave. source#
data(1)	Crescent Serenade (Jazz)	S. Yamamoto	Guitar	2.13
data(2)	For Two (Jazz)	H. Chubachi	Guitar	2.67
data(3)	Jive (Jazz)	M. Nakamura	Piano	1.86
data(4)	Lounge Away (Jazz)	S. Yamamoto	Guitar	4.04
data(5)	For Two (Jazz)	M. Nakamura	Piano	2.34
data(6)	Jive (Jazz)	H. Chubachi	Guitar	1.78
data(7)	Three Gimnopedies no. 1 (Classic)	E. Satie	Piano	2.96
data(8)	Nocturne no.2, op.9-2(Classic)	F. F. Chopin	Piano	1.55

source model, whose energy per unit time given by $\frac{w_k}{Y\phi_k}$ was smaller than a threshold, was considered to be silent. The experimental conditions are shown in detail in Table 4.2.

We chose *'PreFEst' [45] for comparison, as it is one of the most frequently cited works which is dedicated to multipitch analysis. Since PreFEst extracts only the most dominant F_0 trajectory and does not include a specific procedure of estimating the number of sources, we included intensity thresholding as well for the F_0 candidate truncation.

As the HTC method generates F_0, onset time and offset time with continuous values, we quantize them to the closest note and the closest frame number in order to match with the format of the reference. Using the hand-labeled ground truth data as references, F_0 accuracies were computed by

$$\frac{X - D - I - S}{X} \times 100(\%).$$

X : # of the total frames of the voiced parts

D : # of deletion errors

I : # of insertion errors

S : # of substitution errors

*Note that we implemented for the evaluation only the module called 'PreFEst-core', a framewise F_0 likelihood estimation, and not included the one called 'PreFEst-back-end', a multi-agent based F_0 tracking algorithm. Refer to [45] for their details.

Table 4.2 *Experimental Conditions*

frequency analysis	Sampling rate	16 kHz
	frame shift	16 ms
	frequency resolution	12.0 cent
	frequency range	60–3000 Hz
HTC	# of HTC source models: K	60
	# of partials: N	6
	# of kernels in $U_k(t)$: Y	10
	$\bar{v}_n$	$0.6547 \times n^{-2}$
	$\bar{u}_y$	$0.2096 \times e^{-0.2y}$
	d_v, d_u	0.04
	time range of a spectrogram segment	400 frames (6.4 s)
	# of the segments	4 (total time: 25.6 s)
PreFEst [45]	F_0 resolution	20 cent
	# of partials	8
	# of tone models	200
	standard deviation of Gaussian	3.0
	$\bar{r}_n$	$0.6547 \times n^{-2}$
	$\bar{d}$ (prior contribution factor)	3.0

A typical example of the F_0, onset and offset estimates on a particular test data is shown in Fig.4.6 together with the hand-labeled ground truth data. The optimized model and the observed power spectrum time series are shown with 3D and grayscale displays in Fig.4.7.

To validate the performance of the proposed method, we compared the highest accuracy of the HTC method with that of the PreFEst among all the thresholds that were tested, which also shows the limit of the potential capability. The highest accuracies of PreFEst and HTC among all the thresholds we tested are shown in table Table 4.3 together with the number of insertion, deletion and substitution errors, respectively. Comparing these accuracies between PreFEst and HTC, HTC outperforms PreFEst for most of the data, which verifies its potential.

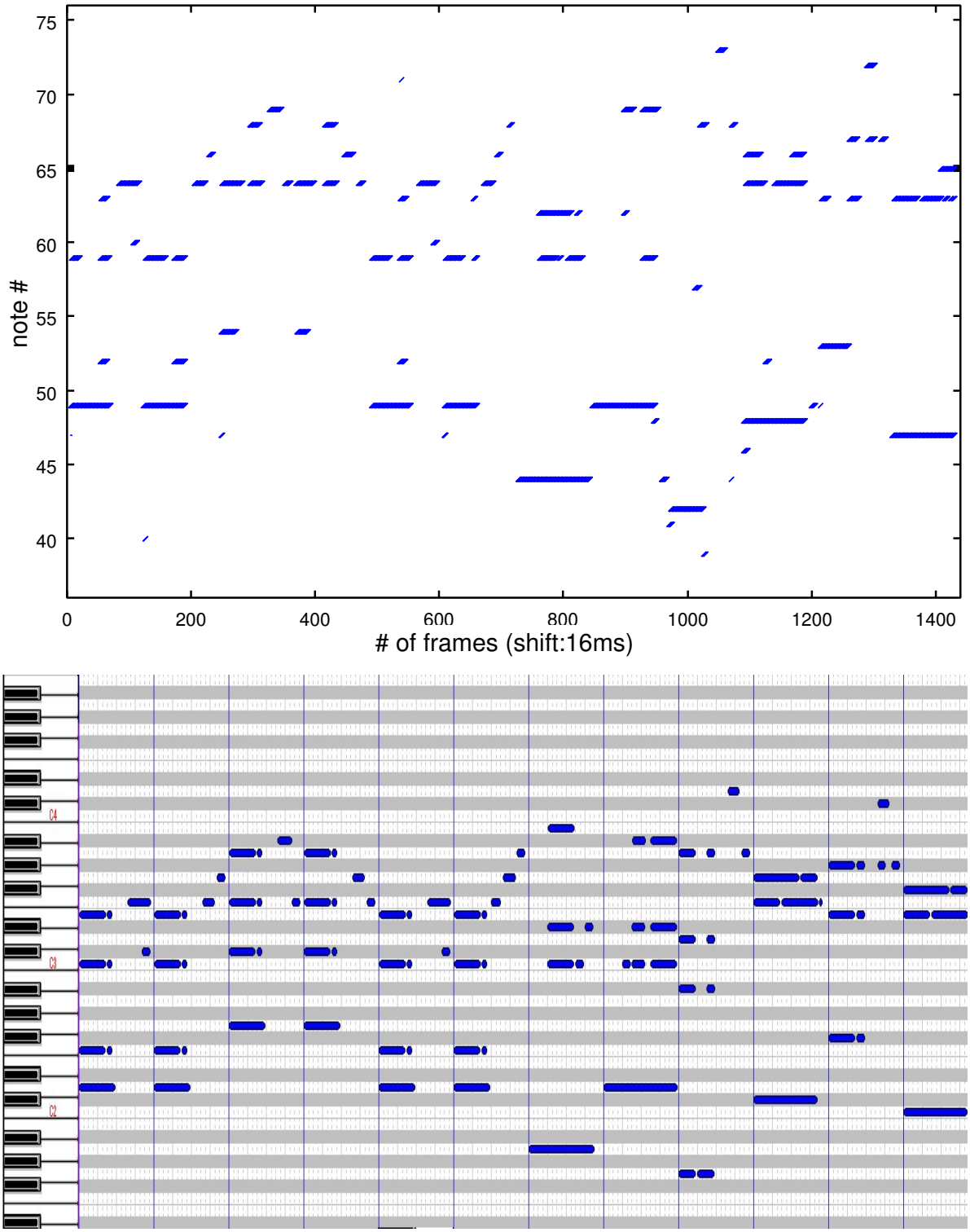


Figure 4.6 *Estimates of* $\mu_{k0}, \tau_k, Y\phi_k$ *(top) and piano-roll display of the reference MIDI (bottom)*

The workstation used to perform the experiments had a Pentium IV processor with 3.2 GHz clock speed and 2 GB memory. With our implementation with the conditions listed in Table 4.2, the computational time for analyzing an acoustic signal of 25.6 seconds length

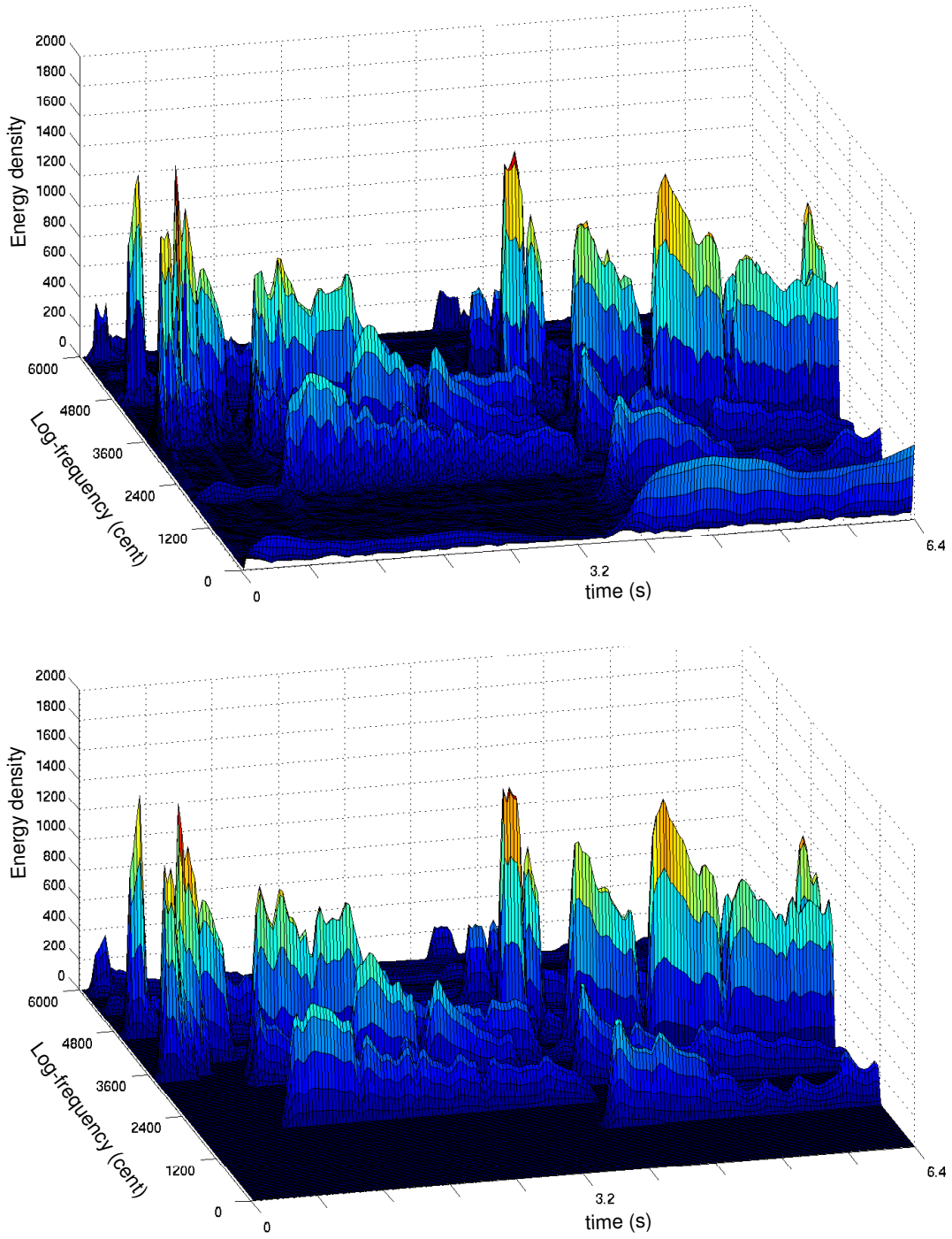


Figure 4.7 *Observed spectrogram (top) and estimated spectro-temporal model (bottom)*

was about 2 minutes. In most cases, the parameters of the HTC source models converged within less than 100 iteration cycles.

We also compared the HTC performances with different conditions: the time range of an analyzing spectrogram segment of 100, 200 and 400 frames, and the number of the HTC source models of 15, 30 and 60, respectively. Comparative results are shown in Table 4.4.

Table 4.3 *Accuracies of the PreFEst [45] and the HTC.*

		conventional 'PreFEst'[45]				*proposed 'HTC'*			
	X	Accuracy (%)	I	D	S	Accuracy (%)	I	D	S
data(1)	3063	74.2	383	327	81	**81.2**	210	312	55
data(2)	3828	71.8	455	397	228	**77.9**	241	397	208
data(3)	2671	55.9	553	500	126	**64.2**	313	524	120
data(4)	5798	**76.2**	476	650	254	75.2	361	769	310
data(5)	3366	**62.3**	565	515	190	62.2	465	627	178
data(6)	2563	48.8	531	597	185	**63.8**	304	476	147
data(7)	4244	53.6	801	830	337	**63.2**	427	734	403
data(8)	2227	57.6	367	482	96	**70.9**	278	291	79

From the results, one can see that the larger the time range of a spectrogram segment, the higher the accuracies. This shows that the domain of definition of t should be as large as possible for a higher performance of the HTC.

4.6.2 F_0 Determination of Single Speech in Clean Environment

We evaluated the accuracy of the F_0 contour estimation of our model on a database of speech recorded together with a laryngograph signal [11], consisting of one male and one female speaker who each spoke 50 English sentences for a total of 0.12h of speech, for the purpose of evaluation of F_0-estimation algorithms.

The power spectrum $\|Y(x,t)\|^2$ was calculated from an input signal digitized at a 16kHz sampling rate (the original data of the database was converted from 20kHz to 16 kHz) using a Gabor wavelet transform with a time resolution of 16ms for the lowest frequency subband. Higher subbands were downsampled to match the lowest subband resolution. The lower bound of the frequency range and the frequency resolution were respectively 50Hz and 14 cent. The spline contour was initially flat and set to 132Hz for the male speaker and 296Hz

Table 4.4 *Comparison of the HTC performances with different ranges of a spectrogram segment and the number of source models.*

		Time range: 100 frames, K: 15				*Time range: 200 frames, K: 30*			
	X	Accuracy (%)	I	D	S	Accuracy (%)	I	D	S
data(1)	3063	68.5	130	677	159	79.4	188	368	76
data(2)	3828	75.1	142	720	93	74.2	218	538	233
data(3)	2671	58.7	271	671	160	61.8	332	549	139
data(4)	5798	60.7	175	1863	243	66.6	232	1376	327
data(5)	3366	55.3	427	926	153	59.6	385	774	201
data(6)	2563	57.7	229	617	239	61.2	270	519	206
data(7)	4244	54.4	309	1226	400	**63.5**	470	619	461
data(8)	2227	58.8	234	598	85	68.2	315	325	69

		Time range: 400 frames, K: 60			
	X	Accuracy (%)	I	D	S
data(1)	3063	**81.2**	210	312	55
data(2)	3828	**77.9**	241	397	208
data(3)	2671	**64.2**	313	524	120
data(4)	5798	**75.2**	361	769	310
data(5)	3366	**62.2**	465	627	178
data(6)	2563	**63.8**	304	476	147
data(7)	4244	63.2	427	734	403
data(8)	2227	**70.9**	278	291	79

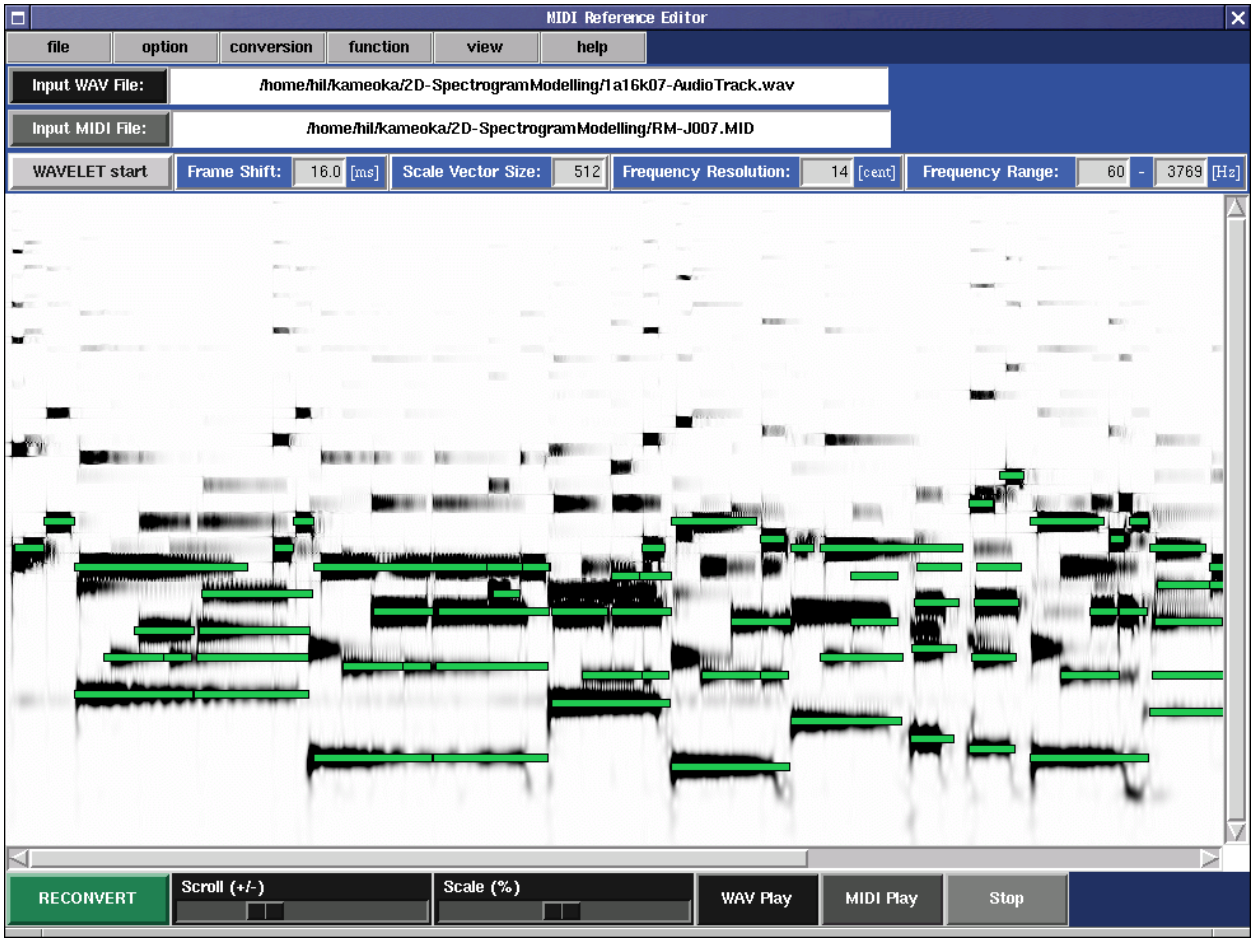
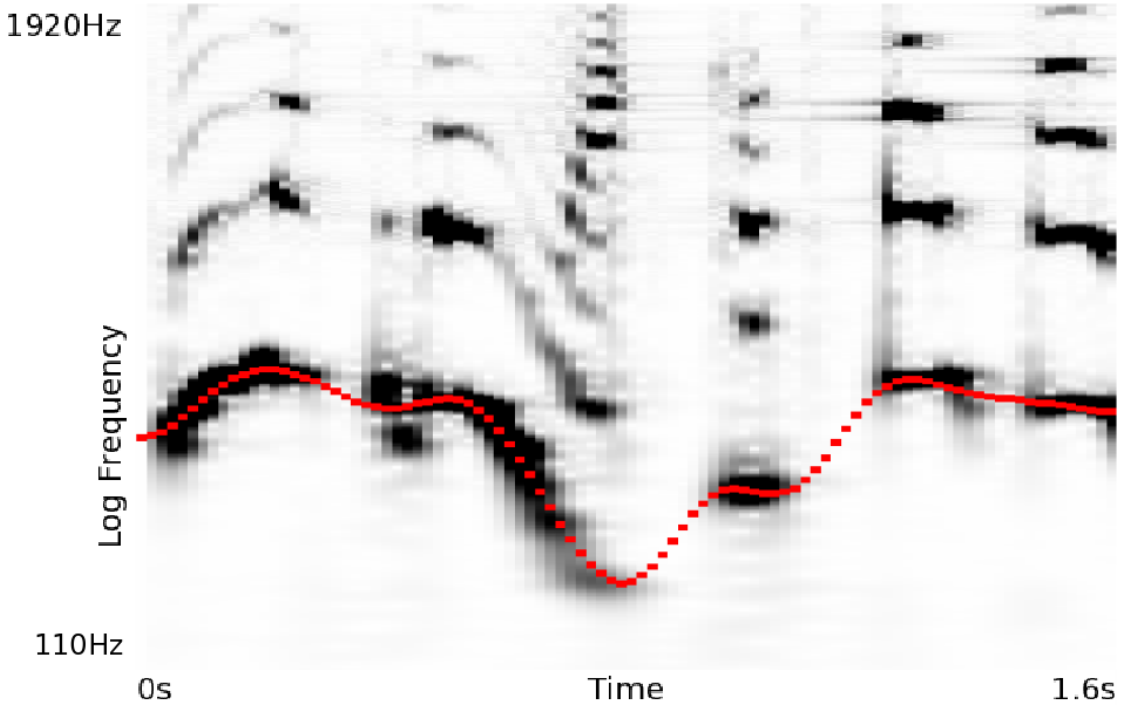


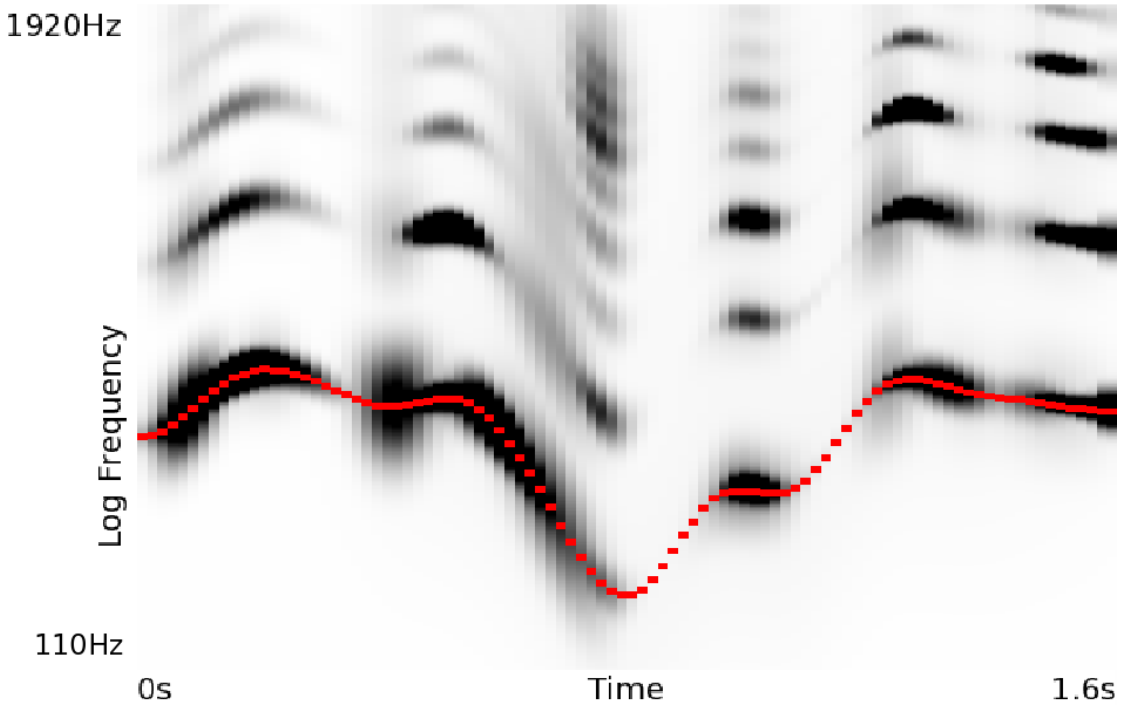
Figure 4.8 *A screenshot of the GUI editor we implemented to create the ground truth data set of note pitches, onsets and durations. The note events of the supplement MIDI data included in the RWC database, which are not temporally aligned with the corresponding real performed signal data, are displayed as rectangular objects over the spectrogram of the real performed signal. We are then able to edit the rectangular objects to align carefully the onset and offset times according to the background spectrogram.*

for the female speaker. The length of the interpolation intervals was fixed to 4 frames. For HTC, we used $K = 10$ source models, each of them with $N = 10$ harmonics. This is enough for F_0 estimation. For a better modeling of the spectrogram, one can use 40 or 60 harmonics for example. Temporal envelope functions were modeled using $Y = 3$ Gaussian kernels. The initial values of w_k, τ_k and ϕ_k were determined uniformly, and σ_k was fixed to 422 cents. For the prior functions, σ_s was fixed to 0.4, d_v to 0.04 and $(\bar{v}_n)_{1\leq n\leq N} = \frac{1}{N}(8, 8, 4, 2, 1, \ldots, 1)$.

We used as ground truth the F_0 estimates and the reliability mask derived by de Cheveigné *et al.* [25] under the following criteria: (1) any estimate for which the F_0 estimate was



(a) Observed spectrogram and estimated F_0 contour



(b) Modeled spectrogram and estimated F_0 contour

Figure 4.9 *Comparison of observed and modeled spectra ("Tsuuyaku denwa kokusai kaigi jimukyoku desu", female speaker). The estimated F_0 contour is reproduced on both the observed and modeled spectrograms to show the precision of the algorithm.*

obviously incorrect was excluded and (2) any remaining estimate for which there was evidence of vocal fold vibration was included. Frames outside the reliability mask were not taken into account during our computation of the accuracy, although our algorithm gives values for every point of the analysis interval by construction. As the spline function gives an analytical expression for the F_0 contour, we compare our result with the reference values at a sampling rate of 20kHz although all the analysis was performed with a time resolution of 16ms.

Deviations over 20% from the reference were deemed to be gross errors. The results can be seen in Table Table 4.5, with for comparison the results obtained by de Cheveigné *et al.* [25] for several other algorithms. Notations stand for the method used, as follows: **ac**: Boersma's autocorrelation method [14] [15], **cc**: cross-correlation [15], **shs**: spectral subharmonic summation [48] [15], **pda**: eSRPD algorithm [11] [120], **fxac**: autocorrelation function (ACF) of the cubed waveform [121], **fxcep**: cepstrum [121], **additive**: probabilistic spectrum-based method [35], **acf**: ACF [25], **nacf**: normalized ACF [25], **TEMPO**: the TEMPO algorithm [64], **YIN**: the YIN algorithm [25]. More details concerning these algorithms can be found in [25]. We can see that our model's accuracy for clean speech is comparable to the best existing single speaker F_0 extraction algorithms designed for that purpose.

4.6.3 Multipitch Estimation of Concurrent Speech

We present here results on the estimation of the F_0 contour of the co-channel speech of two speakers speaking simultaneously with equal average power. We used again the database mentioned above [11], and produced a total of 150 mixed utterances, 50 for each of the "male-male", "female-female" and "male-female" patterns, using each utterance only once and mixing it with another such that two utterances of the same sentence were never mixed together. We used our algorithm in the same experimental conditions as described in 4.6.2 for clean single-speaker speech, but using two spline F_0 contours. The spline contours were initially flat and set to 155Hz and 296Hz in the male-female case, 112Hz and 168Hz in the male-male case, and 252Hz and 378Hz in the female-female case.

The evaluation was done in the following way: only times inside the reliability mask of either of the two references were counted; for each reference point, if either one of the two spline F_0 contours lies within a criterion distance of the reference, we considered the estimation correct. We present scores for two criterion thresholds: 10% and 20% For comparison, tests using the WWB algorithm [115] introduced earlier were also performed, using the code

Table 4.5 *Gross error rates for several F_0 estimation algorithms on clean single speaker speech*

Method	Gross error (%)
pda	19.0
fxac	16.8
fxcep	15.8
ac	9.2
cc	6.8
shs	12.8
acf	1.9
nacf	1.7
additive	3.6
TEMPO	3.2
YIN	1.4
HTC (proposed)	3.5

made available by its authors. YIN could not be used as it does not perform multipitch estimation. Results summarized in Table 4.7 show that our algorithm outperforms the WWB algorithm on this experiment. Fig. 4.6.3 shows the spectrogram of a signal obtained by mixing the two Japanese utterances "oi wo ou" by a male speaker and "aoi" by a female speaker, together with the F_0 contours estimated by our method. One can see from Fig. 4.11 that the spectro-temporal cluster models are separately estimated such that each of them is associated with a single speaker's speech.

Table 4.6 *F_0 estimation of concurrent speech by multiple speakers, gross error for a difference with the reference higher than* 20% *and* 10%

Gross error threshold	20%		10%	
methods	HTC	WWB	HTC	WWB
Male-Female	93.3	81.8	86.8	81.5
Male-Male	96.1	83.4	87.9	69.0
Female-Female	98.9	95.8	95.6	90.8
Total	96.1	87.0	90.2	83.5

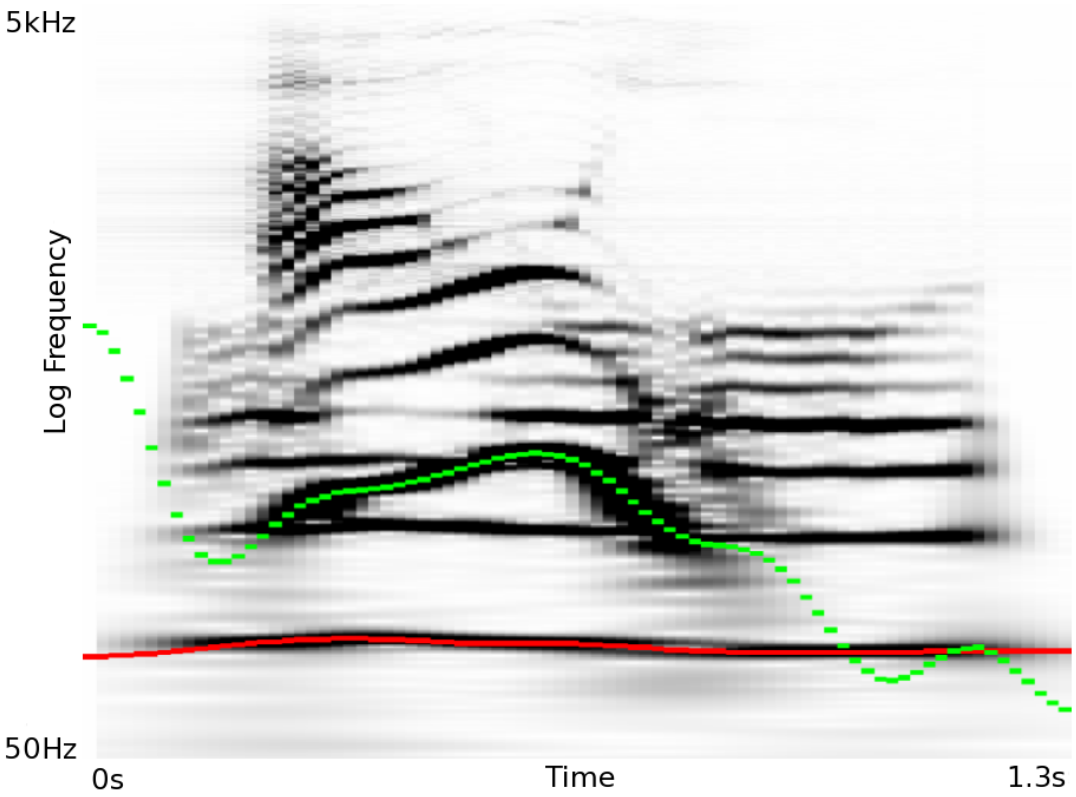
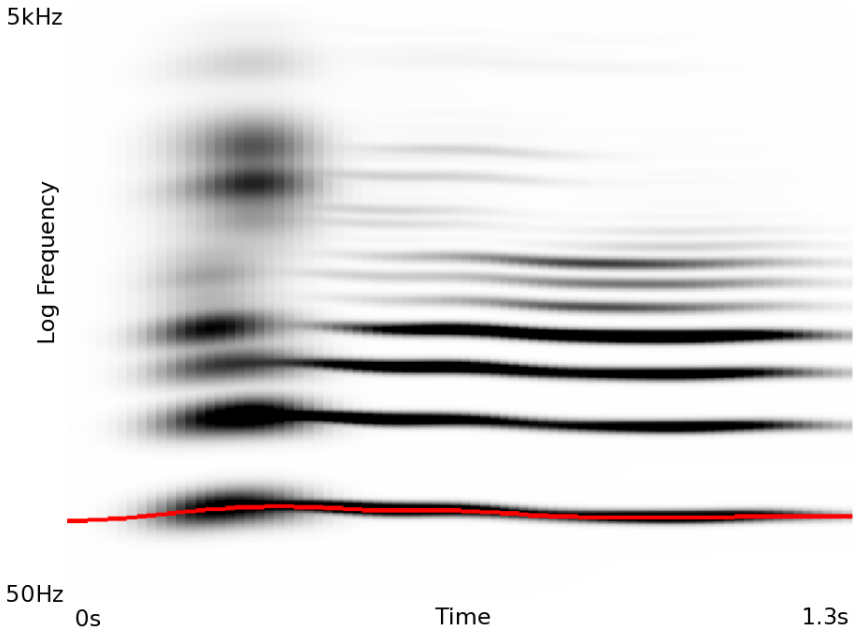


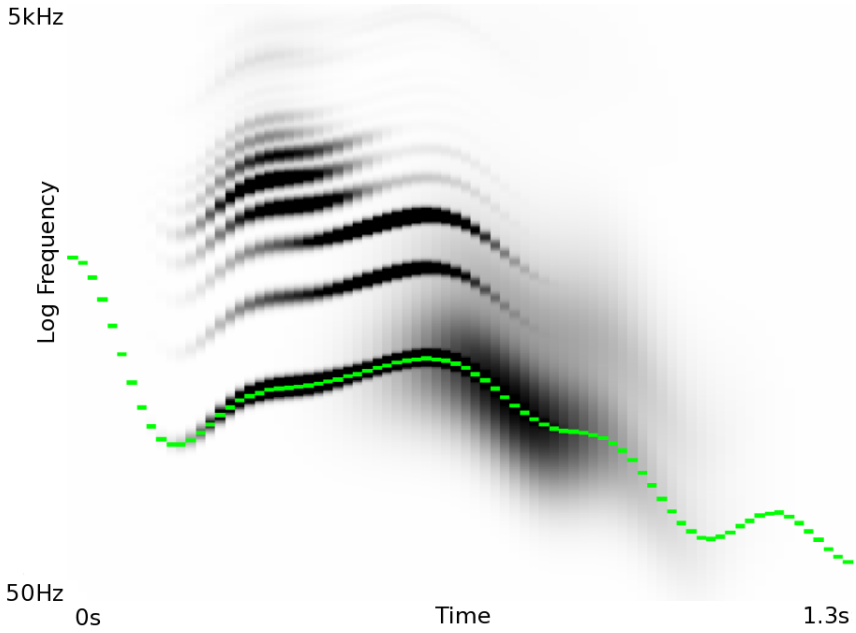
Figure 4.10 *The observed spectrogram of concurrent speech signal of two speakers talking at the same time and the estimated F_0 contour.*

4.7 Summary of Chapter 4

In this chapter, based on Bregman's grouping cues, we proposed a new methodology to estimate simultaneously the spectral structure of each source on the whole time-frequency



(a) *Modeled spectrogram, speaker 1*



(b) *Modeled spectrogram, speaker 2*

Figure 4.11 *Parametric representation of separated spectrograms. Fig. 4.6.3 shows the spectrogram of a signal obtained by mixing the two Japanese utterances "oi wo ou" by a male speaker and "aoi" by a female speaker, together with the F_0 contours estimated by our method. Fig. (a) and Fig. (b) show the parametric representations of the spectrograms of the utterances by the male and female speaker respectively, extracted from the mixed signal shown in Fig. 4.6.3.*

Table 4.7 F_0 *estimation of concurrent speech by multiple speakers, gross error for a difference with the reference higher than* 20% *and* 10%

Gross error threshold	20%		10%	
	HTC	WWB	HTC	WWB
Male-Female	93.3	81.8	86.8	81.5
Male-Male	96.1	83.4	87.9	69.0
Female-Female	98.9	95.8	95.6	90.8
Total	96.1	87.0	90.2	83.5

domain, which we called Harmonic-Temporal Clustering (HTC). Through evaluation experiments on the F_0 estimation of mixed speech signals and music signals, we showed that our method's accuracy outperforms the previous state-of-the-art methods of each of these areas.

Chapter 5

Joint Estimation of Spectral Envelope and Fine Structure

5.1 Introduction

F_0 determination and spectral envelope estimation both have a long history in speech research as they play a very important role in a wide range of speech processing activities such as speech compression, speech recognition and synthesis. Although many efforts have been devoted to both of these topics of research, the problem of determining F_0 and spectral envelope seems to have been tackled independently. The aim of this chapter is to highlight the importance of jointly determining the F_0 and the spectral envelope. From this standpoint, we will propose a new speech analyzer that jointly estimates F_0 and spectral envelope using a parametric speech source-filter model.

Up to now, a number of approaches to spectral envelope estimation have been investigated: LPC (Linear Predictive Coding) [53], PARCOR (Partial Autocorrelation) [55], LSP (Line Spectrum Pair) [56], pole-zero modeling techniques [68, 96, 58, 100], DAP (discrete all-pole) modeling [36], MVDR (minimum variance distortionless response) modeling [73], IAP (iterative all-pole) modeling [78], SEEVOC (spectral envelope estimation vocoder) [80], cepstrum [77] approaches such as LPC cepstrum [9], discrete cepstrum method [42], regularized discrete cepstrum method [20], discrete cepstrum method based on OLC (optimization of the likelihood criterion) [19] and true envelope estimator [52], STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) [63], and others such like [74, 85]. LPC [53, 55, 56] estimates the vocal tract characteristics modeled by an all-pole filter by assuming the excitation source signal of the vocal cords to be

a Gaussian white process, and has been applied with great success in many problems of speech processing. Cepstrum [77] is used to extract the spectral envelope by low-pass filtering a log-amplitude spectrum interpreted as a signal. The fact that MFCC (Mel-Frequency Cepstral Coefficients) [31] has become the most popular feature in speech recognition implies how well the cepstrum-based spectral envelopes express the vocal tract characteristics of speech. Furthermore, LPC cepstrum analyzer [9] is also a well-known and widely used spectral envelope extractor. DAP modeling [36] is an improved method of LPC, that tries to fit an all-pole transfer function to the discrete set of frequency points, and is known to be slightly more accurate than the classical LPC. Discrete cepstrum first presented by Galas and Rodet in [42] is an improved method of cepstrum, that estimates directly the cepstral coefficients through the minimization of a frequency-domain least squares criterion using discrete set of frequency points of the harmonic peaks. Regularized discrete cepstrum [20] is based on the discrete cepstrum approach that makes use of a regularization technique in order to enforce smoothness conditions on spectral envelope estimates. OLC cepstrum [19] is a further improved method that optimizes the cepstral coefficients through a different likelihood criterion, which is considered to be one of today's state-of-the-art methods. Another state-of-the-art technique, called STRAIGHT [63], starts by estimating the F_0 frequency, and then, using an analysis window varying in time according to the F_0 estimate precisely estimates the spectral envelope in a non-parametric way.

Making explicit use of the F_0 estimates via F_0 extractor, as opposed to the classical LPC and cepstrum, is certainly one of the reasons that discrete cepstrum methods [42, 20, 19] and STRAIGHT have been such a high-quality spectral envelope extractor. Accordingly, we can thus expect that the higher the accuracy of F_0 determination the more accurate the spectral envelope estimate.

However, although a huge number of F_0 estimation algorithms have been proposed [49, 50, 25], the reliability of them are still limited. The ambiguity involved in the definition of F_0 makes its determination difficult. In particular, one of the most diffucult problems is how to reduce subharmonic errors, or say, "octave errors". In a mathematical sense, the period of the signal $s(t)$, the inverse of F_0, is defined as the minimum of T such that $s(t) = s(t+T)$. This definition, however, applies strictly only to a perfectly periodic signal but as for speech, that departs from perfect periodicity, one must find the minimum of T such that $s(t) \approx s(t+T)$. The difficulty in F_0 estimation of the acoustic signal in a real environment, in general, stems from the fact that T that is 'likely' to be the smallest member

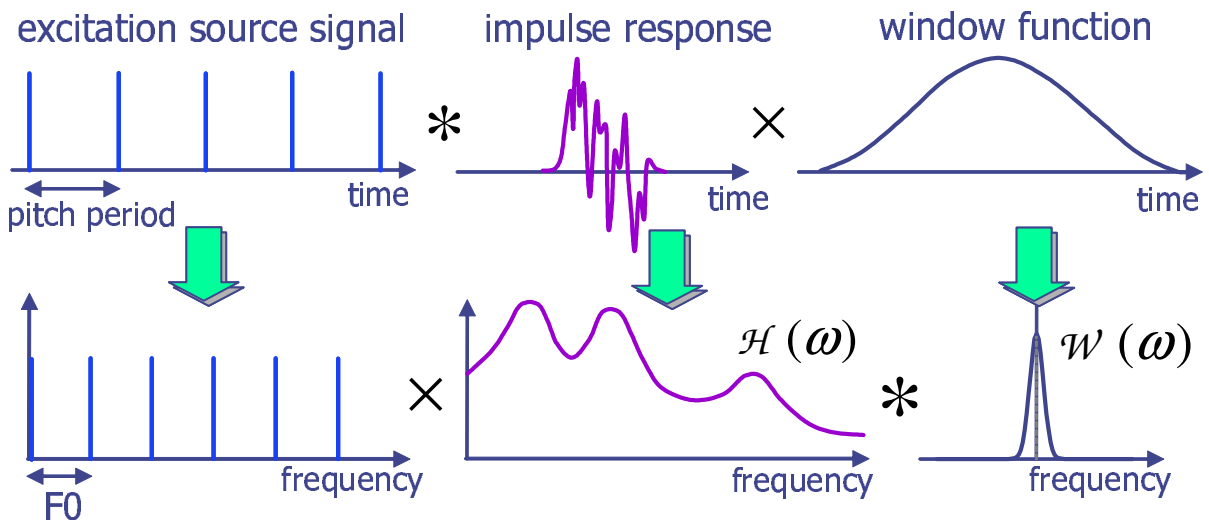


Figure 5.1 The linear system approximation model in the power spectrum domain

of the infinite set of time shifts that leave the signal 'almost' invariant is not always unique, since if T is the true pitch period one obtains $s(t) \approx s(t+nT)$ for all $n \in \mathbb{N}$, where $nT(n \neq 1)$ correspond to the periods of subharmonics. It thus sometimes become difficult to determine which one is the true pitch period, and choosing $nT(n \neq 1)$ instead of T is referred to as the subharmonic error. Making a subharmonic error amounts to misinterpreting as the true spectrum a harmonic structure with zero power for all the odd order harmonics, which is abnormal for usual speech and instruments. Such an error could thus be corrected if we knew in advance the true spectral envelope or at least by assuming that the spectral envelope are usually relatively smooth. For this reason, the spectral smoothness assumption has indeed been used to reduce subharmonic errors in F_0 estimation [10, 67].

So far, we have discussed that the more reliable the F_0 determination the more accurate the spectral envelope estimation will be, and, on the other hand, the more accurate the spectral envelope estimation the more reliable the F_0 determination will be. The F_0 determination and the spectral envelope estimation, having such a chicken and egg relationship, should thus be done jointly rather than independently in succession. This is the standpoint we chose in this chapter to formulate a joint estimation model of the spectral envelope and the fine structure.

5.2 Formulation of the Proposed Method

5.2.1 Speech Spectrum Modeling

A short-time segment of speech signal $y(t)$ can be modeled as an output of the linear system of the vocal tract impulse response $h(t)$ with the source excitation $s(t)$ such that

$$y(t) = \Big(s(t) * h(t)\Big)w(t), \tag{5.1}$$

where t is time and $w(t)$ a window function. In the Fourier domain, the above equation is written as

$$Y(\omega) = \Big(S(\omega)H(\omega)\Big) * W(\omega), \tag{5.2}$$

where ω is the frequency, $Y(\omega)$, $S(\omega)$, $H(\omega)$ and $W(\omega)$ are the Fourier transforms of $y(t)$, $s(t)$, $h(t)$ and $w(t)$. Letting the excitation source signal $s(t)$ be a pulse sequence with pitch period T such that

$$s(t) = \sqrt{\frac{T}{2\pi}} \sum_{n=-\infty}^{\infty} \delta(t - nT), \tag{5.3}$$

the Fourier transform of its analytic signal representation is again a pulse sequence given by

$$\begin{aligned} S(\omega) &= \sqrt{\frac{T}{2\pi}} \left[\frac{2\pi}{T} \sum_{n=0}^{\infty} \delta\Big(\omega - n\frac{2\pi}{T}\Big) \right] \\ &= \sqrt{\mu} \sum_{n=0}^{\infty} \delta(\omega - n\mu), \end{aligned} \tag{5.4}$$

where $\mu \triangleq \frac{2\pi}{T}$ is the F_0 parameter, $\delta(\cdot)$ the Dirac delta function, and n runs over the integers. Multiplying $S(\omega)$ by the vocal tract frequency response $H(\omega)$ and then taking the convolution with the frequency response $W(\omega)$ of the window function yields the complex spectrum of the short-time segment of voiced speech:

$$\begin{aligned} Y(\omega) &= \Big(S(\omega)H(\omega)\Big) * W(\omega) \\ &= \left[\sqrt{\mu} \sum_{n=0}^{\infty} H(n\mu)\delta(\omega - n\mu) \right] * W(\omega) \\ &= \sqrt{\mu} \sum_{n=0}^{\infty} H(n\mu)W(\omega - n\mu). \end{aligned} \tag{5.5}$$

We will use as a model of the speech spectrum the approximation of its power spectrum (Fig. 5.1):

$$\begin{aligned}\left\|Y(\omega)\right\|^2 &= \mu\left(\sum_{n=0}^{\infty}\left\|H(n\mu)\right\|^2\left\|W(\omega-n\mu)\right\|^2 + \sum_{n\neq n'} H^*(n'\mu)H(n\mu)W^*(\omega-n'\mu)W(\omega-n\mu)\right)\\ &\approx \mu\sum_{n=0}^{\infty}\left\|H(n\mu)\right\|^2\left\|W(\omega-n\mu)\right\|^2. \end{aligned} \tag{5.6}$$

This approximation is justified under the sparseness assumption that the power spectrum of the sum of multiple signal components is approximately equal to the sum of the power spectra generated independently from the components. The smaller the interferences between the harmonics, where the cross term $W^*(\omega-n'\mu)W(\omega-n\mu)$ such that $n\neq n'$ is sufficiently smaller than $\left\|W(\omega-n\mu)\right\|^2$, the higher the accuracy of this approximation. If we now suppose the analysis window $w(t)$ to be a Gaussian window, $|W(\omega)|^2$ can then be as well written as a Gaussian distribution function with the frequency spread σ:

$$\left\|W(\omega)\right\|^2 = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\omega^2}{2\sigma^2}\right). \tag{5.7}$$

From Eq. (5.6), one can see that with this model each frequency component power is not free but determined at once through the spectral envelope function $\|H(\omega)\|^2$, each component power being dependent on the rest of the components. As we want $\|H(\omega)\|^2$ to be a smooth and non-negative function of ω and in order to enable a prompt application to the speech synthesis method called "Composite Wavelet Model (CWM)" developed by our group [87], we introduce the following Gaussian mixture function (see Fig. 5.2):

$$\left\|H(\omega)\right\|^2 \triangleq \eta\sum_{m=1}^{M}\frac{\theta_m}{\sqrt{2\pi}\nu_m}\exp\left(-\frac{(\omega-\rho_m)^2}{2\nu_m^2}\right), \tag{5.8}$$

with

$$\sum_{m=1}^{M}\theta_m = 1. \tag{5.9}$$

The scale parameter η determines the level of the spectrum model. From Eqs. (5.6)–(5.8), the speech spectrum can now be written as:

$$\begin{aligned}\left\|Y(\omega)\right\|^2 &= \mu\sum_{n=0}^{N}\frac{\left\|H(n\mu)\right\|^2}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(\omega-n\mu^2)}{2\sigma^2}\right)\\ &= \frac{\mu\eta}{2\pi\sigma}\sum_{n=0}^{N}\left[\sum_{m=1}^{M}\frac{\theta_m}{\nu_m}\exp\left(-\frac{(n\mu-\rho_m)^2}{2\nu_m^2}\right)\right]\exp\left(-\frac{(\omega-n\mu)^2}{2\sigma^2}\right)\\ &= \sum_{n=0}^{N}\sum_{m=1}^{M}\frac{\eta\mu\theta_m}{2\pi\sigma\nu_m}\exp\left(-\frac{(\omega-n\mu)^2}{2\sigma^2}-\frac{(n\mu-\rho_m)^2}{2\nu_m^2}\right). \end{aligned} \tag{5.10}$$

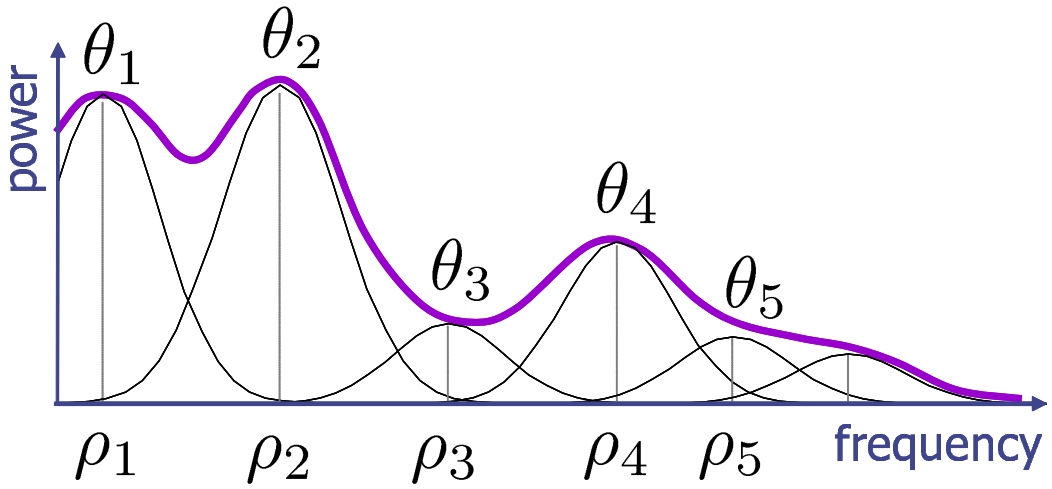


Figure 5.2 Spectral envelope model $\|H(\omega)\|^2$

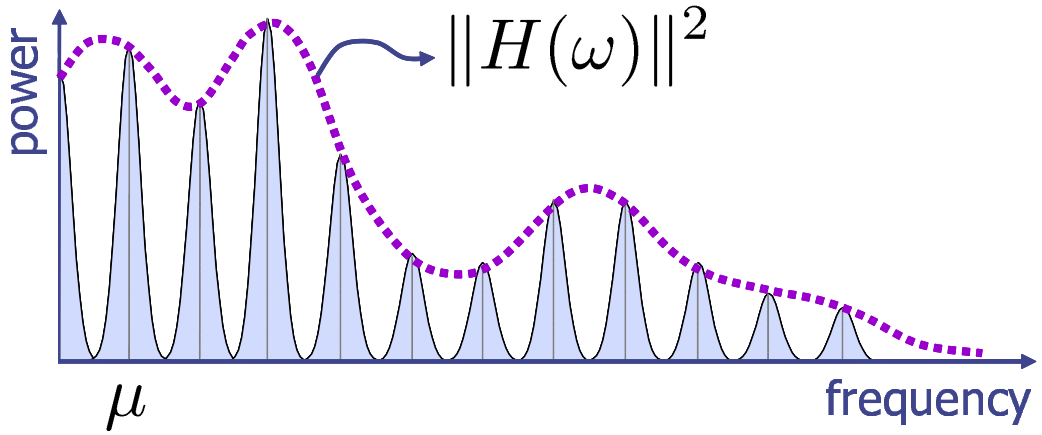


Figure 5.3 Compound model of spectral envelope and fine structure $\|Y(\omega)\|^2$

One notices from Eq. (5.10) that the spectral model we present here is a compound model of two Gaussian mixtures each of which represents the spectral envelope and the spectral fine structure (see Fig. 5.3).

So far we have only discussed voiced speech with a harmonic structure, but by making the up to now constant σ in Eq. (5.10) a free parameter, the model can also be used to approximate reasonably an unvoiced speech spectrum. White noise is indeed generally used as excitation source to synthesize unvoiced speech, but as its power spectrum is a uniform distribution, if in Eq. (5.10) σ becomes large enough such that the tails of adjacent Gaussians cover each other, the harmonic structure disappears and the model appears as a white spectrum. However, as the approximation given in Eq. (5.6) in this case becomes less accurate, a more careful modeling for unvoiced speech should be investigated in the future.

The free parameters of the model are $\mathbf{\Theta} = (\mu, \sigma, \eta, \rho_1, \cdots, \rho_M, \nu_1, \cdots, \nu_M, \theta_1, \cdots, \theta_{M-1})^T$,

and their optimal estimation from a real speech signal is the goal of the following subsection.

5.2.2 Parameter Optimization

Denoting by $F(\omega)$ the observed complex spectrum at a particular short-time segment of speech, the problem we are solving is the minimization of some distortion measure between nonnegative functions $\|Y(\omega)\|^2$ and $\|F(\omega)\|^2$. We will introduce here again as the distortion measure the I divergence of $\|Y(\omega)\|^2$ and $\|F(\omega)\|^2$:

$$J \triangleq \int_{-\infty}^{\infty} \left(\left\|F(\omega)\right\|^2 \log \frac{\left\|F(\omega)\right\|^2}{\left\|Y(\omega)\right\|^2} - \left(\left\|F(\omega)\right\|^2 - \left\|Y(\omega)\right\|^2 \right) \right) \mathrm{d}\omega, \tag{5.11}$$

which henceforth allows us to derive an elegant parameter optimization algorithm. Since the model $\|Y(\omega)\|^2$ is characterized by both the parameters for envelope and fine structures, this optimization leads to a joint estimation of F_0 and the spectral envelope.

Now as $\|Y(\omega)\|^2$ is the sum over n and m of

$$\mathcal{Y}_{n,m}(\omega) \triangleq \frac{\eta\mu\theta_m}{2\pi\sigma\nu_m} \exp\left(-\frac{(\omega - n\mu)^2}{2\sigma^2} - \frac{(n\mu - \rho_m)^2}{2\nu_m^2} \right), \tag{5.12}$$

one must deal with a nonlinear simultaneous equation in order to find the global optimal model parameters, which cannot be solved analytically. However, although any brute force gradient search algorithms are always possible, the model parameters can be efficiently estimated iteratively through the EM algorithm formulation as discussed in the following.

For any weight functions $\lambda_{n,m}(\omega)$ such that

$$\forall n, m, \omega : \ 0 < \lambda_{n,m}(\omega) < 1, \tag{5.13}$$

and

$$\forall \omega : \ \sum_n \sum_m \lambda_{n,m}(\omega) = 1, \tag{5.14}$$

one obtains the following inequation:

$$\begin{aligned}
J &= \int_{-\infty}^{\infty} \left(\left\|F(\omega)\right\|^2 \log \frac{\left\|F(\omega)\right\|^2}{\sum_n \sum_m \mathcal{Y}_{n,m}(\omega)} - \left(\left\|F(\omega)\right\|^2 - \sum_n \sum_m \mathcal{Y}_{n,m}(\omega) \right) \right) \mathrm{d}\omega \\
&= \int_{-\infty}^{\infty} \left(\left\|F(\omega)\right\|^2 \log \frac{\left\|F(\omega)\right\|^2}{\sum_n \sum_m \lambda_{n,m}(\omega) \frac{\mathcal{Y}_{n,m}(\omega)}{\lambda_{n,m}(\omega)}} - \left(\left\|F(\omega)\right\|^2 - \sum_n \sum_m \mathcal{Y}_{n,m}(\omega) \right) \right) \mathrm{d}\omega \\
&\leqq \int_{-\infty}^{\infty} \left(\left\|F(\omega)\right\|^2 \sum_n \sum_m \lambda_{n,m}(\omega) \log \frac{\lambda_{n,m}(\omega) \left\|F(\omega)\right\|^2}{\mathcal{Y}_{n,m}(\omega)} \right. \\
&\qquad \left. - \left(\left\|F(\omega)\right\|^2 - \sum_n \sum_m \mathcal{Y}_{n,m}(\omega) \right) \right) \mathrm{d}\omega,
\end{aligned} \tag{5.15}$$

using Jensen's inequality based on the concavity of the logarithm function such that:

$$\log \sum_i y_i x_i \geqq \sum_i y_i \log x_i, \tag{5.16}$$

where

$$\forall i: \ 0 < y_i < 1, \ \sum_i y_i = 1. \tag{5.17}$$

Denoting by J_λ^+ the upper bound of J, *i.e.*, the right-hand side of the inequation (5.15), equality $J_\lambda^+ = J$ holds if and only if

$$\forall n, \forall m, \forall \omega: \ \lambda_{n,m}(\omega) = \frac{\mathcal{Y}_{n,m}(\omega)}{\sum_n \sum_m \mathcal{Y}_{n,m}(\omega)}. \tag{5.18}$$

Eq. (5.18) is obtained by setting to zero the variation of the functional J_λ^+ with respect to $\lambda_{n,m}(\omega)$. By looking at J_λ^+, one can see that, if $\lambda_{n,m}(\omega)$ is fixed, the minimization of J_λ^+ w.r.t the each element Θ in $\boldsymbol{\Theta}$:

$$\widehat{\Theta} = \underset{\Theta}{\operatorname{argmin}} \, J_\lambda^+ \tag{5.19}$$

can be done analytically, which is impossible with J.

When $\lambda_{n,m}(\omega)$ is given by Eq. (5.18) with arbitrary $\boldsymbol{\Theta}$, the original objective function J is equal to J_λ^+. Then, the parameter $\boldsymbol{\Theta}$ that decreases J_λ^+ with $\lambda_{n,m}(\omega)$ fixed necessarily decreases J, since the original objective function is always guaranteed by the inequation (5.15) to be even smaller than the minimized J_λ^+. Therefore, by repeating the update of $\lambda_{n,m}(\omega)$ by Eq. (5.18) and the update of $\boldsymbol{\Theta}$ by Eq. (5.19), the objective function, bounded below, decreases monotonically and converges to a stationary point.

One notices, however, that the parameter update equation for $\boldsymbol{\Theta}$ cannot be obtained analytically because of the second term in J_λ^+:

$$-\int_{-\infty}^{\infty} \left(\left\| F(\omega) \right\|^2 - \sum_n \sum_m \mathcal{Y}_{n,m}(\omega) \right) \mathrm{d}\omega. \tag{5.20}$$

More specifically, taking the integral of $\mathcal{Y}_{n,m}(\omega)$, one obtains

$$\int_{-\infty}^{\infty} \sum_n \sum_m \mathcal{Y}_{n,m}(\omega) \mathrm{d}\omega = \sum_n \sum_m \int_{-\infty}^{\infty} \frac{\eta \mu \theta_m}{2\pi\sigma\nu_m} \exp\left(-\frac{(\omega - n\mu)^2}{2\sigma^2} - \frac{(n\mu - \rho_m)^2}{2\nu_m^2} \right) \mathrm{d}\omega \tag{5.21}$$

$$= \sum_n \sum_m \frac{\eta \mu \theta_m}{\sqrt{2\pi}\nu_m} \exp\left(-\frac{(n\mu - \rho_m)^2}{2\nu_m^2} \right), \tag{5.22}$$

from which we find that J_λ^+ is nonlinear in μ, ρ_m and ν_m. Since this term essentially amounts to the sum of the heights of the sampled points of $\|H(\omega)\|^2$ with the interval of μ, we shall

find it most convenient to approximate it with the intergral of $\|H(\omega)\|^2$. Approximating the Gaussian integral with the Riemann sums with subintervals of equal length of μ, that is,

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\nu_m} \exp\left(-\frac{(\omega-\rho_m)^2}{2\nu_m^2}\right) \mathrm{d}\omega \approx \mu \sum_n \frac{1}{\sqrt{2\pi}\nu_m} \exp\left(-\frac{(n\mu-\rho_m)^2}{2\nu_m^2}\right), \tag{5.23}$$

one obtains

$$\sum_n \frac{1}{\sqrt{2\pi}\nu_m} \exp\left(-\frac{(n\mu-\rho_m)^2}{2\nu_m^2}\right) \approx \frac{1}{\mu} \tag{5.24}$$

since the left-hand side of Eq. (5.23) is 1. Substituting Eq. (5.24) into Eq. (5.22), it is shown that

$$\int_{-\infty}^{\infty} \sum_n \sum_m \mathcal{Y}_{n,m}(\omega)\mathrm{d}\omega \approx \sum_m \frac{\eta\mu\theta_m}{\mu} = \eta. \tag{5.25}$$

Therefore, it became apparrent that the second term of the I divergence (and J_λ^+) depends very weakly on μ, ρ_m and ν_m. The update equation for the parameters except for η can thus be obtained approximately by simply minimizing

$$\int_{-\infty}^{\infty} \left\|F(\omega)\right\|^2 \sum_n \sum_m \lambda_{n,m}(\omega) \log \frac{\lambda_{n,m}(\omega)\left\|F(\omega)\right\|^2}{\mathcal{Y}_{n,m}(\omega)} \mathrm{d}\omega. \tag{5.26}$$

Now the parameter update equations obtained through Eq. (5.19) for μ, ρ_m, θ_m, σ and ν_m are derived as follows:

$$\begin{pmatrix} \mu^{(i)} \\ \rho_1^{(i)} \\ \vdots \\ \rho_M^{(i)} \end{pmatrix} = \left(\begin{array}{c|ccc} a & b_1 & \cdots & b_M \\ \hline b_1 & c_1 & & \mathbf{0} \\ \vdots & & \ddots & \\ b_M & \mathbf{0} & & c_M \end{array}\right)^{-1} \begin{pmatrix} d \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \tag{5.27}$$

$$a \triangleq \sum_{n=0}^{N} n^2 \sum_{m=1}^{M} \left(\frac{1}{\sigma^{(i-1)2}} + \frac{1}{\nu_m^{(i-1)_2}}\right) \int_{-\infty}^{\infty} \lambda_{n,m}(\omega)\left\|F(\omega)\right\|^2 \mathrm{d}\omega,$$

$$b_m \triangleq -\frac{1}{\nu_m^{(i-1)_2}} \sum_{n=0}^{N} n \int_{-\infty}^{\infty} \lambda_{n,m}(\omega)\left\|F(\omega)\right\|^2 \|\mathrm{d}\omega,$$

$$c_m \triangleq \frac{1}{\nu_m^{(i-1)_2}} \sum_{n=0}^{N} \int_{-\infty}^{\infty} \lambda_{n,m}(\omega)\left\|F(\omega)\right\|^2 \mathrm{d}\omega,$$

$$d \triangleq \frac{1}{\sigma^{(i-1)2}} \sum_{n=0}^{N} n \sum_{m=1}^{M} \int_{-\infty}^{\infty} \lambda_{n,m}(\omega)\left\|F(\omega)\right\|^2 \omega \mathrm{d}\omega,$$

$$\theta_m^{(i)} = \frac{\sum_{n=0}^{N} \int_{-\infty}^{\infty} \lambda_{n,m}(\omega) \left\| F(\omega) \right\|^2 \mathrm{d}\omega}{\sum_{n=0}^{N} \sum_{m=1}^{M} \int_{-\infty}^{\infty} \lambda_{n,m}(\omega) \left\| F(\omega) \right\|^2 \mathrm{d}\omega} \tag{5.28}$$

$$\sigma^{(i)} = \left(\frac{\sum_{n=0}^{N} \sum_{m=1}^{M} \int_{-\infty}^{\infty} \lambda_{n,m}(\omega) \left\| F(\omega) \right\|^2 \left(\omega - n\mu^{(i)}\right)^2 \mathrm{d}\omega}{\sum_{n=0}^{N} \sum_{m=1}^{M} \int_{-\infty}^{\infty} \lambda_{n,m}(\omega) \left\| F(\omega) \right|^2 \mathrm{d}\omega} \right)^{1/2} \tag{5.29}$$

$$\nu_m^{(i)} = \left(\frac{\sum_{n=0}^{N} \left(n\mu^{(i)} - \rho_m^{(i)}\right)^2 \int_{-\infty}^{\infty} \lambda_{n,m}(\omega) \left\| F(\omega) \right\|^2 \mathrm{d}\omega}{\sum_{n=0}^{N} \int_{-\infty}^{\infty} \lambda_{n,m}(\omega) \left\| F(\omega) \right\|^2 \mathrm{d}\omega} \right)^{1/2} \tag{5.30}$$

$$\eta^{(i)} = \frac{\sqrt{2\pi} \int_{-\infty}^{\infty} \left\| F(\omega) \right\|^2 \mathrm{d}\omega}{\sum_{n=0}^{N} \sum_{m=1}^{M} \frac{\theta_m^{(i)}}{\nu_m^{(i)}} \exp\left(- \frac{(n\mu^{(i)} - \rho_m^{(i)})^2}{2\nu_m^{(i)2}} \right)}. \tag{5.31}$$

where the superscript i refers to the iteration cycle. Some examples of the estimated envelope $\|H(\omega)\|^2$ with $M = 15$ can be seen in Fig. 5.4.

5.3 Experimental Evaluations

5.3.1 Single Voice F_0 Determination

To confirm its performance as a F_0 extractor, we tested our method on 10 Japanese speech data of male ('`myi`') and female ('`fym`') speakers from the ATR speech database and chose the well-known F_0 extractor "YIN"[25] for comparison. All power spectra were computed with a sampling rate of 16kHz, a frame length of 32ms and a frame shift of 10ms. The spectral model was made using $N+1 = 60$ Gaussians, and the envelope model was made using $M = 15$ Gaussians. The number of free parameters is thus $3 + 15 \times 3 = 48$. The initial values of μ were set to 47Hz, 94Hz and 141Hz, respectively, and among these conditions, the converged parameter set that gave the minimum of J was considered as the global optimum. The initial values of θ_m were determined uniformly, and σ and ν_m were initialized to 31Hz and 313Hz, respectively. For an F_0 estimation task, we defined two error criteria: deviations over

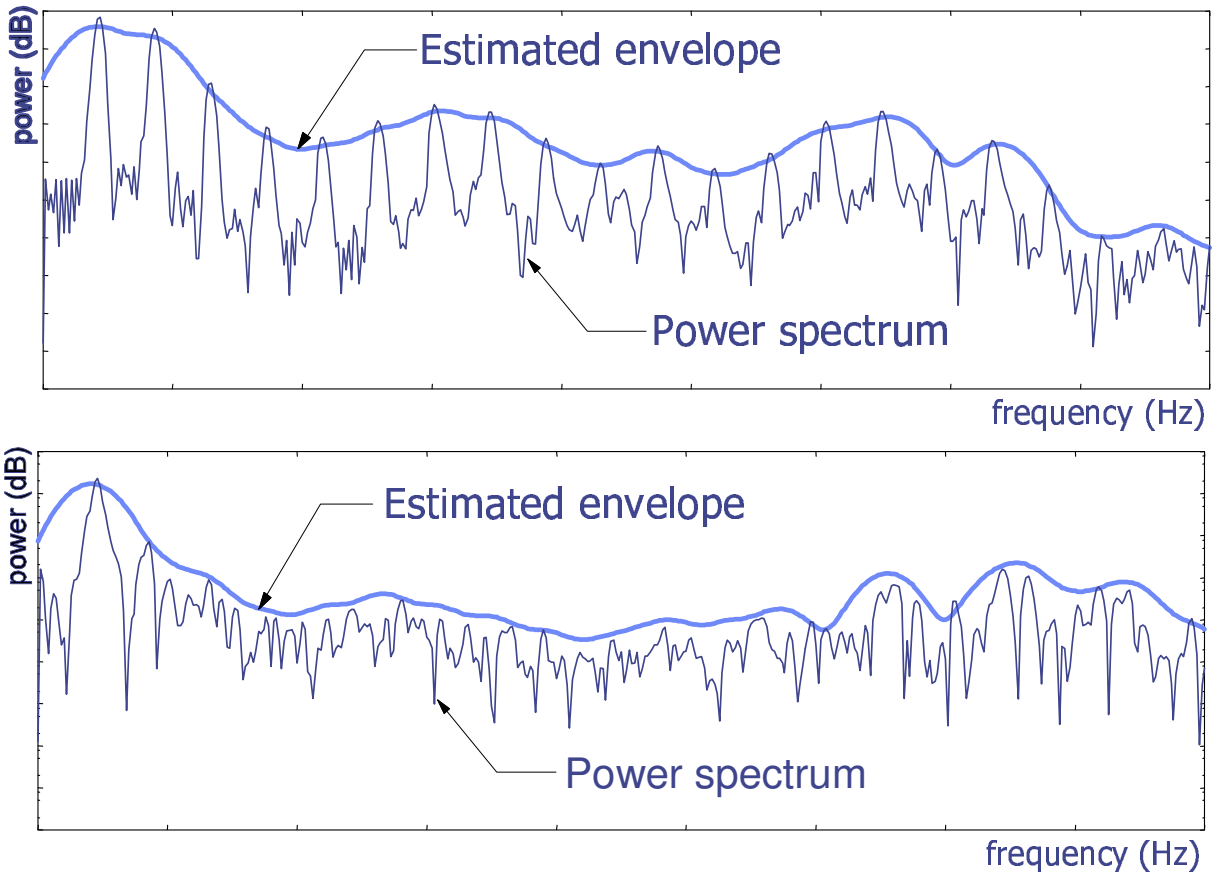


Figure 5.4 *Observed power spectra of voiced (top) and unvoiced (bottom) speech and the corresponding spectral envelope estimates.*

5% and 20% from the hand-labeled F_0 reference as fine and gross errors, respectively. The former criterion shows how precisely the proposed analyzer is able to estimate F_0 and the latter shows the robustness against the double/half pitch errors. The areas where reference F_0s are given by zero were not considered in the computation of the accuracy. As a second evaluation, we took the average of the cosine measures between $\|Y(\omega)\|^2$ and $\|F(\omega)\|^2$ on the whole analysis interval to verify how well the choices of the distortion measure to minimize and of the model for expressing actual speech power spectra are. These results can be seen in Table 5.1. The numbers in the brackets in Table 5.1 are the results obtained with YIN. The source code was kindly provided to us by its authors. One can verify from the results that our method is as accurate as YIN when it comes to roughly estimate F_0 and significantly outperforms YIN for precise estimation. Thus, our method would be especially useful for situations in which a highly precise F_0 estimate is required, which is exactly the case in the spectral envelope estimation algorithms that use F_0 estimates. We should note

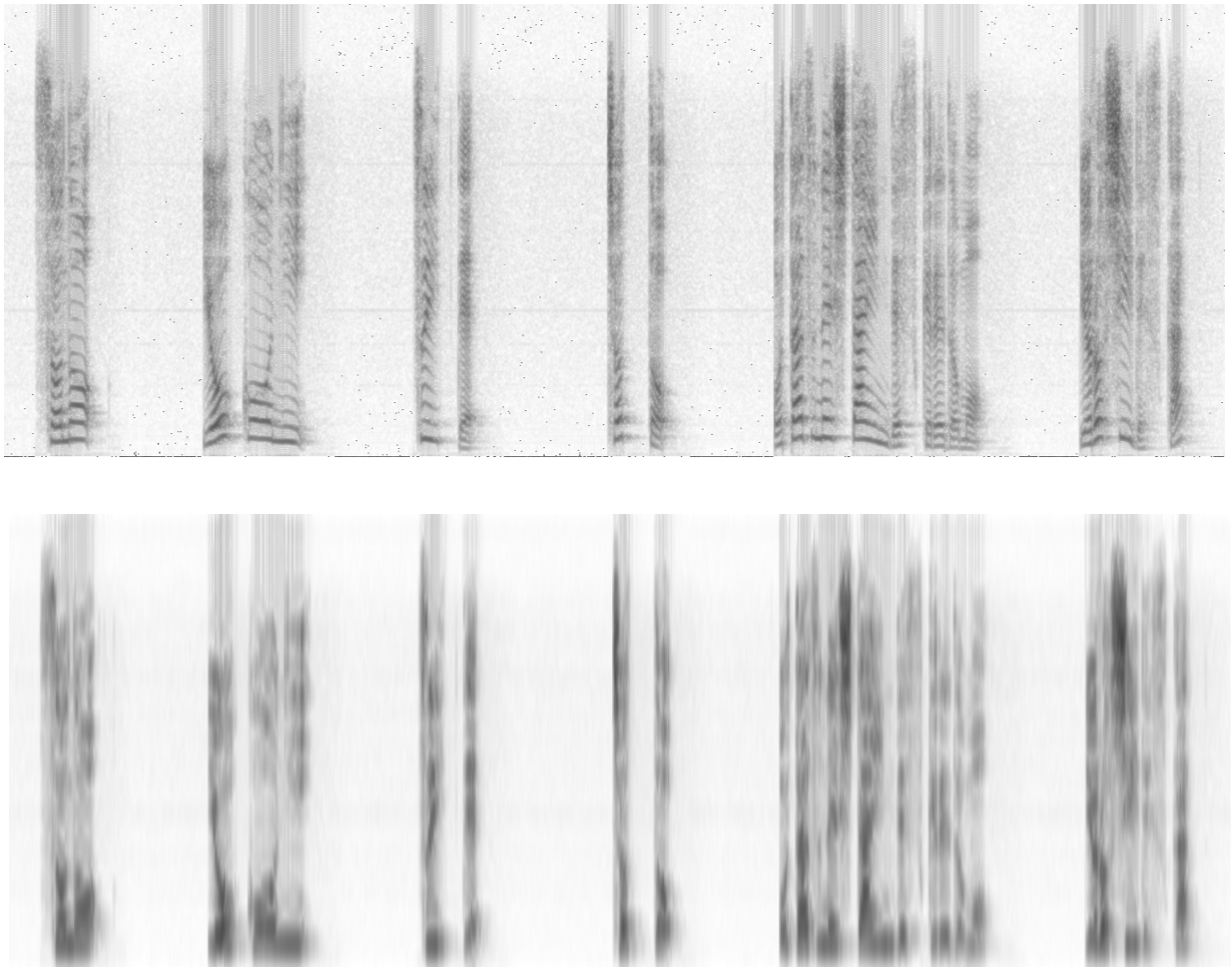


Figure 5.5 *A spectrogram of a female speech (top) and a gray-scale display of the spectral envelope estimates (bottom).*

however that the parameters used for YIN may not do it full justice. The results seem to be rather good for a frame-by-frame algorithm, which encourages us to embed this envelope structured model into the parametric spectrogram model proposed in [**?**, **?**] to exploit the temporal connectivity of speech attributes for a further improvement.

5.3.2 Synthesis and Analysis

We evaluate here the accuracies of spectral envelope estimation. To do so, we need to use speech signals whose true spectral envelope is known in advance as the experimental data. For this we purpose, we created several synthetic speech signals. The synthetic signals were made using three types of linear filter: all-zero filter, all-pole filter and pole-zero filter, and

Table 5.1 *Accuracies of F_0 determination*

Speech File	F_0 accuracy (%)		Cosine (%)
	±5%	±20%	
myisda01	98.4 (85.3)	98.6 (98.6)	96.7
myisda02	93.3 (82.6)	97.8 (97.8)	98.0
myisda03	94.2 (79.9)	97.5 (96.9)	96.0
myisda04	98.0 (86.3)	99.0 (95.1)	96.8
myisda05	93.7 (71.7)	97.8 (96.1)	95.9
fymsda01	97.2 (87.0)	98.0 (98.0)	98.3
fymsda02	96.8 (88.5)	98.1 (98.1)	97.6
fymsda03	95.4 (84.6)	98.5 (98.5)	98.2
fymsda04	97.0 (88.2)	98.1 (98.1)	98.2
fymsda05	95.7 (86.5)	99.2 (98.5)	98.1

the input excitation. The input excitation we used here is a linear chirped single pulse signal, whose F_0 modulates linearly from 100Hz to 400Hz within 2 seconds. The characteristics of the filters were chosen as follows:

All-zero (1):

$$\widetilde{H}(z) = 1 - 0.1z^{-1} + 0.3z^{-2} + 0.4z^{-3} + 0.2z^{-4} + 0.4z^{-5} + 0.2z^{-6} + 0.1z^{-7} + 0.5z^{-8},$$

All-zero (2):

$$\widetilde{H}(z) = -1.2 + 0.1z^{-1} + 0.3z^{-2} + 0.1z^{-3} + 0.2z^{-4} + 0.4z^{-5} - 0.2z^{-6} + 0.8z^{-7} + 1.2z^{-8},$$

All-pole:

$$\widetilde{H}(z) = \frac{1}{1 - 0.5z^{-1} + 0.4z^{-2} - 0.1z^{-3} + 0.3z^{-4} - 0.3z^{-5}},$$

Pole-zero:

$$\widetilde{H}(z) = \frac{-1.2 + 0.1z^{-1} + 0.3z^{-2} + 0.1z^{-3} + 0.2z^{-4} + 0.4z^{-5} - 0.2z^{-6} + 0.8z^{-7} + 1.2z^{-8}}{1 - 0.5z^{-1} + 0.4z^{-2} - 0.1z^{-3} + 0.3z^{-4} - 0.3z^{-5}}.$$

We chose as the measure to assess the accuracy of the spectral envelope estimation the "*Spectral Distortion* (SD)", defined by

$$\frac{1}{I}\sum_{i=1}^{I}\Big(\log\big\|H(\omega_i)\big\| - \log\big\|\widetilde{H}(e^{j\omega_i})\big\|\Big)^2, \tag{5.32}$$

where i refers to the index of the frequency-bin, $\|\widetilde{H}(e^{j\omega_i})\|$ the true (reference) spectral envelope and $\|H(\omega_i)\|$ the spectral envelope estimate.

The experimental results are shown in Fig. 5.6. Fig. 5.6 (a), (b), (c) and (d) are the results when testing with the data created respectively by *all-zero (1)*, *all-zero (2)*, *all-pole* and *pole-zero*. Each graph shows the transitions of SD values within two seconds during which the F_0 of the input excitation modulates from 100Hz to 400Hz. One sees from these graphs that as the F_0 of the input gets higher, conventional methods such as 40-order LPC and LPC cepstrum tend to obtain poorer results. This is perhaps because the envelope estimates descend down into the space between the partials for high F_0. The accuracies of the envelope estimates obtained by the proposed method does not seem to become poor even in high F_0. This is obviously because the proposed method tries to estimate the spectral fine structure at the same time. On the other hand, the 14-order LPC envelope is too smooth to make a good fit to the true envelope.

5.3.3 Analysis and Synthesis

We compared through a psychological experiment the processing capacity and the intelligibility of the synthesized speech restored from the parameters obtained via the proposed and LPC analyzers. The parameters extracted via the proposed analyzer were transformed to a synthesized speech using the *CWM method [87]. As a test set, we used speech data of 5 vowels (/a/, /i/, /u/, /e/, /o/) and 40 randomly chosen words uttered by a female speaker excerpted from the same database. Analyses were done with a sampling rate of 16kHz, a frame shift of 10ms and a frame length of 32ms for the proposed method and 30 ms for the

*CWM synthesizes speech by spacing composite Gabor functions, transformed from a Gaussian mixture envelope, by a pitch period interval.

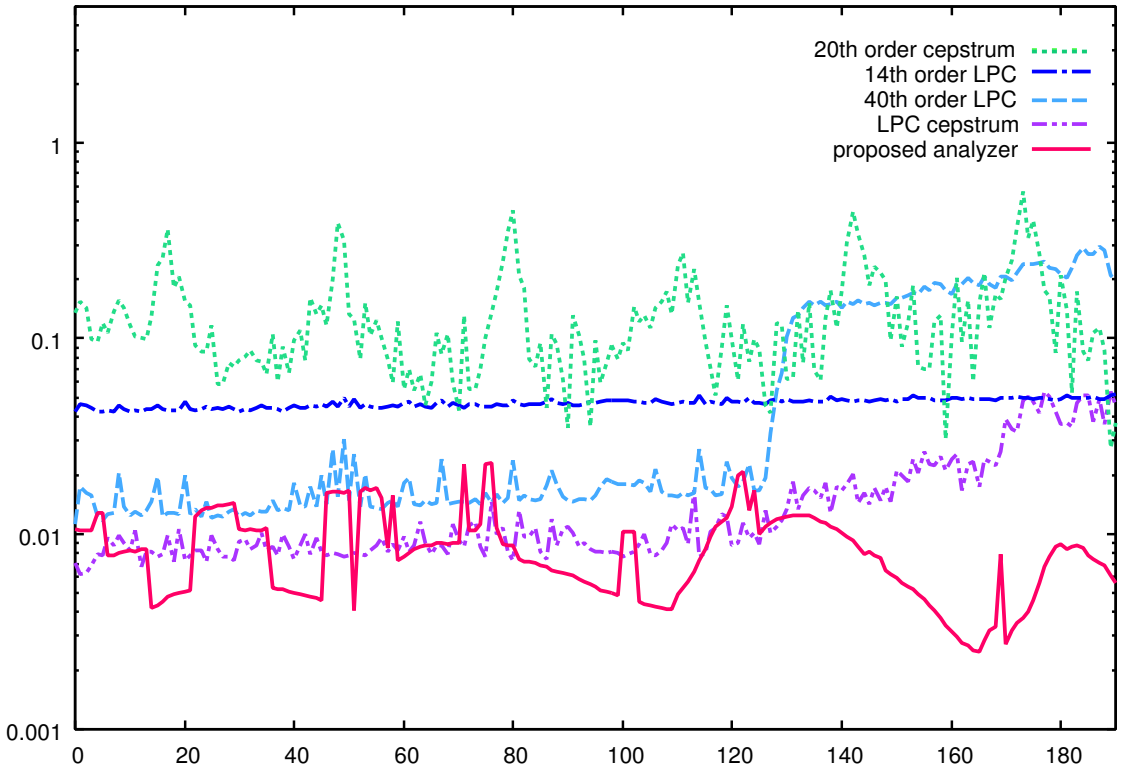
Table 5.2 *Preference score(%) of the synthesized speech generated by CWM[87] using the parameter estimates of the proposed model.*

listener	vowel	word
A	60	84
B	60	83
C	40	68
D	80	80
E	60	95
F	80	96
G	100	100
H	40	64
I	80	94
J	60	88
Ave.	66	83

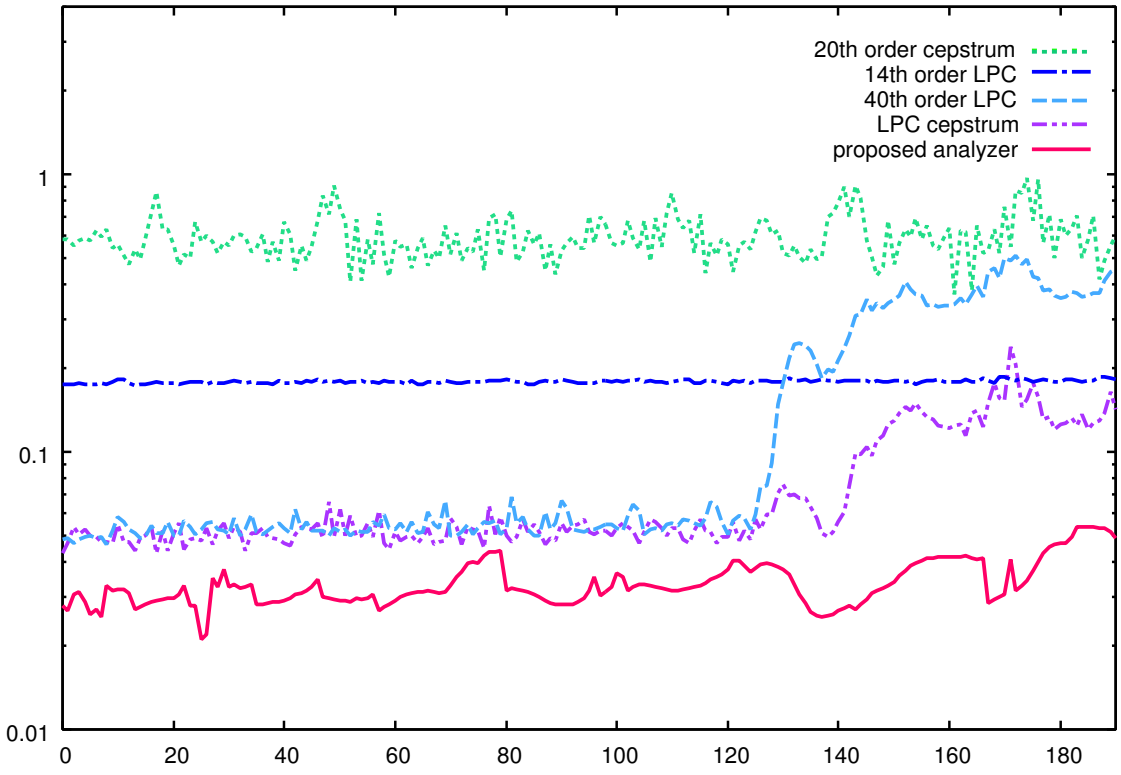
LPC. The dimension of the parameters for the proposed model and the LPC's were both set to 45. For the LPC analysis, the F_0s were extracted via the supplementary F_0 extraction tool included in the Snack Sound Toolkit. Each synthesized speech used for the evaluation was excited with an estimated vocal tract characteristic by a pulse sequence at intervals of a different pitch period from the original one. The pitch periods were modified to 80% and 120% of the pitch periods obtained from the original speech. We let 10 listeners choose the one they thought was more intelligible and obtained a preference score of the results via the proposed analyzer. The preference score, shown in Table 5.2, shows that the processing capacity and the intelligibility of the synthesized speech generated through the proposed analyzer are higher than that from through LPC analyzer.

5.4 Summary of Chapter 5

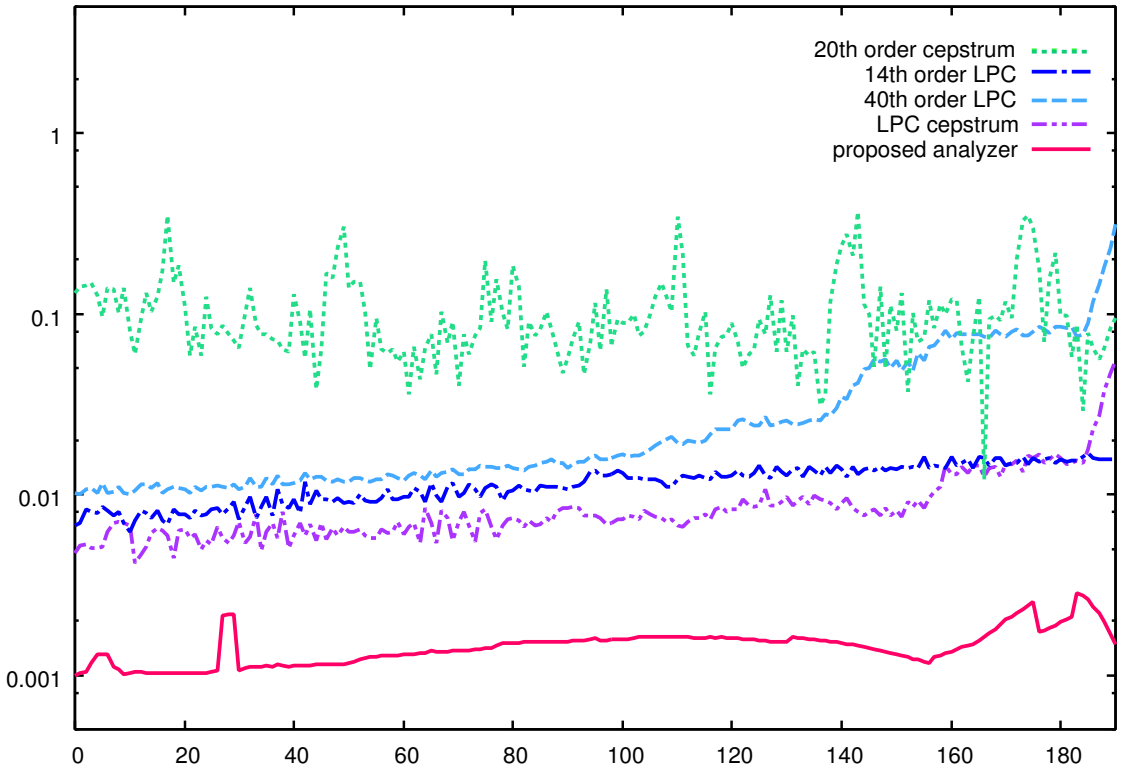
In this chapter, we formulated the estimation of F_0 and the spectral envelope as a joint optimization of a composite function model of the spectral envelope and the fine structure, and confirmed through experiments the effectiveness of this method. Encouraged by the results, we are planning to apply this idea to Harmonic-Temporal Clustering in the future.



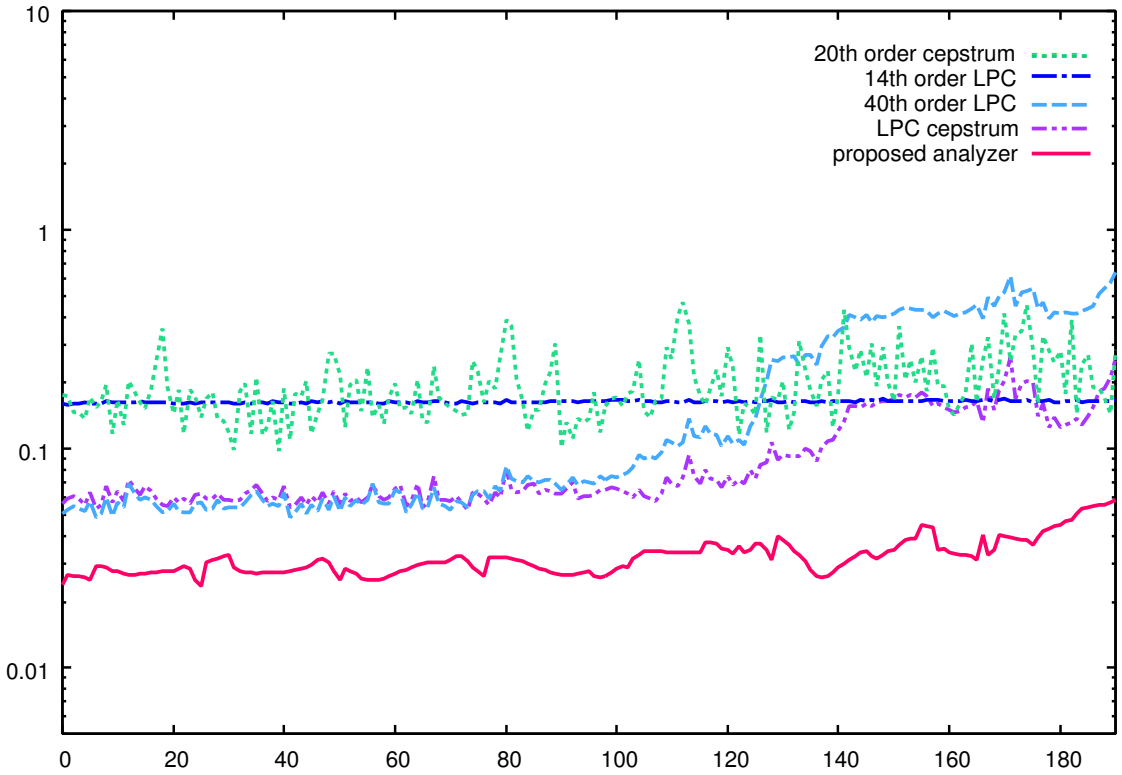
(a) *All-zero (1)*



(b) *All-zero (2)*



(c) *All-pole*



(d) *Pole-zero*

Figure 5.6 *Comparison of the accuracies of spectral envelope estimation between the proposed method and the conventional methods. Each graph shows the transitions of SD values during two seconds.*

Chapter 6

Parameter Optimization of Sinusoidal Signal Model

6.1 Introduction

The approaches of the preceding chapters are based on the approximate assumption of additivity of the power spectra (neglecting the terms corresponding to interferences between frequency components), but it becomes usually difficult to infer F_0s when two voices are mixed with close F_0s as far as we are only looking at the power spectrum. In this case not only the harmonic structure but also the phase difference of each signal becomes an important cue for separation. Moreover, having in mind future source separation methods designed for multi-channel signals of multiple sensory input, analysis methods in the complex spectrum domain taking into account the phase estimation are indispensable.

After McAulay et al. [71] showed that the sinusoidal signal model could be applied to Analysis-by-Synthesis systems to obtain high-quality synthesized speech, the range of application of this model has widened to Text-To-Speech synthesis, speech modification, coding, etc. In particular, as the possibility to generate high-quality synthesized speech shows that the sinusoidal signal model represents extremely well acoustic signals such as speech and music, we can have high expectations for its application to source separation.

Independently of the situation of application, the common point of this framework (signal analysis using sinusoidal signal model) is that the most important problem resides in how to accurately estimate the parameters of the sinusoidal signal model, and this estimation accuracy is directly related to the performance of every application. The sinusoidal signal model used by McAulay et al. is the superposition of K complex sinusoids which are assumed

to have constant frequency and amplitude:

$$s(t) \triangleq \sum_{k=1}^{K} A_k e^{j\mu_k t}, \quad t \in (-\infty, \infty), \tag{6.1}$$

where μ_k and A_k represent respectively the frequency and complex amplitude of the k-th sinusoidal component. In addition, the arguments $\arg(A_k)$ represent the phases at time $t = 0$ (initial phase). If we denote the target analytic signal on the short-time analysis interval $t \in [-T, T]$ by $\widetilde{y}(t)$, and if we assume that it can be expressed as

$$\widetilde{y}(t) = s(t) + \epsilon(t), \quad t \in [-T, T], \tag{6.2}$$

where $\epsilon(t)$ is a Gaussian white noise $\epsilon(t) \sim \mathcal{N}(\mathbf{0}, \Sigma)$ with $\Sigma = \nu^2 \boldsymbol{I}$, then the problem is to obtain the maximal likelihood parameters $\boldsymbol{\Theta} = \{\mu_k, a_k, \varphi_k\}_{k=1}^{K}$. In this case, as $\epsilon(t) \sim \mathcal{N}(\mathbf{0}, \Sigma)$, the log-likelihood of $\boldsymbol{\Theta}$ can be written

$$\int_{-T}^{T} \left(\log \frac{1}{\sqrt{2\pi}\nu} - \frac{1}{2\nu^2} \left\| \widetilde{y}(t) - s(t) \right\|^2 \right) \mathrm{d}t, \tag{6.3}$$

and finally the solution of the minimization of the L^2 norm error corresponds to the maximal likelihood parameter:

$$\widehat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \int_{-T}^{T} \left\| \widetilde{y}(t) - s(t) \right\|^2 \mathrm{d}t \tag{6.4}$$

$$= \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \int_{-\infty}^{\infty} \left\| w(t) \Big(\widetilde{y}(t) - s(t) \Big) \right\|^2 \mathrm{d}t, \tag{6.5}$$

where $w(t)$ is the rectangular window

$$w(t) = \begin{cases} 1 & |t| \leqq T \\ 0 & |t| > T \end{cases}. \tag{6.6}$$

As shown in Eq. (6.1), the sinusoidal signal model depends linearly on A_k, but non-linearly on μ_k, and thus it is straightforward to analytically obtain the maximum likelihood solution for A_k when μ_k is fixed, but even when A_k is fixed the maximum likelihood solution for μ_k cannot be obtained analytically. This point is the essence of the difficulty of the parameter optimization of the sinusoidal signal model, and methods to obtain the maximum likelihood solution for μ_k have been the subject of intensive research for many years in the area of statistical signal processing [71, 84, 98, 34, 118, 6, 17, 5, 99, 22, 57, 43].

In McAulay et al. [71], in order to obtain the estimation of the parameters $\boldsymbol{\Theta} = \{\mu_k, A_k\}_{k=1}^{K}$, a simple method is used which consists in repeating K times the operation of determining

the frequency, amplitude and phase of the peak element maximizing the discrete power spectrum density (periodogram) of the target signal and of subtracting this peak element from the signal. The fact pointed out by Rife et al. [83, 81, 82] that the frequency giving the maximum value of the periodogram of a single sinusoid is a maximum likelihood estimator and that this estimator is an unbiased estimator is one argument for the validity of the above frequency estimation method as an approximate solution of Eq. (6.5). Considering that (1) the peaks of the discrete periodogram do not necessarily correspond to the maximal values of the original continuous periodogram, (2) when there are several frequency components the above theory does not stand anymore because of the interferences between frequency components, (3) when several frequency components are close to each other it happens that the detection of each peak can not be done correctly because of the energy dispersion, it is natural to hope for the development of an estimation method with a higher efficiency than the above simple frequency estimation method can be expected. In such a perspective, methods [98, 34, 118, 6, 17, 5] trying to obtain a more efficient parameter estimation by not directly considering the peak frequency as an estimation value but by looking for the maximal point of a curve interpolating several points in the neighborhood of the peak have been used particularly often recently for their simplicity. However, these methods still do not solve the problems (2) and (3) mentioned above, and as they only give, similarly to McAulay et al.'s method, an approximate solution of (6.5), zero-padding and window function design methods to increase the accuracy of this approximation are the main object of their discussions [17, 5]. Meanwhile, non-linear optimization methods such as gradient search methods (e.g., steepest descent or Newton's method), and methods based on statistical sampling (Gibbs sampler or Markov chain Monte-Carlo (MCMC) method) are also proposed to search numerically for the solution of Eq. (6.5) [1, 99, 22, 57, 43].

While the method of McAulay et al. is a mixture of K pure tone signals, one can also consider in the same way the case of an analytic signal which is the superposition of K harmonic signals (signal composed of N harmonic components, where the n-th harmonic component's frequency is n times the fundamental frequency μ_k):

$$s(t) \triangleq \sum_{k=1}^{K} \sum_{n=1}^{N} A_{k,n} e^{jn\mu_k t}, \quad t \in (-\infty, \infty). \tag{6.7}$$

This model is often used for 1ch source separation when the target mixed signal is only composed of harmonic signals [22, 57, 43]. Eq. (6.7) with $N = 1$ corresponds to assuming the same model as Eq. (6.1), and McAulay et al.'s model is thus a special case of Eq. (6.7).

However, as in this model each sinusoid's frequency does not have its own degree of freedom but is constrained to be a multiple $n\mu_k$ of the fundamental frequency, obtaining the maximum likelihood solution for μ_k becomes even harder than in McAulay et al's model. For example, some methods from the first type presented above try to estimate μ_k based on peak extraction, but it then becomes necessary to rely on ad hoc threshold setting to determine to which harmonic component of which harmonic signal the extracted peak belongs, and the discussion on the obtained maximum likelihood solution for μ_k becomes complicated. For that reason, the source separation approaches which used this model are most often from the second type (gradient search and sampling methods)[22, 57, 43]. However, this kind of numerical computation is often beset with local optimum problems. A global minimum for Eq. (6.5) is not guaranteed to be obtained unless, in the case of the gradient search method the iterative computation is led to convergence for an infinity of initial points, or in the case of the stochastic sampling an infinite number of trial is performed. For that reason, the problem is to know if the search for the solution can be performed with a low computation cost (the lower the computation cost, the more searches can be performed from different initial parameter conditions), but as for now only brute force numerical computations such gradient search method and sampling method have been proposed.

As explained above, albeit the sinusoidal signal model represents extremely well acoustic signals such as speech and music, room was left for discussion on how to estimate its parameters. Against this background, the goal of this chapter to is derive a new optimization algorithm to obtain the maximum likelihood parameter of the sinusoidal signal model.

6.2 Abstract and Organization of Chapter 6

The parameter optimization algorithm for sinusoidal signal model, proposed in this section, is based on a principle of the iterative method that uses an auxiliary function. This principle was inspired by the essential idea of EM algorithm. Let $\Phi(\mathbf{\Theta})$ be the objective function one wants to minimize with respect to its parameters $\mathbf{\Theta} = (\Theta_1, \cdots, \Theta_I)$, and define by $\Phi^+(\mathbf{\Theta}, m)$ the *auxiliary function* of $\Phi(\mathbf{\Theta})$, and $m = (m_1, \cdots, m_J)$ the *auxiliary parameters* if $\Phi^+(\mathbf{\Theta}, m)$ satisfies

$$\Phi\big(\mathbf{\Theta}\big) \leqq \Phi^+\big(\mathbf{\Theta}, m\big). \tag{6.8}$$

$\Phi(\mathbf{\Theta})$ can then be decreased monotonically by the iteration consisting of the two steps: minimization of the auxiliary function with respect to the auxiliary parameters m, and as

well with respect to the parameters $\boldsymbol{\Theta}$.

In the next section, we introduce the sinusoidal signal model and the objective function that we will deal with through this chapter. We show the principle of the auxiliary function method in Section 6.4, and derive the auxiliary function from the Feder's lemma in Subsection 6.4.2. As one sees that it is still impossible to obtain analytically the update equation for the F_0 parameter μ_k, we show in Subsection 6.4.3 that one can derive a further auxiliary function by introducing a theorem for concave functions. This auxiliary function enables us to derive analytically the update equation for μ_k as will be mentioned in Subsection 6.4.4.

6.3 Problem Setting

6.3.1 Pseudoperiodic Signal Model

Consider as the time-varying acoustic signal the sum of pseudoperiodic signal models given in an analytic signal representation by

$$s(t) = \sum_{k=1}^{K} \sum_{n=1}^{N} A_{k,n}(t) e^{jn\theta_k(t)}, \quad t \in (-\infty, \infty), \tag{6.9}$$

where the instantaneous phase $\theta_k(t)$ of the fundamental component, and the instantaneous complex amplitude $A_{k,n}(t)$ of the n^{th} are the unknown parameters. $\mu_k(t) = \dot{\theta}_k(t)$ amounts to the instantaneous F_0 and $a_{k,n}(t) = |A_{k,n}(t)|$ the instantaneous amplitude, which are both assumed here to change gradually over time. These are the free parameters that one wants to estimate, which we denote for convenience by $\boldsymbol{\Theta}$:

$$\boldsymbol{\Theta} = \left\{ \theta_k(t), \left\{ A_{k,n}(t) \right\}_{1 \le n \le N} \right\}_{1 \le k \le K}. \tag{6.10}$$

Now letting $y(t)$ be the observed signal of interest, we assume the following model:

$$y(t) = s(t) + n(t), \quad t \in (-\infty, \infty), \tag{6.11}$$

where $n(t)$ is a Gaussian white noise. The maximum likelihood estimate of $\boldsymbol{\Theta}$ can thus be obtained by minimizing the L^2 norm of the error signal in $t \in (-\infty, \infty)$:

$$\int_{-\infty}^{\infty} \left\| y(t) - s(t) \right\|^2 \mathrm{d}t. \tag{6.12}$$

We now show that this time domain objective can be equivalently defined in the time-frequency domain. As short-time Fourier transform (STFT) is one of the most popular ways of time-frequency decomposition, we show the following lemma, which gives us the objective function in the time-frequency domain by the Gabor transform (STFT).

6.3.2 Objective Function Defined on Gabor Transform Domain

Lemma 1 (L^2 norm in STFT domain). *The time-frequency components of $y(t)$ and $s(t)$ by Gabor transform is by definition given by*

$$G_y(\omega,t) \triangleq \Big\langle y(u), \psi_{\omega,t}(u) \Big\rangle_{u\in\mathbb{R}}, \tag{6.13}$$

$$G_s(\omega,t) \triangleq \Big\langle s(u), \psi_{\omega,t}(u) \Big\rangle_{u\in\mathbb{R}}, \tag{6.14}$$

where $\psi_{\omega,t}(u)$ is the Gabor function, which is a nonorthogonal basis used to measure the component of frequency ω at time t, and defined as the product of the complex sinusoid with frequency of ω and the Gaussian window centered at time t:

$$\psi_{\omega,t}(u) = e^{-d(u-t)^2+j\omega(u-t)}, \tag{6.15}$$

where d is the time spread parameter of the Gaussian window, that can be chosen arbitrarily. Though trivial, we then have

$$\int_{-\infty}^{\infty} \Big\| y(t) - s(t) \Big\|^2 dt = \eta \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Big\| G_y(\omega,t) - G_s(\omega,t) \Big\|^2 \mathrm{d}\omega\mathrm{d}t, \tag{6.16}$$

where η is a constant that depends neither on ω nor on t.

Proof. By definition, $G_y(\omega,t)$ can be written as

$$\begin{aligned} G_y(\omega,t) &= \Big\langle y(u), \psi_{\omega,t}(u) \Big\rangle_{u\in\mathbb{R}} && (6.17)\\ &= \int_{-\infty}^{\infty} y(u)\psi^*_{\omega,t}(u)\mathrm{d}u && (6.18)\\ &= \int_{-\infty}^{\infty} y(u)e^{-d(u-t)^2}e^{-j\omega(u-t)}\mathrm{d}u && (6.19)\\ &= \int_{-\infty}^{\infty} y(u)e^{-d(u-t)^2+j\omega t}e^{-j\omega u}\mathrm{d}u && (6.20)\\ &= e^{j\omega t}\mathscr{F}\Big[y(u)e^{-d(u-t)^2}\Big]_u. && (6.21) \end{aligned}$$

G_y can as well be written as $G_y(\omega,t) = e^{j\omega t}\mathscr{F}\big[s(u)e^{-d(u-t)^2}\big]_u$. Therefore,

$$\begin{aligned} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Big\| G_y(\omega,t) - G_s(\omega,t) \Big\|^2 \mathrm{d}\omega\mathrm{d}t &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Big\| e^{j\omega t}\mathscr{F}\Big[\Big(y(u)-s(u)\Big)e^{-d(u-t)^2}\Big]_u \Big\|^2 \mathrm{d}\omega\mathrm{d}t\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Big\| \mathscr{F}\Big[\Big(y(u)-s(u)\Big)e^{-d(u-t)^2}\Big]_u \Big\|^2 \mathrm{d}\omega\mathrm{d}t\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Big\| \Big(y(u)-s(u)\Big)e^{-d(u-t)^2} \Big\|^2 \mathrm{d}u\mathrm{d}t\\ &= \int_{-\infty}^{\infty} \Big\| y(u)-s(u) \Big\|^2 \int_{-\infty}^{\infty} e^{-2d(t-u)^2}\mathrm{d}t\mathrm{d}u. \end{aligned} \tag{6.22}$$

The equality in the third line follows from the Parseval's theorem. Using the result of the Gaussian integral, one obtains

$$\int_{-\infty}^{\infty} e^{-2d(t-u)^2} \mathrm{d}t = \sqrt{\frac{\pi}{2d}}, \tag{6.23}$$

which immediately proves that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Big\| G_y(\omega,t) - G_s(\omega,t) \Big\|^2 \mathrm{d}\omega\mathrm{d}t = \sqrt{\frac{\pi}{2d}} \int_{-\infty}^{\infty} \Big\| y(u) - s(u) \Big\|^2 \mathrm{d}u. \tag{6.24}$$

□

One sees from this result that minimization of Eq. (6.12) is equivalent to minimizing

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Big\| G_y(\omega,t) - G_s(\omega,t) \Big\|^2 \mathrm{d}\omega\mathrm{d}t. \tag{6.25}$$

Recall that $G_s(\omega,t)$ is the Gabor transform of $s(t)$, such that, from Eq. (6.9),

$$G_s(\omega,t) = \int_{-\infty}^{\infty} \sum_{k=1}^{K}\sum_{n=1}^{N} A_{k,n}(u) e^{jn\phi_k(u)} e^{-d(u-t)^2 - j\omega(u-t)} \mathrm{d}u. \tag{6.26}$$

As the dominant part of the Gabor function $e^{-d(u-t)^2-j\omega(u-t)}$ is localized only around time t, the result of the integral in Eq. (6.26) depends heavily on the portion of $\theta_k(u)$ and $A_{k,n}(u)$ near t. Recalling that we have assumed that the instantaneous phase $\theta_k(u)$ and the instantaneous complex amplitude $A_{k,n}(t)$ change gradually over time, approximating $\theta_k(u)$ and $A_{k,n}(u)$ by zero and first order Taylor series expansions around time t:

$$\theta_k(u) \approx \theta_k(t) + \mu_k(t)(u-t) \tag{6.27}$$

$$A_{k,n}(u) \approx A_{k,n}(t) \tag{6.28}$$

may not affect significantly the result of Eq. (6.26). $\mu_k(t) \triangleq \dot{\theta}_k(t)$ is the instantaneous F_0. $G_s(\omega,t)$ can then be written as

$$G_s(\omega,t) = \sum_{k=1}^{K}\sum_{n=1}^{N} A_{k,n}(t) e^{-\frac{(\omega - n\mu_k(t))^2}{4d}}, \tag{6.29}$$

where $A_{k,n}(t) = \widetilde{A}_{k,n}(t) e^{jn\theta_k(t)}/\sqrt{2d}$.

In the case of discrete-time observations, we shall consider as the problem of interest the minimization of

$$\int_{-\infty}^{\infty} \Big\| G_y(\omega,t) - G_s(\omega,t) \Big\|^2 \mathrm{d}\omega, \tag{6.30}$$

with respect to

$$\mathbf{\Theta}_t = \left\{\theta_k, \left\{A_{k,n}\right\}_{1\le n\le N}\right\}_{1\le k\le K} \tag{6.31}$$

at each discrete time point. The problem can thus be summarized as follows.

We model the acoustic signal by a stationary sinusoidal model

$$s(t) = \sum_{k=1}^{K}\sum_{n=1}^{N} \widetilde{A}_{k,n} e^{jn\mu_k t}, \quad t \in (-\infty, \infty), \tag{6.32}$$

where the F_0 μ_k, and the complex amplitude $A_{k,n}$ of the n^{th} partial are the unknown parameters. These are the free parameters corresponding to the instantaneous features at $t = 0$, that one wants to estimate, which we denote for convenience by $\mathbf{\Theta}$:

$$\mathbf{\Theta} = \left\{\mu_k, \left\{\widetilde{A}_{k,n}\right\}_{1\le n\le N}\right\}_{1\le k\le K}. \tag{6.33}$$

Letting $y(t)$ be the observed signal of interest, the problem we are solving is to estimate the instantaneous feature $\mathbf{\Theta}$ in $y(t)$ near $t = 0$. This can be achieved by finding $\mathbf{\Theta}$ that minimizes

$$\Phi(\mathbf{\Theta}) = \int_{-\infty}^{\infty} \left\| Y(\omega) - \sum_{k=1}^{K}\sum_{n=1}^{N} A_{k,n} e^{-\frac{(\omega - n\mu_k)^2}{4d}} \right\|^2 \mathrm{d}\omega, \tag{6.34}$$

where $A_{k,n} = \frac{\widetilde{A}_{k,n}}{\sqrt{2d}}$, d is the time spread parameter of the Gaussian window, and $Y(\omega)$ the simplified notation of $G_y(\omega, 0)$ (we shall emphasize that it is not meant to be the Fourier transform of $y(t)$). From the next section, we will derive the parameter optimization algorithm that finds the maximum likelihood estimate of $\mathbf{\Theta}$.

6.4 Parameter Optimization Algorithm

6.4.1 Auxiliary Function Method

The parameter optimization algorithm we propose in this chapter is based on a principle called the *auxiliary function method*, which was inspired by the idea of the EM algorithm. We first define the auxiliary function and then show the lemma for the iterative algorithm, the auxiliary function method.

Definition 1 (Auxiliary function). *Let $\Phi(\mathbf{\Theta})$ be the objective function that one wants to minimize with respect to the parameter $\mathbf{\Theta} = (\Theta_1, \cdots, \Theta_I)$. We then define $\Phi^+(\mathbf{\Theta}, m)$ as the*

auxiliary function of $\Phi(\boldsymbol{\Theta})$*, and* $m = (m_1, \cdots, m_J)$ *as the auxiliary parameter if* $\Phi^+(\boldsymbol{\Theta}, m)$ *satisfies*

$$\Phi\big(\boldsymbol{\Theta}\big) \leqq \Phi^+\big(\boldsymbol{\Theta}, m\big), \tag{6.35}$$

or

$$\Phi\big(\boldsymbol{\Theta}\big) = \min_{m} \Phi^+\big(\boldsymbol{\Theta}, m\big). \tag{6.36}$$

Lemma 2 (Auxiliary function method). *Denoting by* $\Phi(\boldsymbol{\Theta})$ *the objective function, and by* $\Phi^+(\boldsymbol{\Theta}, m)$ *the auxiliary function of* $\Phi(\boldsymbol{\Theta})$*, then the objective function* $\Phi(\boldsymbol{\Theta})$ *can be decreased monotonically by minimizing* $\Phi^+(\boldsymbol{\Theta}, m)$ *iteratively with respect to* $m = (m_1, \cdots, m_J)$ *and with respect to* $\Theta_1, \cdots, \Theta_I$*:*

$$\widehat{m} = \operatorname*{argmin}_{m} \Phi^+\big(\boldsymbol{\Theta}, m\big) \tag{6.37}$$

$$\forall i,\ \widehat{\Theta}_i = \operatorname*{argmin}_{\Theta_i} \Phi^+\big(\widehat{\Theta}_1, \cdots, \widehat{\Theta}_{i-1}, \Theta_i, \cdots, \Theta_I, m\big). \tag{6.38}$$

If $\Phi(\boldsymbol{\Theta})$ *is bounded below, then the parameter* $\boldsymbol{\Theta}$ *converges to a stationary point.*

Proof. Suppose we set the parameter to an arbitrary value $\boldsymbol{\Theta}^{(0)}$. We will prove that $\Phi(\boldsymbol{\Theta})$ necessarily decreases after the update Eq. (6.37) and Eq. (6.38). From Eq. (6.37), one obtains

$$\Phi(\boldsymbol{\Theta}^{(0)}) = \Phi^+(\boldsymbol{\Theta}^{(0)}, \widehat{m}), \tag{6.39}$$

and it is obvious from Eq. (6.38) that

$$\Phi^+(\boldsymbol{\Theta}^{(0)}, \widehat{m}) \geqq \Phi^+(\widehat{\boldsymbol{\Theta}}, \widehat{m}). \tag{6.40}$$

By definition, one sees from Eq. (6.35) that

$$\Phi^+(\widehat{\boldsymbol{\Theta}}, \widehat{m}) \geqq \Phi(\widehat{\boldsymbol{\Theta}}). \tag{6.41}$$

Therefore, we can immediately prove that

$$\Phi(\boldsymbol{\Theta}^{(0)}) = \Phi^+(\boldsymbol{\Theta}^{(0)}, \widehat{m}) \geqq \Phi^+(\widehat{\boldsymbol{\Theta}}, \widehat{m}) \geqq \Phi(\widehat{\boldsymbol{\Theta}}). \tag{6.42}$$

□

Having in mind applying this method to some optimization problem, it is important to design an auxiliary function such that the update equations for both the auxiliary parameter and the model parameters can be obtained analytically. It should be emphasized here that the EM algorithm can be considered as a special case of this method.

6.4.2 Inequality for L^2 norm

One possible auxiliary function of Eq. (6.34) can be made using the inequality for L^2 norm suggested for example by Feder *et al.* [39].

Lemma 3 (Inequality for L^2 norm). *If some complex function $m_i(x)$ satisfies*

$$\forall x,\ \sum_{i=1}^{I} m_i(x) = \sum_{i=1}^{I} m_i^*(x) = 1, \tag{6.43}$$

then

$$\int_{-\infty}^{\infty} \left\| y(x) - \sum_{i=1}^{I} s_i(x) \right\|^2 \mathrm{d}x \leqq \sum_{i=1}^{I} \frac{1}{\beta_i} \int_{-\infty}^{\infty} \left\| m_i(x)y(x) - s_i(x) \right\|^2 \mathrm{d}x, \tag{6.44}$$

and the equality holds if and only if

$$m_i(x) = \frac{1}{y(x)} \left[s_i(x) + \beta_i \left(y(x) - \sum_{i=1}^{I} s_i(x) \right) \right]. \tag{6.45}$$

β_i is an arbitrary constant such that

$$\sum_{i=1}^{I} \beta_i = 1 \tag{6.46}$$

$$0 < \beta_i < 1, \quad i \in \{1, \cdots, I\}. \tag{6.47}$$

Proof. We prove that the minimum of the right-hand side with respect to $m_i(x)$ is equal to the left-hand side using the variational method. Consider here the functional

$$J[m] \triangleq \sum_{i=1}^{I} \frac{1}{\beta_i} \int_{-\infty}^{\infty} \left\| m_i(x)y(x) - s_i(x) \right\|^2 \mathrm{d}x - \int_{-\infty}^{\infty} \lambda(x) \left(\sum_{i=1}^{I} m_i^*(x) - 1 \right) \mathrm{d}x \tag{6.48}$$

where the second term is the Lagrange multiplier term corresponding to the condition Eq. (6.43). The variation of $J[m]$ with respect to $m_i^*(t)$ is given as

$$\delta J[m] = \sum_{i=1}^{I} \int_{-\infty}^{\infty} \left(\frac{\partial J[m]}{\partial m_i^*} \right) \delta m_i^* \mathrm{d}x, \tag{6.49}$$

which should be 0 at the minimum point. In order to let this be identically 0, one must solve $\frac{\partial J[m]}{\partial m_i^*} = 0$. Hence, setting

$$\frac{\partial J[m]}{\partial m^*} = \frac{1}{\beta_i} y^*(x) \Big(m_i(x)y(x) - s_i(x) \Big) - \lambda(x) \tag{6.50}$$

to 0, one obtains

$$m_i(x) = \frac{1}{\left\|y(t)\right\|^2}\Big(\beta_i\lambda(x) + y^*(x)s_i(x)\Big). \tag{6.51}$$

From Eq. (6.43),

$$\sum_{i=1}^{I} m_i(x) = \sum_{i=1}^{I} \frac{1}{\left\|y(x)\right\|^2}\Big(\beta_i\lambda(x) + y^*(x)s_i(x)\Big) = \frac{1}{\left\|y(x)\right\|^2}\left(\lambda(x) + y^*(x)\sum_{i=1}^{I} s_i(x)\right) = 1. \tag{6.52}$$

Therefore,

$$\lambda(x) = \left\|y(x)\right\|^2 - y^*(x)\sum_{i=1}^{I} s_i(x). \tag{6.53}$$

Substituting this result into Eq. (6.51), the extreme value is determined uniquely as:

$$m_i(x) = \frac{1}{y(x)}\left[s_i(x) + \beta_i\left(y(x) - \sum_{i=1}^{I} s_i(x)\right)\right]. \tag{6.54}$$

One immediately notices that the sign of equality in Eq. (6.44) holds when $m_i(x)$ is given by this result. Whether this extreme value is the minimum solution or not can be shown easily by checking that the Hessian of $J[m]$ with respect to $\boldsymbol{m}(x)$, given by $\mathrm{diag}\big(\frac{\|y(x)\|^2}{\beta_1}, \cdots, \frac{\|y(x)\|^2}{\beta_I}\big)$, is obviously positive definite. □

Putting $S_{k,n}(\omega) \triangleq A_{k,n}e^{-\frac{(\omega - n\mu_k)^2}{4d}}$ for simplicity of notation, then by the Lemma 3 and from Eq. (6.34) we have the following inequality:

$$\begin{aligned}\Phi(\boldsymbol{\Theta}) &= \int_{-\infty}^{\infty}\left\|Y(\omega) - \sum_{k=1}^{K}\sum_{n=1}^{N} S_{k,n}(\omega)\right\|^2 \mathrm{d}\omega \\ &\leqq \sum_{k=1}^{K}\sum_{n=1}^{N}\frac{1}{\beta_{k,n}}\int_{-\infty}^{\infty}\Big\|m_{k,n}(\omega)Y(\omega) - S_{k,n}(\omega)\Big\|^2 \mathrm{d}\omega,\end{aligned} \tag{6.55}$$

where $\beta_{k,n} \in (0,1)$, $\sum_{k,n}\beta_{k,n} = 1$. The sign of the equality holds when

$$m_{k,n}(\omega) = \frac{1}{Y(\omega)}\left[S_{k,n}(\omega) + \beta_{k,n}\left(Y(\omega) - \sum_{k=1}^{K}\sum_{n=1}^{N} S_{k,n}(\omega)\right)\right]. \tag{6.56}$$

Let $\Phi^+(\boldsymbol{\Theta}, m)$ be the right-hand side of Eq. (6.55). By Definition 1, $\Phi^+(\boldsymbol{\Theta}, m)$ is an auxiliary function of the objective $\Phi(\boldsymbol{\Theta})$, and $m_{k,n}(\omega)$ is an auxiliary parameter, respectively. Eq. (6.56) corresponds to the update equation for Eq. (6.37) in Lemma 2.

This inequality implies that L^2 norm of the error between the observed signal and the sinusoidal model is the lower limit of the weighted sum of L^2 norm of each error between

an arbitrarily decomposed component $m_{k,n}(\omega)Y(\omega)$ and the single sinusoid $S_{k,n}(\omega)$. As the auxiliary parameter $m_{k,n}(\omega)$ acts as sort of a filter that decomposes the observed signal, we will henceforth call it a *decomposing filter*. Eq. (6.56) implies that letting $m_{k,n}(\omega)Y(\omega)$ be the sum of the $\{k,n\}^{\text{th}}$ sinusoid and the portion of the error between the observed signal and the sinusoidal model is said to be the "optimal" way of separating $Y(\omega)$.

By Lemma 2, we consider next to minimize $\Phi^+(\boldsymbol{\Theta}, m)$ with respect to $\boldsymbol{\Theta}$. As $\Phi^+(\boldsymbol{\Theta}, m)$ can be written as

$$\Phi^+(\boldsymbol{\Theta}, m) = \sum_{k=1}^{K}\sum_{n=1}^{N}\frac{1}{\beta_{k,n}}\int_{-\infty}^{\infty}\left(\left\|m_{k,n}(\omega)Y(\omega)\right\|^2+\left\|S_{k,n}(\omega)\right\|^2-2e^{-\frac{(\omega-n\mu_k)^2}{4d}}\mathrm{Re}\Big[A_{k,n}m_{k,n}^*(\omega)Y^*(\omega)\Big]\right)\mathrm{d}\omega, \tag{6.57}$$

from which we see that the integral of the second term inside the parenthesis can be calculated straightforwardly using the Gaussian integral:

$$\int_{-\infty}^{\infty}\left\|S_{k,n}(\omega)\right\|^2\mathrm{d}\omega = \sqrt{2\pi d}\left\|A_{k,n}\right\|^2. \tag{6.58}$$

Hence, this term does not depend on μ_k. Eq. (6.57) can thus be written as follows:

$$\begin{aligned}\Phi^+(\boldsymbol{\Theta}, m) = \sqrt{2\pi d}\sum_{k=1}^{K}\sum_{n=1}^{N}\frac{\left\|A_{k,n}\right\|^2}{\beta_{k,n}} \\ +\sum_{k=1}^{K}\sum_{n=1}^{N}\frac{1}{\beta_{k,n}}\int_{-\infty}^{\infty}\left(\left\|m_{k,n}(\omega)Y(\omega)\right\|^2-2e^{-\frac{(\omega-n\mu_k)^2}{4d}}\mathrm{Re}\Big[A_{k,n}m_{k,n}^*(\omega)Y^*(\omega)\Big]\right)\mathrm{d}\omega.\end{aligned} \tag{6.59}$$

One notices from Eq. (6.59) that one still cannot obtain analytically the update equation for μ_k because μ_k appears inside the exponential. In the next subsection, we will derive another auxiliary function that enables the analytical expression of the update equation for μ_k, using the property of exponential function.

6.4.3 Theorem on Differentiable Concave Functions

Not being able to obtain analytically the update equation for μ_k is because of the nonlinear part $\exp(-\frac{(\omega-n\mu_k)^2}{4d})$ in Eq. (6.59). To further derive another auxiliary function such that the update equation for μ_k can be obtained analytically, we focused on two points: $-e^{-x}$ is a continuously differentiable concave function, and we have the following theorem about continuously differentiable concave function.

Lemma 4 (Inequality on differentiable concave functions). *Let $f(x)$ be a real function of x that is continuously differentiable and concave. Then, for any point $\alpha \in \mathbb{R}$,*

$$f(x) \leqq f(\alpha) + (x-\alpha)f'(\alpha), \tag{6.60}$$

where $f'(\alpha) = \frac{\mathrm{d}f(x)}{\mathrm{d}x}\big|_{x=\alpha}$.

Proof. By definition, for any two points $x, \alpha \in \mathbb{R}$ and for any real number $\gamma \in (0,1)$, if

$$f\Big(\alpha x + (1-\gamma)\alpha\Big) \geqq \gamma f(x) + (1-\gamma)f(\alpha), \tag{6.61}$$

then $f(x)$ is said to be a concave function. This inequality can be rewritten as

$$\frac{f\Big(\gamma x + (1-\gamma)\alpha\Big) - f(\alpha)}{\gamma} \geqq f(x) - f(\alpha). \tag{6.62}$$

Since $\gamma x + (1-\gamma)\alpha = \alpha + \gamma(x-\alpha)$,

$$(x-\alpha)\frac{f\Big(\alpha + \gamma(x-\alpha)\Big) - f(\alpha)}{\gamma(x-\alpha)} \geqq f(x) - f(\alpha). \tag{6.63}$$

As $f(x)$ is assumed to be differentiable, when $\gamma \to 0$,

$$\lim_{\gamma\to 0} \frac{f\Big(\alpha + \gamma(x-\alpha)\Big) - f(\alpha)}{\gamma(x-\alpha)} = f'(\alpha). \tag{6.64}$$

Substituting this expression into Eq. (6.63), one obtains

$$(x-\alpha)f'(\alpha) \geqq f(x) - f(\alpha). \tag{6.65}$$

□

Since $-e^{-x}$ is a differentiable concave function, using Lemma 4 we have

$$-e^{-x} \leqq -e^{-\alpha} + \big(x-\alpha\big)e^{-\alpha}, \tag{6.66}$$

for any point $\alpha \in \mathbb{R}$. Replacing x with $\frac{(\omega-n\mu_k)^2}{4d}$ and α with a real function $\alpha_{k,n}(\omega)$, then

$$-e^{-\frac{(\omega-n\mu_k)^2}{4d}} \leqq -e^{-\alpha_{k,n}(\omega)} + \left(\frac{(\omega-n\mu_k)^2}{4d} - \alpha_{k,n}(\omega)\right) e^{-\alpha_{k,n}(\omega)}. \tag{6.67}$$

From Eq. (6.59) and Eq. (6.67),

$$\begin{aligned}\Phi^+(\boldsymbol{\Theta}, m) \leq \sqrt{2\pi d}\ \sum_{k=1}^{K}\sum_{n=1}^{N} \frac{\big\|A_{k,n}\big\|^2}{\beta_{k,n}} + \\ \sum_{k=1}^{K}\sum_{n=1}^{N} \frac{1}{\beta_{k,n}} \int_{-\infty}^{\infty} \Bigg[\big\|Y_{k,n}(\omega)\big\|^2 + 2\mathrm{Re}\Big[A_{k,n} m_{k,n}^*(\omega) Y^*(\omega)\Big] \\ \Bigg\{ -e^{-\alpha_{k,n}(\omega)} + e^{-\alpha_{k,n}(\omega)} \left(\frac{(\omega-n\mu_k)^2}{4d} - \alpha_{k,n}(\omega)\right) \Bigg\} \Bigg] \mathrm{d}\omega.\end{aligned} \tag{6.68}$$

Denoting by $\widetilde{\Phi}^{+}(\boldsymbol{\Theta}, m, \alpha)$ the right-hand side of this inequation, $\widetilde{\Phi}^{+}(\boldsymbol{\Theta}, m, \alpha)$ can also be considered as an auxiliary function of $\Phi(\boldsymbol{\Theta})$ because

$$\Phi(\boldsymbol{\Theta}) \leqq \Phi^{+}(\boldsymbol{\Theta}, m) \leqq \widetilde{\Phi}^{+}(\boldsymbol{\Theta}, m, \alpha). \tag{6.69}$$

In such case, both $m_{k,n}(\omega)$ and $\alpha_{k,n}(\omega)$ are the corresponding auxiliary parameters. Equality $\Phi(\boldsymbol{\Theta}) = \widetilde{\Phi}^{+}(\boldsymbol{\Theta}, m, \alpha)$ holds if and only if $m_{k,n}(\omega)$ is given by Eq. (6.56) and $\alpha_{k,n}(\omega)$ by

$$\alpha_{k,n}(\omega) = \frac{(\omega - n\mu_k)^2}{4d}. \tag{6.70}$$

6.4.4 Update Equations for Sinusoidal Parameters

There are two advantages worth mentioning of deriving this auxiliary function. One is that this enables the analytical expression of the update equation for the F_0 parameter μ_k, allowing a complex-spectrum-domain EM-like multipitch estimation.

Setting to 0 the partial derivative of $\widetilde{\Phi}^{+}(\boldsymbol{\Theta}, m, \alpha)$ with respect to μ_k:

$$\frac{\partial \widetilde{\Phi}^{+}(\boldsymbol{\Theta}, m, \alpha)}{\partial \mu_k} = \sum_{n=1}^{N} \frac{1}{\beta_{k,n}} \int_{-\infty}^{\infty} e^{-\alpha_{k,n}(\omega)} \mathrm{Re}\Big[A_{k,n} m_{k,n}^{*}(\omega) Y^{*}(\omega)\Big] \frac{-n\,(\omega - n\mu_k)}{d} \mathrm{d}\omega, \tag{6.71}$$

if

$$\sum_{n=1}^{N} \frac{n^2}{\beta_{k,n}} \int_{-\infty}^{\infty} e^{-\alpha_{k,n}(\omega)} \mathrm{Re}\Big[A_{k,n} m_{k,n}^{*}(\omega) Y^{*}(\omega)\Big] \mathrm{d}\omega \neq 0, \tag{6.72}$$

then one obtains

$$\mu_k = \frac{\displaystyle\sum_{n=1}^{N} \frac{n}{\beta_{k,n}} \int_{-\infty}^{\infty} e^{-\alpha_{k,n}(\omega)} \mathrm{Re}\Big[A_{k,n} m_{k,n}^{*}(\omega) Y^{*}(\omega)\Big] \omega \mathrm{d}\omega}{\displaystyle\sum_{n=1}^{N} \frac{n^2}{\beta_{k,n}} \int_{-\infty}^{\infty} e^{-\alpha_{k,n}(\omega)} \mathrm{Re}\Big[A_{k,n} m_{k,n}^{*}(\omega) Y^{*}(\omega)\Big] \mathrm{d}\omega}. \tag{6.73}$$

Secondly, the so far constant d can be regarded as a free variable $d_{k,n}$ for each k, n sinusoidal component and its update equation can be derived analytically. The ML estimate of $d_{k,n}$ itself is not important to us as its true value is already known, but by updating $d_{k,n}$ in parallel to the other parameters, we expect $d_{k,n}$ to play a similar role to the variance parameter in GMM, which often helps other parameters getting out of local optima during the parameter learning. The update equation for $d_{k,n}$ is given as

$$d_{k,n} = \left(\frac{\int_{-\infty}^{\infty} e^{-\alpha_{k,n}(\omega)} \mathrm{Re}\Big[A_{k,n} m_{k,n}^{*}(\omega) Y^{*}(\omega)\Big] (\omega - n\mu_k)^2 \mathrm{d}\omega}{\sqrt{2\pi} \big\|A_{k,n}\big\|^2} \right)^{2/3} \tag{6.74}$$

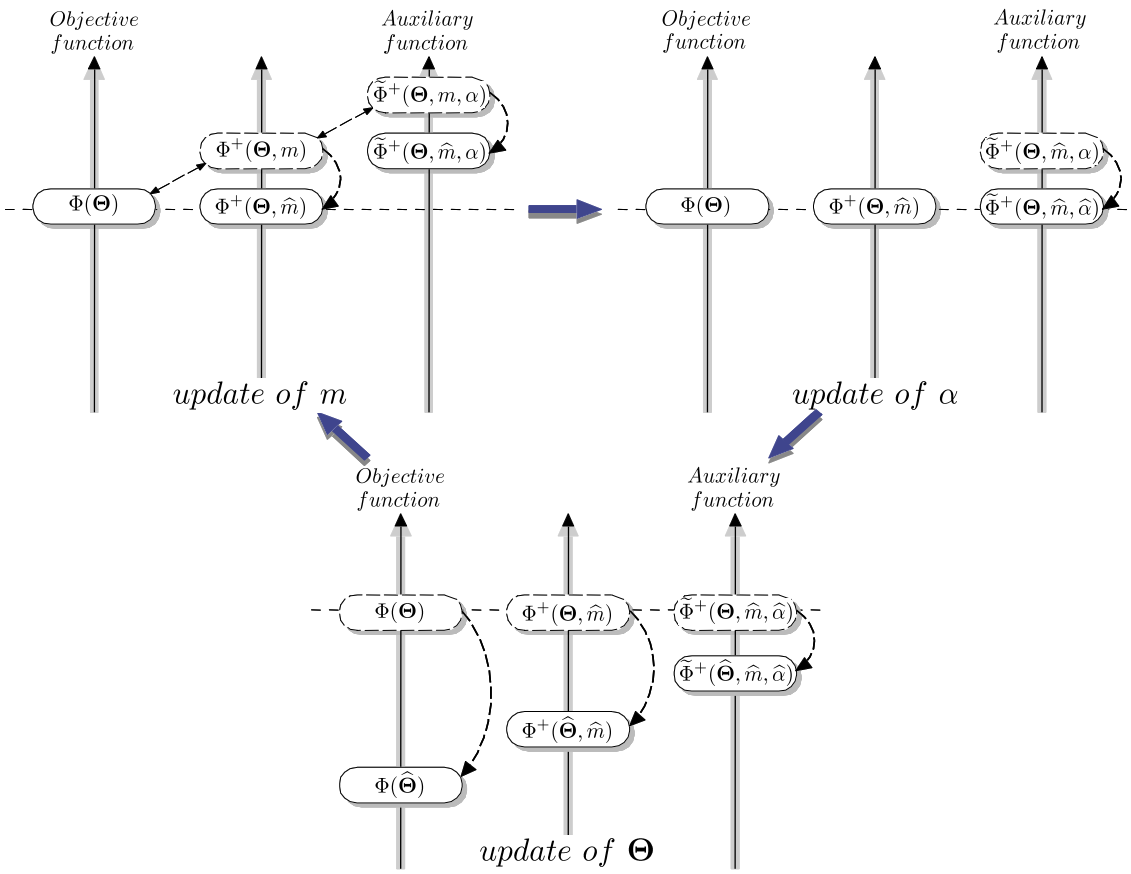


Figure 6.1 An illustration of the proposed parameter optimization algorithm

We should mention also that the update equation for $A_{k,n}$ can be derived analytically. Setting to 0 the partial derivative of $\widetilde{\Phi}^+(\mathbf{\Theta}, m, \alpha)$ with respect to $A^*_{k,n}$:

$$\frac{\partial \widetilde{\Phi}^+(\mathbf{\Theta}, m)}{\partial A^*_{k,n}} = \frac{\sqrt{2\pi d} A_{k,n}}{\beta_{k,n}} - \frac{1}{\beta_{k,n}} \int_{-\infty}^{\infty} \left\{ -e^{-\alpha_{k,n}(\omega)} + e^{-\alpha_{k,n}(\omega)} \left(\frac{(\omega - n\mu_k)^2}{4d} - \alpha_{k,n}(\omega) \right) \right\} m_{k,n}(\omega) Y(\omega) \mathrm{d}\omega, \quad (6.75)$$

one immediately obtains

$$A_{k,n} = \frac{1}{\sqrt{2\pi d}} \int_{-\infty}^{\infty} \left\{ -e^{-\alpha_{k,n}(\omega)} + e^{-\alpha_{k,n}(\omega)} \left(\frac{(\omega - n\mu_k)^2}{4d} - \alpha_{k,n}(\omega) \right) \right\} m_{k,n}(\omega) Y(\omega) \mathrm{d}\omega. \quad (6.76)$$

The amplitude and the starting phase of the each sinusoidal component can be expressed using $A_{k,n}$ as $a_{k,n} = \left| A_{k,n} \right|$ and $\varphi_{k,n} = \arg\left(A_{k,n} \right)$, respectively.

6.4.5 Overview of the Algorithm

We summarize here the global structure of the algorithm for the optimization of the sinusoidal signal model parameters. The transitions between the objective functions of each

step of the iteration are represented in Fig. 6.1.

Step 0 Initial setting of $\{\mu_k, \{A_{k,n}\}_{1\leq n\leq N}\}_{1\leq k\leq K}$.

Step 1 Update $m_{k,n}(\omega)$ through Eq. (6.56).

Step 2 Update $e^{-\alpha_{k,n}(\omega)}$ through Eq. (6.70).

Step 3 Update $A_{k,n}$ through Eq. (6.76)

Step 4 Update μ_k through Eq. (6.73) and go back to Step 1.

6.5 Experimental Evaluation

6.5.1 Convergence Properties of the Algorithm

The goal of this subsection is to compare the dependency on the initial parameter and the convergence speed of the gradient search method and the proposed method. The gradient search method-based parameter estimation method we use here as a comparison (hereafter simply called gradient method) is based on the Jinachitra's method [57] and composed of three steps, the update of $m_{k,n}(\omega)$ through Eq. (6.37), the update of $A_{k,n}$ through Eq. (6.76) and the decrease through steepest descent update of μ_k through Eq. (6.59). From the comparison of this method to the proposed method, we show the effectiveness of the pitch frequency estimation method proposed in this chapter in terms of ability to avoid local solutions.

In this comparative experiment, a signal which parameters are already known (synthetic signal) is analyzed. Specifically, two periodic signals (with pitch frequencies 207Hz and 200Hz) composed of 10 harmonic components, with each component's amplitude and phase determined by random generation, were added together to create a mixed signal. The interval of definition for the random generation of the amplitude and phase of the n-th harmonic component were respectively $[\frac{1}{n}, \frac{3}{n})$ and $[0, 2\pi)$. In the sinusoidal signal model, we set $K = 2$ and $N = 10$. A Gabor transform with diffusion parameter $d = 0.067$ was performed on this synthetic signal (16kHz sampling frequency) to obtain the short-time complex spectrum $Y(\omega)$.

The courses of the update of the pitch frequencies μ_1, μ_2 as they are updated at each step through the proposed method and the gradient method, starting from various initial parameter conditions, are shown in Fig. 6.2 and Fig. 6.3 (the update pattern of the parameters

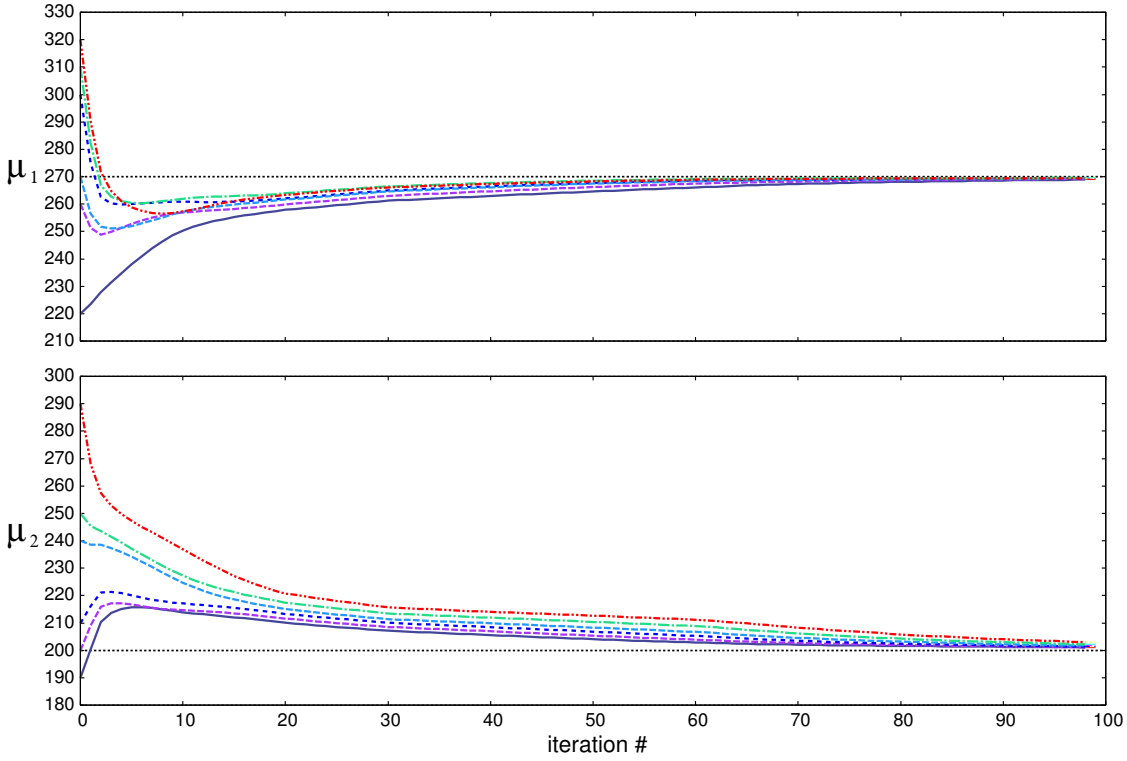


Figure 6.2 *Course of the pitch frequency update for the proposed method*

except the pitch frequencies is omitted). The transitions of the update values of μ_1 and μ_2 corresponding to the same iterative computation are shown in each figure respectively in the upper and lower part with the same color and same line type. The initial value for the amplitude $A_{k,n}$ was set to 0.

One sees from Fig. 6.2 and Fig. 6.3 that the gradient method often gets trapped into stationary points different from the true values for initial values of μ_1, μ_2 which are not sufficiently close to the true values (270Hz、270Hz), while the proposed method converges quickly from any initial points in a large interval to the true values. The result of this simulation is one illustration of the fact that the proposed method outperforms the previous works using gradient methods in terms of ability to avoid local solutions and convergence speed.

6.5.2 1ch Blind Source Separation of Concurrent Speech

Next, we confirm here the basic performance of our method for 1ch blind source separation. We use the ATR B-set speech database to build the mixed signals by adding together the

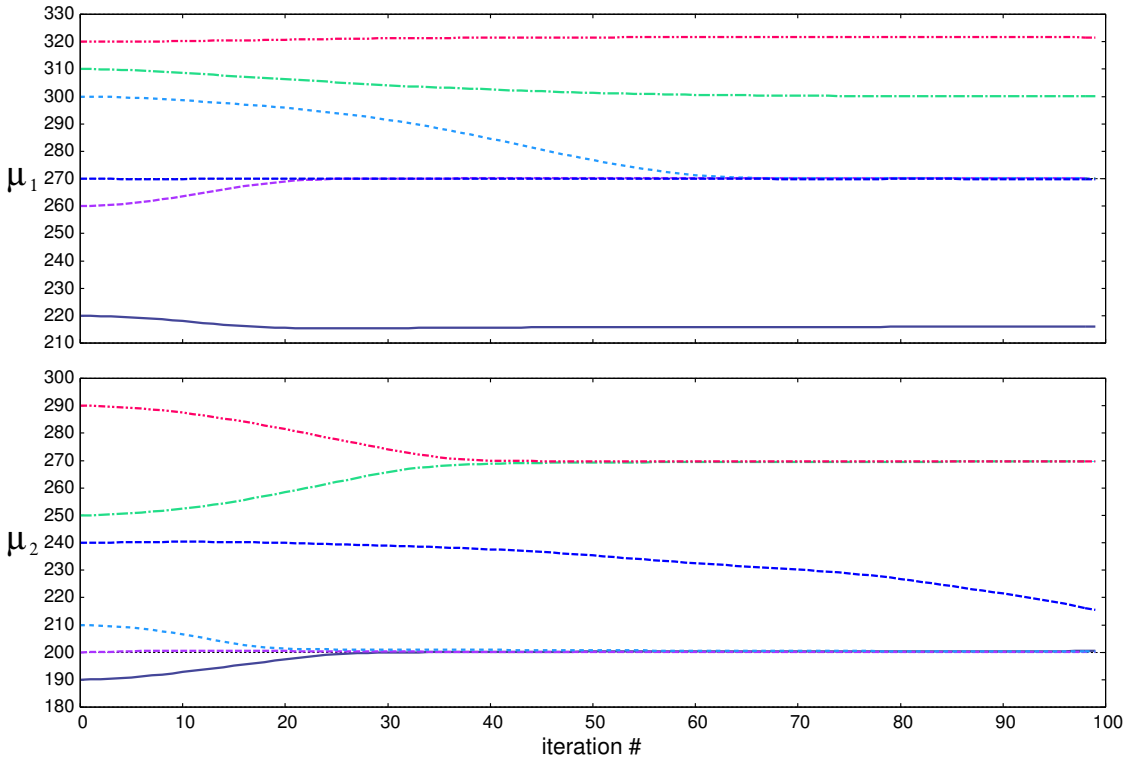


Figure 6.3 *Course of the pitch frequency update for the steepest descent method*

waveforms of utterances from two male speakers, two female speakers, or a male speaker and a female speaker.

For all the speech data the sampling rate was 16kHz, and the frequency analysis was done using a Gabor transform with a frame interval of 10ms. As in the preceding subsection, the diffusion parameter was set to 0.067. The number of harmonic components N of each harmonic signal of the sinusoidal signal model was set to 30.

The overview of the algorithm used here is as follows: starting from a sinusoidal signal model with an initial number K of harmonic signals equal to 10, in the process of the iterative estimation of the parameters, if the pitch frequency parameters of several harmonic signal models (1) come closer than a fixed value or (2) see their ratio become almost integer, the harmonic signal with the lowest pitch frequency only is kept and the other discarded. After convergence, the two harmonic signals with the largest total power are kept and parameter estimation is performed once again. The two harmonic signals thus obtained eventually are the separated signals. The initial values for μ_k are obtained by finding all the frequencies giving a minimum or a maximum of the real part or the imaginary part of the complex spectrum of the observed mixed signal, and selecting the 10 frequencies which correspond

to the largest power. The procedure described above estimates the separated signals on each short-time window (frame), but we do not determine here to which source the signal separated at each frame correspond. In this experiment, in order to check the basic source separation performance in the situation where this source determination would be dealt with, we determine to which source the separated signals correspond by looking at their proximity to each signal prior to the mixing.

Under the above conditions, an example of actual results of the separation of the mixed signal shown in Fig. 6.4 is shown in Fig. 6.5. After separation performed on the mixed signal of the male speaker A and the female speaker B (with a SNR of -0.3dB seen from the male speaker A), the SNRs for the speakers were respectively 7.2dB and 6.4dB (improvement of 7.5dB and 6.1dB), after separation performed on the mixed signal of the female speaker A and the female speaker B (with a SNR of 1.5dB seen from the female speaker A), the SNRs for the speakers were respectively 6.0dB and 4.8dB (improvement of 4.5dB and 6.3dB), after separation performed on the mixed signal of the male speaker A and the male speaker B (with a SNR of -0.3dB seen from the male speaker A), the SNRs for the speakers were respectively 4.8dB and 4.3dB (improvement of 5.1dB and 4.0dB). As in our method the difference between the pitch frequencies of the two speakers is clue for the source separation, the fact that the separation accuracy on mixed signal with speakers of the same gender is slightly lower than the accuracy on mixed signals with speakers of different gender is a result which corresponds to what we expected.

As the method presented in this chapter estimates the parameter independently for each frame, it happens quite often that the phase change of the separated signals is not continuous or the amplitude of varies abruptly. In the future, if a coordinated parameter estimation accross several adjacent frames could be performed, we shall expect a substantial reduction of the musical noise and an improvement of the SNR.

6.6 Summary of Chapter 6

In this chapter, focusing on the fact that the essential difficulty of the single tone frequency estimation or the fundamental frequency estimation, which are at the core of the parameter estimation problem for the sinusoidal signal model, comes from the non-linearity of the sinusoidal signal model in the frequency parameter, we introduced a new iterative estimation algorithm using an auxiliary function. Contrary to the power spectrum domain multi-pitch

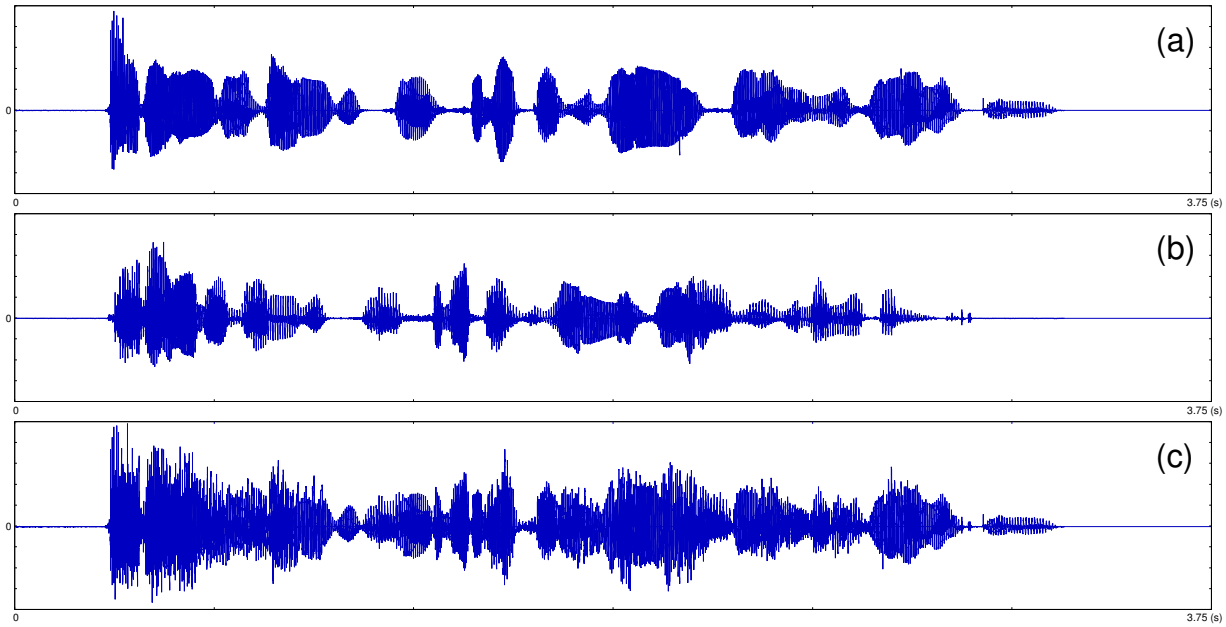


Figure 6.4 *Utterance by a female speaker (a), a male speaker (b) and their mixed signal (c).*

analysis methods discussed in the preceding chapters, this method does not assume that there is no interference between the components of different sources of between the harmonic components of a same source, and could become, depending on the accuracy of the parameter estimation, a very accurate method for the separation of frequency components which are close to each other.

In the present implementation, we derived the update equation of the fundamental frequency by transfering the objective function to the STFT domain, using the fact that through Parseval equality the L^2 norm of the error defined in the time domain is equal to the L^2 norm of the error in the STFT domain. From the analogy with the performance of the multipitch analysis methods in the power spectrum domain presented in the preceding chapters, one can think that it is highly probable that a higher performance could be obtained by performing the parameter update in the time-frequency domain obtained through constant Q filterbank. We thus plan to concentrate heavily in the future on investigating the possibility to obtain a formulation for which convergence is guaranteed and to derive parameter update equations in this domain.

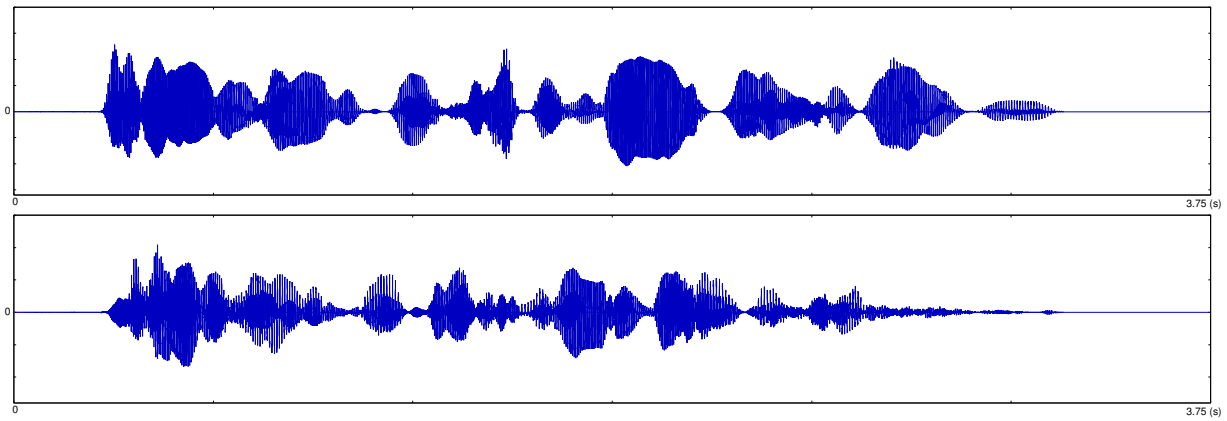


Figure 6.5 *Separated signals corresponding to the female and the male speaker.*

Chapter 7

Conclusion

The objective of this paper was to propose a unified methodological framework, in which one can handle (1) source separation, (2) multipitch estimation, (3) estimation of the number of sources, (4) estimation of the continuous temporal trajectories of F_0s and amplitudes, and (5) spectral envelope estimation, at the same time.

We introduced in Chapter 2 a method called "Harmonic Clustering". The method searches for the optimal spectral masking function and the optimal F_0 estimate for each source by performing the *source separation* step and the F_0 *estimation* step iteratively. In Chapter 3, we generalized the Harmonic Clustering method and then reformulated it from a Bayesian point of view. This Bayesian reformulation enabled us to derive a model selection criterion, that leaded to estimating the number of sources. We confirmed through experiments the effectiveness of the two techniques introduced in Chapter 3: multiple F_0 estimation and source number estimation.

In Chapter 4, based on Bregman's grouping cues, we proposed a new methodology to estimate simultaneously the spectral structure of each source on the whole time-frequency domain, which we called the "Harmonic-Temporal Clustering (HTC)". Through experimental evaluations on the F_0 estimation of mixed speech signals and music signals, we showed that our method's accuracy outperforms the previous state-of-the-art methods of each of these areas.

As F_0 estimation and spectral envelope estimation could be considered as "chicken and egg" problems, we formulated in Chapter 5 the estimation of F_0 and the spectral envelope as a joint optimization of a compound model of the spectral envelope and the fine structure. We found through experiments a significant advantage of jointly estimating F_0 and spectral

envelope in both F_0 estimation and spectral envelope estimation.

Taking into account the fact that it becomes usually difficult to estimate F_0s or to separate frequency components that are close to each other only based on the power spectrum, we considered that not only the harmonic structure but also the phase difference of each signal could be an important cue for separation. The main topic of Chapter 6 was the development of a non-linear optimization algorithm to obtain the maximum likelihood parameter of the sinusoidal signal model. We introduced a new iterative estimation algorithm using an auxiliary function, eventually allowing a complex-spectrum-domain EM-like multi-pitch estimation, which was inspired by the idea of the EM algorithm. Through simulation experiments, we showed that this parameter optimization algorithm outperformed existing gradient descent-based methods in the ability to avoid local solutions and the convergence speed. We also confirmed the basic performance of our method through 1ch speech separation experiments on real speech signal.

Acknowledgement (in Japanese)

本論文は、筆者が東京大学大学院情報理工学系研究科システム情報学専攻修士課程および博士課程に在学中に、嵯峨山研究室で行った研究をまとめたものであります。

まず始めに、卒業論文、修士論文、そしてこの博士論文の完成に至るまで長い間指導して頂いた、我が恩師であり、かつ、本論文の学位審査の主査である嵯峨山茂樹教授に敬意を込めて感謝します。本論文第 2 章で述べられている手法は筆者がまだ学部生だった頃に既に先生が考案していたものであり、筆者がこれに興味をもち、修士課程における研究テーマとして選んだことが本研究にとりかかるきっかけとなりました。当初は長く険しい道のりだと思っていた 5 年間の大学院生活を楽しく充実して過ごせたのは、筆者の自発性やアイディアを常に尊重、奨励してくれる嵯峨山先生の日頃の温かい指導方法のおかげであったように思います。また、本研究を進めていく過程で、幾度と無く研究の進むべき方向を正して頂いたり、数え切れないほどの示唆に富んだご助言を頂きました。

学位審査員を務めて頂いた嵯峨山研究室の講師でいらっしゃる小野順貴講師には、本研究を進めるにあたりいつも適切な助言を頂きました。また、小野先生のご自身の研究に対する真摯な姿勢や問題を数理的に理路整然と思考する様子には大変感銘を受け、そうした姿を傍らで拝見しながら、大きな刺激を受けるとともに数多くのことを学ぶことができました。第 6 章のアイディアを初めて伝えた際に先生から頂いた祝福のお言葉は忘れません。

お忙しい中にも関わらず、学位審査員を引き受けて下さったばかりでなく、予備審査会においては数々の有益な助言や指導を頂きました副査の舘暲教授、審査員の安藤繁教授、広瀬啓吉教授に感謝します。おかげで本審査までに論文をより充実させることができました。あまりにも偉大な御三方の先生に本論文を審査して頂けたことを、とても誇りに思います。

筆者が博士課程 1 年時まで嵯峨山研究室の連携講座客員教授でいらっしゃった守谷健弘博士 (NTT R&D フェロー) には、在任中、本研究に対し、嵯峨山先生とはまた違った切口からの大変に思慮深いご助言を数多く頂きました。任期終了後も、ときどきお会いした際には研究や就職の相談にいつも乗って頂きました。また、守谷先生には、これまで筆者の研究を財団賞と学会賞の 2 つの研究賞に推薦して頂きました。いつも筆者をごひいきにして頂いていることにも深く感謝しており、いつか何らかの形で恩返しできればと考えている次第です。

筆者が修士課程 1 年時まで嵯峨山研究室の助教授でいらっしゃり、かつ、当時の指導教官

である篠田浩一助教授 (東工大) には、論理的に隙のない研究の進め方やプレゼンテーションの重要性を幾度も説いて頂きました。

嵯峨山研究室の助手でいらっしゃる西本卓也助手、酒向慎司特任助手は、日頃より本研究に関してご助言頂いただけでなく、常に快適な研究環境を維持することにご尽力されておりました。御二方がいなければ本研究は決して成し得なかったと心より思います。

東大特任助手でいらっしゃった田原鉄也博士 (山武) は、筆者が戦略ソフトウェア創造人材養成プログラム (以後、戦略ソフト) でソフトウェア開発に携わっていた際に最も御世話になった人物です。同氏との議論を通じてウェーブレット変換に関する理解を深めることができ、本研究においてもその知識が大いに役立っております。平木敬教授、稲葉真理特任助教授、鈴木隆文講師ら (東大) には戦略ソフトの定例ミーティングにおいて、「動くものを作る」ことの重要性を何度も説いて頂きました。御三方から頂いた貴重な指導を今後に活かしていきたいと思います。

学外で最初に筆者の研究に関心を示して頂いたのは後藤真孝博士 (産総研) でした。同時に、本研究において最も関連の深い革新的な先行研究論文の著者であり、お会いする以前から憧れていた人物でした。筆者は、同氏と交わす議論は大変好きで、新しいアイディアを思いつくたびにいつも最初に同氏に意見を仰いでいました。第 4 章のアイディアを初めてお話したときに同氏より頂いた賛辞のお言葉は忘れません。また、同氏には、産総研実習に誘って頂き大学院以外の研究環境を実際に体験できる機会を与えて頂いたり、研究の新規性を打ち出せず気が滅入っていた時期には親身に励まして頂いたりと公私にわたり本当に御世話になりました。今後も同氏と何らかの形で関わりを持ちづけられればと思う次第です。

渡部晋治博士 (NTT) は、若くして豊かな知識と業績をもち、他分野の研究者ながら本研究に関心を示して頂き、含蓄のある指摘を多数下さりました。同氏と交わした議論は本論文に存分に活かされています。また、同氏には、NTT コミュニケーション科学基礎研究所 (CS 研) の学外実習に誘って頂き、そこで変分ベイズ法をはじめベイズ理論を学びました。

柏野邦夫博士 (NTT) も後藤氏同様、修士時代から拝読していた研究論文の著者であり、初めてお会いしたときに感激したのを覚えております。同氏より頂いた指摘やコメントも本論文完成には欠かせないものです。また、同氏は筆者に CS 研で講演する機会を与えて下さり、そのおかげで CS 研の面々と本研究に関する濃密な議論を交わすことができました。

Alain de Cheveigné 教授 (ENS) は、海外の研究者の中で最も親しくして頂いている人物であり、同時に本研究に非常に関連の深い研究者です。本論文の導入や研究背景の章は、同氏が執筆した著書を参考にしました。また、パリで開かれた聴覚に関する国際ワークショップ「New Ideas in Hearing」にディスカッサントとして旅費つきで招待して頂き、聴覚研究の活発な議論に触れられる貴重な機会を与えて下さりました。

浅野太博士、麻生英樹氏、緒方淳博士 (産総研) には、筆者の産総研での夏期実習時に大変御世話になりました。特に、浅野氏と麻生氏には、本研究に対し大変有益なご助言を頂きました。また、緒方氏は、音声情報処理分野では渡部氏と並んで新進気鋭と目される若手研究者で、筆者にとって格闘技の話題で盛り上がれる数少ない研究者仲間でもあります。

荒木章子氏、石塚健太郎氏、上田修功博士、大庭隆伸氏、木下慶介氏、Eric McDermott 博士、Michael Schuster 博士、澤田宏博士、中谷智広博士、中村篤博士、引地孝文博士、牧野昭二博士、南泰浩博士、三好正人博士 (NTT) には、NTT CS 研での学外実習時に大変御世話になりました。特に、実習中、中谷氏と南氏とは議論を交わす機会が幾度かあり、二方からはさまざまなことを教えて頂きました。また、荒木氏、大庭氏、木下氏らは年齢が近いこともあり、なんとなく不安な気持ちでいた実習開始間もない時期に仲良く接して頂き、そのおかげですぐに周囲と打ち解けることができました。

安部素嗣博士 (ソニー) は、後藤氏や柏野氏同様、本研究に関連する重要な先行研究論文の著者であり、あまりに難解なその内容から、著者は一体どのような方なのかと一度お会いしてみたいと思っていた人物でした。筆者が博士課程 3 年の時にようやくその念願が叶い、小野先生とソニーに訪問した際にお会いして議論を交わす機会がありました。まさにイメージしていた通りの才気溢れる方という印象で、同氏には、本論文における論理展開の不備の指摘や、改善するためのご助言を頂きました。

小坂直敏教授 (東京電機大)、片寄晴弘教授 (関西学院大)、菅野由弘教授 (早大)、長嶋洋一教授 (静岡大)、平田圭二博士 (NTT)、堀内靖雄助教授 (千葉大) をはじめ、情報処理学会音楽情報科学研究会の主要メンバーには大変御世話になりました。特に、片寄先生はいつも筆者のことを実力以上に評価していて下さり、本研究に対して頂いたご助言や激励のお言葉は数知れません。今後ともお付き合いできればと思う次第です。

嵯峨山先生は音声情報処理の研究者の間で国内外に非常に顔が広く、そのおかげで、板倉文忠教授 (名城大)、Wolfgang J. Hess 教授 (Bonn 大)、Frank K. Soong 博士 (Microsoft Asia)、Manfred R. Shröeder 博士ら錚々たる面々とディスカッションする貴重な機会がありました。修士 2 年時に板倉先生と議論を交わしていた際、嵯峨山先生と板倉先生の両者から「スペクトルは確率分布ではないから本来 EM アルゴリズムは適用できないのでは?」との指摘を受けたことがありました。その言葉の真の意味を理解するのに時間がかかりましたが、時間を隔てて本論文でその回答ができたことを嬉しく思います。また、この素朴な疑問をもつきっかけを頂いたことが、第 3 章以降の定式化とアイディア、および最終的に第 6 章 (EM アルゴリズムの拡張原理) の着想に至ったことは幸運だったとしか言いようがありません。

本研究並びに筆者の大学院での学生生活をずっと支えて頂いた、横坂満さん、森田俊哉さん、笠間邦子さんらシステム情報学専攻の事務職員一同、および嵯峨山研究室秘書でいらっ

しゃる金子玲子さん、尾関直子さんに感謝します。また、旧嵯峨山研秘書の小畠新子さん、河村裕子さんにも御世話になりました。

筆者が修士 1 年だった頃より、研究室先輩として基礎知識や計算機の扱い方を丁寧に教えて頂いた武田晴登氏 (関西学院大) と五十川賢造氏 (東芝) に感謝します。特に、武田さんは嵯峨山研究室メンバーの中では最も付き合いの長い人物です。多才な同氏から受けた影響は大きく、尊敬する先輩とともに嵯峨山研究室を創成期から一緒に支えてこれたことは今の大きな自信につながっています。菅原啓太氏 (ソニーエリクソン)、高橋佳吾氏 (警察庁)、山本仁氏 (NEC) は、筆者の修士時代の貴重な嵯峨山研究室同期生です。三方とは、今でも交流があり、今後もこの付き合いを大切にしていきたいと思います。

井上和士氏 (コルグ)、鎌本優氏 (NTT)、中潟昌平氏 (富士通) は、筆者にとって初めての研究室の後輩でした。とはいえ、井上氏と鎌本氏は年齢が同じだったこともあり、まるで同期のように仲良くして頂きました。井上氏は、筆者が博士 1 年時に隣席だったこともあり、当時最も多く研究の議論や雑談を日常的に交わしていた相手でした。この優秀な三人の後輩の存在が、当時の研究室の活気につながっていたことは幸運でした。槻武也氏、山本遼氏、齊藤翔一郎氏とは、ときに白熱した議論を、ときには実に他愛のない雑談を交わし合いました。三方の研究には筆者も興味があり、問題に直面しては一緒に考えたりしておりました。また、それらを通して筆者も学んだことが多くありました。槻氏は修士在学中に音声合成の研究に取り組んでおり、その過程で音声のスペクトル包絡を GMM で近似する何らかの方法が必要だという状況になり、それが筆者が第 5 章の手法を思いつくきっかけとなりました。山本遼氏は、槻氏と同期の現博士課程の学生であり、研究室の雰囲気を明るく楽しくすることに努力を惜しまない大変好ましい人物です。同氏が時々投げかけてくる質問は実に奥が深く、理解していたつもりの事柄を改めて考えさせられたことが何度もありました。齊藤氏とは、時が経つのを忘れてしまうくらい議論が白熱してしまうことが多々ありました。一方で、たまに同氏が発するジョークはときには筆者のツボにはまってしまうことがあり、そのせいで研究に集中できないことも多々ありました。

本論文の完成は、Jonathan Le Roux 氏の協力なくしてはありえませんでした。第 4 章に相当する研究は同氏とともに行ったものであり、音声分析の実験結果は同氏の実装によるものです。また、本論文の英語化と英文添削を献身的に手伝って頂きました。同氏いわく「謝金のためではなく友情のため」とのことですが真実は謎です。ともに研究できたこと、隣席であったこともあり、今ではすっかり嵯峨山研究室で最も仲の良い友人 (漫才コンビ) です。

筆者が在学する最後の一年を和泉洋介氏、松本恭輔氏、宮本賢一氏といった素質溢れる面々とともに過ごせたのは幸運でした。彼らの研究はいずれも非常にレベルの高いものであり、三方とホワイトボードで交わす議論はいつも刺激的でした。松本氏と宮本氏からは時々

ドキッとするような鋭い質問や指摘を受けることがあり、最も議論し甲斐のある後輩たちでした。和泉氏には公私にわたり御世話になりました。博士課程に進学するはずの同氏のことは今後も応援し続けるつもりですが、もし博士学生特有の悩みに直面したときにはいつでも相談に乗ってあげようと思います。隣の研究室の学生でありながら、藤田悠哉氏 (東大) とは雑談や研究に関連することがらの議論を交わしたりしました。同氏の研究の話をいつも楽しく聞かせて頂きました。

安東弘泰氏 (東大) とは、学部時代から博士課程までの 7 年間、別々の専攻に在籍していましたが、ともに切磋琢磨しながら大学院生活を送ってきました。筆者は、博士課程に進学した後、時として言いようのない孤独感に苛まれるときがありました。そうしたときにはいつも、同氏も同じ状況で頑張っているのだと思うことで、自然と自分自身を奮い立たせることができました。同氏とともに学位を取得することは、実に感慨深いものがあります。

大石康智氏 (名大)、北原鉄朗氏 (京大)、中野倫靖氏 (筑波大)、藤原弘将氏 (京大)、吉井和佳氏 (京大)、吉岡拓也氏 (NTT) は、学会や学外実習を通じて知り合った同志たちです。筆者が学会デビューを果たしたときに同じセッションで発表していたのが北原氏で、筆者と同学年にも関わらず当時既に何度も学会発表をこなしていたせいか同氏が醸し出していたエネルギーに圧倒されたのを今でも鮮明に記憶しております。産総研実習において筆者と同時期に実習生だったことがきっかけで仲良くなった大石氏と藤原氏は今や同分野を代表する若手研究者となりました。実習中最も多くの議論を交わしたのがこの両氏でした。彼らとの議論はとても楽しく、直接会える機会をいつも楽しみにしております。中野氏とは長い付き合いで、公私ともに仲の良い同分野の研究者仲間の一人です。同氏には見習うべきところが多く、特に、人一倍強い研究への情熱にはとても影響を受けました。吉岡氏には NTT CS 研実習中に御世話になりました。同氏との議論は大変刺激的で、議論を通じて同氏から学んだことが多くありました。研究者として生きていくのは、孤独との戦いの連続だとはじめは思っており、これだけ優秀で、かつ素晴しい仲間ができることは想像だにしていませんでした。同分野の同世代のこうした新進気鋭の研究者の活躍が発奮材料となったことも本論文完成の一端を担っています。今後とも学生時代にできたこの仲間たちとは、同分野で切磋琢磨していける良い関係でいたいと願います。

この他にも、本研究は、数多くの方々からの支援や議論なくしては成しえませんでした。本研究に関して議論して頂いたすべての方々に御礼申し上げます。また、このあまりにも長文の謝辞を最後まで読んで頂いた読者に感謝します。

最後に、大学院での研究を何不自由なく行えるよう筆者を常に陰から惜しみなく支え続けてくれた両親に、感謝の意を込めて本論文を捧げたいと思います。

平成 19 年 2 月 23 日

Bibliography

[1] T. J. Abatzoglou, "A Fast Maximum Likelihood Algorithm for Frequaency Estimation of a Sinusoid Based on Newton's Method," *IEEE Trans. Acoust., Speech, Signal Process.*, Vol. ASSP-33, No. 1, pp. 77–89, 1985.

[2] S. A. Abdallah, and M. D. Plumbley, "Polyphonic Music Transcription by Non-negative Sparse Coding of Power Spectra," In *Proc. ISMIR2004*, pp. 318–325, 2004.

[3] M. Abe, and S. Ando, "Auditory Scene Analysis Based on Time-Frequency Integration of Shared FM and AM (I): Lagrange Differential Features and Frequency-Axis Integration," *IEICE Trans.*, Vol. J83-D-II, No. 2, pp. 458–467, 2000 (in Japanese).

[4] M. Abe, and S. Ando, "Auditory Scene Analysis Based on Time-Frequency Integration of Shared FM and AM (II): Optimum Time-Domain Integration and Stream Sound Reconstruction," *IEICE Trans.*, Vol. J83-D-II, No. 2, pp. 468–477, 2000 (in Japanese).

[5] M. Abe, and J. O. Smith III, "Design Criteria for Simple Sinusoidal Parameter Estimation Based on Quadratic Interpolation of FFT Magnitude Peaks – For Quasi-Stationary Sinusoidal Components –," *Technical Report of IEICE*, EA2004-61/SIP2004-65/SIS2004-32, pp. 7–12, 2004 (in Japanese).

[6] S. S. Abeysekera, "An Efficient Hilbert Transform Interpolation Algorithm for Peak Position Estimation," In *Proc. IEEE Workshop on Statistical Signal Processing (SSP)*, pp. 417–420, 2001.

[7] H. Akaike, "On Entropy Maximization Principle," In Proc. Applications of Statistics, P. R. Krishnaiah, Ed. Amsterdam, North-Holland, pp. 27–41, 1977.

[8] S. Araki, S. Makino, H. Sawada and R. Mukai, "Reducing Musical Noise by a Fine-Shift Overlap-Add Method Applied to Source Separation Using a Time-Frequency Mask," In *Proc. ICASSP2005*, Vol. 3, pp. 81–84, 2005.

[9] B. S. Atal, "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification," *J. Acoust. Soc. Amer.*, Vol. 55, No. 6, pp. 1304–1312, 1974.

[10] F. Bach, and M. Jordan, "Discriminative Training of Hidden Markov Models for Multiple Pitch Tracking," In *Proc. IEEE ICASSP2005*, Vol. 5, pp. 489–492, 2005.

[11] P. C. Bagshaw, S. M. Hiller and M. A. Jack, "Enhanced Pitch Tracking and the Processing of F_0 Contours for Computer and Intonation Teaching," In *Proc. Eurospeech'93*, pp. 1003–1006, 1993.

[12] J. G. Beerends, "Pitches of Simultaneous Complex Tones," PhD thesis, Technical University of Eindhoven, 1989.

[13] J. G. Beerends, and A. J. M. Houtsma, "Pitch Identification of Simultaneous Diotic and Dichotic Two-tone Complexes," *J. Acoust. Soc. Am.*, Vol. 85, pp. 813–819, 1989.

[14] P. Boersma, "Accurate Short-term Analysis of the Fundamental Frequency and the Harmonics-to-noise Ratio of a Sampled Sound," In *Proc. Institute of Phonetic Sciences*, Vol. 17, pp. 97–110, 1993.

[15] P. Boersma and D. Weenin, "Praat system," `http://www.fon.hum.uva.nl/praat/`.

[16] A. S. Bregman, Auditory Scene Analysis, MIT Press, Cambridge, 1990.

[17] J. C. Brown *et al.*, "A High Resolution Fundamental Frequency Determination Based on Phase Changes of the Fourier Transform," *J. Acoust. Soc. Am.*, Vol. 94, No. 2, pp. 662–667, 1993.

[18] G. J. Brown and M. Cooke, "Computational Auditory Scene Analysis," Comp. Speech & Lang., No. 8, pp. 297–336, 1994.

[19] M. Campedel-Oudot, O. Cappé, and E. Moulines, "Estimation of the Spectral Envelope of Voiced Sounds Using a Penalized Likelihood Approach," *IEEE Trans. Speech and Audio Processing*, Vol. 9, No. 5, pp. 469–481, 2001.

[20] O. Cappé, J. Laroche and E. Moulines, "Regularized Estimation of Cepstrum Envelope from Discrete Frequency Points," In *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 1995.

[21] Cauwenberghs, "Monaural Separation of Independent Acoustical Components," In *IEEE Symp. Circuit and Systems (ISCAS)*, 1999.

[22] D. Chazan, Y. Stettiner and D. Malah, "Optimal Multi-Pitch Estimation Using the EM Algorithm for Co-Channel Speech Separation," In *Proc. ICASSP'93*, Vol. 2, pp. 728–731, 1993.

[23] A. de Cheveigné, "Separation of Concurrent Harmonic Sounds: Fundamental Frequency Estimation and a Time-domain Cancellation Model of Auditory Processing," *J. Acoust. Soc. Am.*, Vol. 93, pp. 3271–3290, 1993.

[24] A. de Cheveigné, and H. Kawahara, "Multiple Period Estimation and Pitch Perception Model," *Speech Communication*, Vol. 27, pp. 175–185, 1999.

[25] A. de Cheveigné and H. Kawahara, "YIN, a Fundamental Frequency Estimator for Speech and Music," *J. Acoust. Soc. Am.*, 111(4), pp. 1917–1930, 2002.

[26] A. de Cheveigné, and A. Baskind, "F0 Estimation of One or Several Voices," In *Proc. Eurospeech2003*, pp. 833–836, 2003.

[27] A. de Cheveigné, "Pitch Perception Models," In C. J. Plack, A. Oxenham, R. R. Fay, A. N. Popper, editors, *Pitch – Neural coding and perception*, Springer, New York, 2005.

[28] A. de Cheveigné, "Multiple F_0 Estimation," in *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, D. -L. Wang, G. J. Brown Eds., IEEE Press / Wisely, 2006.

[29] M. Cooke, G. J. Brown, M. Crawford and P. Green, "Computational Auditory Scene Analysis – Listening to Several Things at Once," Endevour New Series, Vol. 17, No. 4, pp. 186–190, 1993.

[30] I. Csiszar, "I-Divergence Geometry of Probability Distributions and Minimization Problems," *The Annals of Probability*, Vol. 3, No. 1, pp. 146–158, 1975.

[31] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, No. 4, pp. 357–366, 1980.

[32] M. Davy and S. Godsill, "Bayesian Harmonic Models for Musical Signal Analysis," In Bayesian Statistics 7, pp. 105–124. Oxford University Press, Oxford, 2003.

[33] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. of Royal Statistical Society Series B*, Vol. 39, pp. 1–38, 1977.

[34] Ph. Depalle and T. Hélie, "Extraction of Spectral Peak Parameters Using a Short-Time Fourier Transform Modeling and No Sidelobe Windows," In *Proc. IEEE WASPAA'97*, 1997.

[35] B. Doval, "Estimation de la fréquence fondamentale des signaux sonores," *Ph. D. Thesis at Université Pierre et Marie Curie*, 1994.

[36] A. El-Jaroudi, and J. Makhoul, "Discrete All-Pole Modeling," *IEEE Trans. Signal Process.*, Vol. 39, No. 2, pp. 411–423, 1991.

[37] D. P. W. Ellis, "A Computer Implementation of Psychoacoustic Grouping Rules," In *Proc. IEEE ICPR'94*, pp. 108–112, 1994.

[38] D. Ellis, "Prediction-driven Computational Auditory Scene Analysis," Ph.D. thesis, MIT, 1996.

[39] M. Feder and E. Weinstein, "Parameter Estimation of Superimposed Signals Using the EM Algorithm," *IEEE Trans. Acoust., Speech, Signal Process.*, Vol. ASSP-36, No. 4, pp. 477–489, 1988.

[40] A. Fishbach, "Primary Segmentation of Auditory Scene," In Proc. IEEE ICPR'94, pp. 113–117, 1994.

[41] D. H. Friedman, "Pseudo-Maximum-Likelihood Speech Pitch Extraction," *IEEE Trans. Acoust., Speech, Signal Process.*, Vol. AASP-25, No. 3, 1977.

[42] T. Galas and X. Rodet, "An Improved Cepstral Method for Deconcolution of Source-Filter Systems with Discrete Spectra: Application to Musical Sound Signals," In *Proc. ICMC'90*, pp. 82–84, 1991.

[43] S. Godsill and M. Davy, "Baysian Harmonic Models for Musical Pitch Estimation and Analysis," In *Proc. IEEE ICASSP2002*, Vol. 2, pp. 1769–1772, 2002.

[44] M. Goto, H. Hashiguchi, T. Nishimura and R. Oka "RWC Music Database: Popular, Classical, and Jazz Music Database," In *Proc. ISMIR 2002*, pp. 287–288, 2002.

[45] M. Goto, "A Real-Time Music-Scene-Description System: Predominant-F0 Estimation for Detecting Melody and Bass Lines in Real-World Audio Signals," *ISCA Journal*, Vol. 43, No. 4, pp. 311–329, 2004.

[46] S. Haykin, Unsupervised Adaptive Filtering, John Wiley & Sons, 2000.

[47] A. Hyvärinen, J. Karhunen and E. Oja, Independent Component Analysis, John Wiley & Sons, 2001.

[48] D. J. Hermes, "Measurement of Pitch by Subharmonic Summation," *J. Acoust. Soc. Am.*, Vol. 83, pp. 257–264, 1988.

[49] W. J. Hess, Pitch Determination of Speech Signals, (Springer-Verlag, Berlin), 1983.

[50] W. J. Hess, "Pitch and Voicing Determination," in *Advances in Speech Signal Processing*, edited by S. Furui and M. M. Sohndi (Marcel Dekker, New York), pp. 3–48, 1992.

[51] D. Huron, "Voice Denumerability in Polyphonic Music of Homogenous Timbres," *Music Perception*, Vol. 6, pp. 361–382, 1989.

[52] S. Imai and Y. Abe, "Spectral Envelope Extraction by Improved Cepstral Method," *Electron. and Commun. in Japan*, Vol. 62-A No. 4, pp. 10–17, 1979.

[53] F. Itakura and S. Saito, "Analysis Synthesis Telephony based upon the Maximum Likelihood Method," In *Proc. 6th ICA*, C-5-5, C17–20, 1968.

[54] F. Itakura, "A Study on Speech Analysis and Synthesis Based on a Statistical Method," Ph.D. Thesis, Nagoya University, 1972 (in Japanese).

[55] F. Itakura, S. Saito, T. Koike, H. Sawabe and M. Nishikawa, "An Audio Response Unit based on Partial Autocorrelation," *IEEE Trans. Commun.,* Vol. COM-20, pp. 792–797, 1972.

[56] F. Itakura, "Line Spectrum Representation of Linear Predictive Coefficients of Speech Signals," *J. Acoust. Soc. Am.*, Vol. 57, No. 1, S35(A), 1975.

[57] P. Jinachitra, "Constrained EM Estimates for Harmonic Source Separation," In *Proc. ICASSP2003*, Vol. 6, pp. 609–612, 2003.

[58] R. E. Kalman, "Design of a Self-optimizing Control System," Trans. ASME, Vol. 80, pp. 468–478, 1958.

[59] M. Karjalainen and T. Tolonen. Multipitch and periodicity analysis model for sound separation and auditory scene analysis. In *Proc. ICASSP'99*, pp. 929–932, 1999.

[60] K. Kashino, K. Nakadai, T. Kinoshita, and H. Tanaka, "Organization of Hierarchical Perceptual Sounds: Music Scene Analysis with Autonomous Processing Modules and a Quantitative Information Integration Mechanism," In *Proc. IJCAI'95*, Vol. 1, pp. 158–164, 1995.

[61] K. Kashino, K. Nakadai, T. Kinoshita, and H Tanaka, "Application of the Bayesian Probability Network to Music Scene Analysis," In D.F. Rosenthal and H.G. Okuno, editors, Computational Auditory Scene Analysis, pp. 115–137. Lawrence Erlbaum Associates, 1998.

[62] K. Kashino, and S. J Godsill, "Bayesian Estimation of Simultaneous Musical Notes Based on Frequency Domain Modelling," In *Proc. ICASSP'04*, Vol. 4, pp. 305–308, 2004.

[63] H. Kawahara, "Speech Respresentation and Transformation Using Adaptive Interpolation of Weighted Spectrum: Vocoder Revisted," In *Proc. ICASSP '97*, Vol. 2, pp. 1303–1306, 1997.

[64] H. Kawahara and H. Katayose and A. de Cheveigné and R. D. Patterson, "Fixed Point Analysis of Frequency to Instamtaneous Frequency Mapping for Accurate Estimation of F_0 and Periodicity," In *Proc. Eurospeech'99*, Vol. 6, pp. 2781–2784, 1999.

[65] H. Kawahara, I. Masuda-Katsuse and A. de Cheveigné, "Restructuring Speech Representations Using a Pitch-Adaptive Time-Frequency Smoothing and an Instantaneous-Frequency-Based F0 Extraction," *Speech Communication*, Vol. 27, pp. 187–207, 1999.

[66] A. Klapuri, "Signal Processing Methods for the Automatic Transcription of Music," Ph.D. thesis, Tampere University, 2002.

[67] A. P. Klapuri, "Multiple Fundamental Frequency Estimation Based on Harmonicity and Spectral Smoothness," *IEEE Trans. Speech, Audio, Process.*, Vol. 11, No. 6, pp. 804–816, 2003.

[68] G. E. Kopec, A. V. Oppenheim and J. M. Tribolet, "Speech Analysis by Homomorphic Prediction," *IEEE Trans. Acoust. Speech, Signal Process.*, ASSP-25, pp. 40–49, 1977.

[69] R. J. Leistikow, H. D. Thornburg, J. O. Smith III, and J. Berger, "Bayesian Identification of Closely-spaced Chords from Single-frame STFT Peaks," In *Proc. DAFX*, pp. 5–8, 2004.

[70] M. D. Macleod, "Fast Nearly ML Estimation of the Parameters of Real or Complex Single Tones or Resolved Multiple Tones," *IEEE Trans. Signal Process.*, Vol. SP-46, No. 1, pp. 141–148, 1998.

[71] R. J. McAulay and T. F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," *IEEE Trans. Acoust., Speech, Signal Process.*, Vol. ASSP-34, No. 4, pp. 744–754, 1986.

[72] R. Meddis, and M. J. Hewitt, "Modeling the Identification of Concurrent Vowels with Different Fundamental Frequencies, *J. Acoust. Soc. Am.*, Vol. 91, pp. 233–245, 1992.

[73] M. N. Murthi, and B. D. Rao, "Minimum Variance Distortionless Response (MVDR) Modeling of Voiced Speech," In *Proc. ICASSP'97*, 1997.

[74] T. Nakajima and T. Suzuki, "Speech Power Spectrum Envelope (PSE) Analysis Based on the F0 Interval Sampling," *IEICE Technical Report*, Vol. SP86, No. 94, pp. 55–62, 1987 (in Japanese).

[75] T. Nakatani, M. Goto and H. G. Okuno, "Localization by Harmonic Structure and Its Application to Harmonic Sound Segregation," In *Proc. IEEE ICASSP'96*, pp. 653–656, 1996.

[76] K. Nishi, M. Abe, and S. Ando, "Multiple Pitch Tracking and Harmonic Segregation Algorithm for Auditory Scene Analysis," *SICE Trans.*, Vol. 34, No. 6, pp. 483–490, 1998 (in Japanese).

[77] A. V. Oppenheim and R. W. Schafer, "Homomorphic Analysis of Speech," *IEEE Trans. Audio Electroacoust.*, Vol. AU-16, No. 2, pp. 221–226, 1968.

[78] C. I. Parris, D. Wong, F. Chambon, "A Robust 2.4kb/s LP-MBE with Iterative LP Modeling," In *Proc. Eurospeech'95*, pp. 677–690, 1995.

[79] T. W. Parsons, "Separation of Speech from Interfering Speech by Means of Harmonic Selection," *J. Acoust. Soc. Am.*, Vol. 60, pp. 911–918, 1976.

[80] D. B. Paul, "The Spectral Envelope Estimation Vocoder," *IEEE Trans. Acoust., Speech, Signal Process.*, Vol. ASSP-29, pp. 786–794, 1981.

[81] B. G. Quinn, "Estimating Frequency by Interpolation Using Fourier Coefficients," *IEEE Trans. Signal Process.*, Vol. SP-42, pp. 1264–1268, 1994.

[82] B. G. Quinn, "Estimation of Frequency, Amplitude, and Phase from the DFT of a Time Series," *IEEE Trans. Signal Process.*, Vol. SP-45, No. 3, pp. 814–817, 1997.

[83] D. C. Rife and R. R. Boorstyn, "Single-Tone Parameter Estimation from Discrete-Time Observations," *IEEE Trans. Info. Theory*, Vol. IT-20, No. 5, pp. 591–598, 1974.

[84] D. C. Rife and R. R. Boorstyn, "Multiple Tone Parameter Estimation from Discrete-Time Observations," *Bell System Technical Journal*, Vol. 55, No. 9, pp. 1389–1410, 1976.

[85] A. Röbel and X. Rodet, "Efficient Spectral Envelope Estimation and Its Application to Pitch Shifting and Envelope Preservation," In *Proc. of the 18th Int. Conf. on Digital Audio Effects (DAFx05)*, pp. 30–35, 2005.

[86] N. Roman and D. Wang, "Binaural Sound Segregation for Multisource Reverberant Environments," In *Proc. ICASSP2004*, pp. 373–386, 2004.

[87] T. Saikachi, K. Matsumoto, S. Sako and S. Sagayama, "Speech Analysis and Synthesis Based on Composite Wavelet Model," *Technical Report of IEICE*, Vol. 105, No. 372, pp. 1–6, in Japanese, 2005.

[88] S. Saito, H. Kameoka, N. Ono, and S. Sagayama, "POCS-based Common Harmonic Structure Estimation for Specmurt Analysis," *IPSJ SIG Technical Report*, 2006-MUS-65, pp. 13–18, 2006.

[89] S. Saito, H. Kameoka, N. Ono, and S. Sagayama, "Iterative Multipitch Estimation Algorithm for MAP Specmurt Analysis," *IPSJ SIG Technical Report*, 2006-MUS-66, pp. 85–92, 2006 (in Japanese).

[90] L. K. Saul, F. Sha, and D. D. Lee, "Statistical Signal Processing with Nonnegativity Constraints," In *Proc. Eith European Conference on Speech Communication and Technology*, Vol. 2, pp. 1001–1004, 2003.

[91] H. Sawada, R. Mukai, S. Araki and S. Makino, "Estimating the Number of Sources Using Independent Component Analysis," *Acoustical Science and Technology, the Acoustical Society of Japan*, Vol. 26, No. 5, 2005.

[92] H. Sawada, R. Mukai, S. Araki and S. Makino, "A Robust and Precise Method for Solving the Permutation Problem of Frequency-Domain Blind Source Separation," *IEEE Trans. Speech and Audio Process.*, Vol. 12, No. 5, pp. 530–538, 2004.

[93] G. Schwarz, "Estimating the Dimension of a Model," Annals of Statistics, 6, pp. 461–464, 1978.

[94] F. Sha and L. K. Saul, "Real-time Pitch Determination of One or More Voices by Nonnegative Matrix Factorization," In *Proc. NIPS*, pp. 1233–1240, 2004.

[95] F. Sha, and L. K. Saul, "Real-time Pitch Determination of One or More Signals by Nonnegative Matrix Factorization," In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Processing Systems 17*, MIT Press, Cambridge, MA, 2005.

[96] J. L. Shanks, "Recursion Filters for Digital Signal Processing," Geophysics, Vol. 32, pp. 33–51, 1967.

[97] P. Smaragdis, "Discovering Auditory Objects through Non-negativity Constraints," In *ISCA Tutorial and research workshop on statistical and perceptual audio processing – SAPA2004*, 2004.

[98] J. O. Smith III and X. Serra, "PARSHL: A Program for the Analysis/Synthesis of Inharmonic Sounds Based on a Sinusoidal Representation," In *Proc. ICMC'87*, pp. 290–297, 1987.

[99] D. Starer and A. Nehorai, "Newton Algorithms for Conditional and Unconditional Maximum Likelihood Estimation of the Parameters of Exponential Signals in Noise," *IEEE Trans. Acoust., Speech, Signal Process.*, Vol. ASSP-40, No. 6, pp. 1528–1534, 1992.

[100] K. Steiglitz, "On the Simulateous Estimation of Poles and Zeros in Speech Analysis," *IEEE Trans. Acoust., Speech, Signal Process.*, Vol. ASSP-25, No. 3, pp. 229–234, 1977.

[101] R. J. Stubbs, and Q. Summerfield "Algorithms for Separating the Speech of Interfering Talkers: Evaluations with Voiced Sentences, and Normal-hearing and Hearing-impaired Listeners," *J. Acoust. Soc. Am.*, Vol. 87, pp. 359–372, 1990.

[102] K. Takahashi, T. Nishimoto, and S. Sagayama, "F0 Multi-Pitch Analysis Using Deconvolution of Log-Frequency Spectrum," *IPSJ SIG Technical Report*, 2003-MUS-53, pp. 61–66, 2003 (in Japanese).

[103] F. J. Theis, C. G. Puntonet and E. W. Lang, "A Histogram-based Overcomplete ICA Algorithms," In *Proc. ICA2003*, pp. 1071–1076, 2003.

[104] D. J. Thomson, "Spectrum Estimation and Harmonic Analysis," In *Proc. IEEE*, Vol. 70, No. 9, pp. 1055–1096, 1982.

[105] T. Tolonen, and M. Karjalainen, "Computationally Efficient Multiplitch Analysis Model," *IEEE Trans. Speech, Audio Process.*, Vol. 8, pp. 708–716, 2000.

[106] M. Unoki, and M. Akagi, "A Method of Extracting the Harmonic Tone from Noisy Signal Based on Auditory Scene Analysis," *IEICE Trans.*, Vol. J82-A, No. 10, pp. 1497–1507, 1999 (in Japanese).

[107] T. Virtanen, and A. Klapuri, "Separation of Harmonic Sounds Using Multipitch Analysis and Iterative Parameter Estimation," In *Proc. IEEE WASPAA*, pp. 83–86, 2001.

[108] T. Virtanen, and A. Klapuri, "Separation of Harmonic Sounds Using Linear Models for the Overtone series," In *Proc. IEEE ICASSP2002*, Vol. 2, pp. 1757–1760, 2002.

[109] T. Virtanen, "Sound Source Separation Using Sparse Coding with Temporal Continuity Objective," In *ICMC2003*, pp. 231–234, 2003.

[110] T. Virtanen, "Separation of Sound Sources by Convolutive Sparse Coding," In *ISCA Tutorial and research workshop on statistical and perceptual audio processing – SAPA 2004*, 2004.

[111] M. Weintraub, "A Theory and Computational Model of Auditory Menaural Sound Separation," Ph.D. thesis, Stanford University, 1985.

[112] P.J. Walmsley, S. Godsill, and P. J. W. Rayner, "Bayesian Graphical Models for Polyphonic Pitch Tracking," In Diderot Forum, pp. 1–26, Vienna, 1999.

[113] M. Wu, "Pitch Tracking and Speech Enhancement in Noisy and Reverberant Environments," Ph.D. thesis, Ohio State University, 2003.

[114] M. Wu, D. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," In *Proc. IEEE ICASSP'02*, Vol. 1, pp. 369–372, 2002.

[115] M. Wu, D. L. Wang and G. J. Brown, "A Multipitch Tracking Algorithm for Noisy Speech," *IEEE Trans., Speech, Audio Process.*, Vol. 11, pp. 229–241, 2003.

[116] C. Yeh, A. Roebel, and X. Rodet, "Multiple Fundamental Frequency Estimation of Polyphonic Music Signals," In *Proc. ICASSP2005*, Vol. 3, pp. 225–228, 2005.

[117] Ö. Yilmaz and S. Rickard, "Blind Separation of Speech Mixtures via Time-Frequency Masking," *IEEE Trans. Signal Process.*, Vol. 52, No. 7, 2004.

[118] Y. V. Zakharov and T. C. Tozer, "Frequency Estimator with Dichotomous Search of Periodogram Peak," *IEEE Electronic Letters*, Vol. 35, No. 19, pp. 1608–1609, 1999.

[119] P. Zolfaghari, S. Watanabe, A. Nakamura and S. Katagiri, "Bayesian Modelling of the Speech Spectrum Using Mixture of Gaussians," In *Proc. ICASSP 2004*, Vol. 1, pp. 553–556, 2004.

[120] "Edinburgh Speech Tools Library," `http://www.cstr.ed.ac.uk/`.

[121] "Speech Filing System," `http://www.phon.ucl.ac.uk/resource/sfs/`.

Appendix A

List of Publications

Journal papers

[J1] Hirokazu Kameoka, Takuya Nishimoto, Shigeki Sagayama, "A Multipitch Analyzer Based on Harmonic Temporal Structured Clustering," *IEEE Transactions on Audio, Speech and Language Processing*, in Press, 2007.

[J2] Jonathan Le Roux, Hirokazu Kameoka, Nobutaka Ono, Alain de Cheveigné and Shigeki Sagayama, "Single and Multiple Pitch Contour Estimation through Parametric Spectrogram Modeling of Speech in Noisy Environments," *IEEE Transactions on Audio, Speech and Language Processing*, in Press, 2007.

[J3] Shigeki Sagayama, Haruto Takeda, Kameoka Hirokazu, Takuya Nishimoto, "Music Information Processing Using Speech Recognition Techniques," in J. Acoust. Soc. Jpn., vol.6, No.8, pp. 454–460, Aug, 2005. (in Japanese)

International Conferences

[C1] Hirokazu Kameoka, Takuya Nishimoto, Shigeki Sagayama, "Extraction of Multiple Fundamental Frequencies from Polyphonic Music Using Harmonic Clustering," In *Proc. 18th International Congress on Acoustics (ICA2004)*, in CD-ROM, 2004.

[C2] Hirokazu Kameoka, Takuya Nishimoto, Shigeki Sagayama, "Accurate F0 Detection Algorithm for Concurrent Sounds Based on EM Algorithm and Information Criterion," In *Proc. Special Workshop in MAUI (SWIM)*, in CD-ROM, 2004.

[C3] Hirokazu Kameoka, Takuya Nishimoto, Shigeki Sagayama, "Multi-pitch Detection Algorithm Using Constrained Gaussian Mixture Model and Information Criterion for Simultaneous Speech," In *Proc. Speech Prosody (SP2004)*, pp. 533–536, 2004.

[C4] Hirokazu Kameoka, Takuya Nishimoto, Shigeki Sagayama, "Separation of Harmonic Structures Based on Tied Gaussian Mixture Model and Information Criterion for Concurrent Sounds," In *Proc. IEEE, International Conference on Acoustics, Speech and Signal Processing (ICASSP2004)*, Vol. 4, pp. 297–300, 2004. (awarded the Best Student Paper Award)

[C5] Hirokazu Kameoka, Takuya Nishimoto, Shigeki Sagayama, "Multi-Pitch Trajectory Estimation of Concurrent Speech Based on Harmonic GMM and Nonlinear Kalman Filtering," In *Proc. Interspeech2004 International Conference on Spoken Language Processing (ICSLP2004)*, in CD-ROM, 2004.

[C6] Shigeki Sagayama, Keigo Takahashi, Hirokazu Kameoka, Takuya Nishimoto, "Specmurt Anasylis: A Piano-Roll-Visualization of Polyphonic Music Signal by Deconvolution of Log-Frequency Spectrum," In *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing (SAPA2004)*, in CD-ROM, 2004.

[C7] Hirokazu Kameoka, Takuya Nishimoto, Shigeki Sagayama, "Audio Stream Segregation of Multi-Pitch Music Signal Based on Time-Space Clustering Using Gaussian Kernel 2-Dimensional Model," In *Proc. IEEE, International Conference on Acoustics, Speech and Signal Processing (ICASSP2005)*, Vol. 3, pp. 5–8, 2005. (Selected as a finalist of the Student Paper Contest)

[C8] Shigeki Sagayama, Hirokazu Kameoka, Shoichiro Saito, Takuya Nishimoto, "'Specmurt Anasylis' of Multi-Pitch Signals," In *Proc. IEEE-EURASIP, International Workshop on Nonlinear Signal and Image Processing (NSIP2005)*, in CD-ROM, 2005.

[C9] Hirokazu Kameoka, Takuya Nishimoto, Shigeki Sagayama, "Harmonic-temporal structured clustering via deterministic annealing EM algorithm for audio feature extraction," In *Proc. International Conference on Music Information Retrieval (ISMIR2005)*, pp. 115–122, 2005.

[C10] Shoichiro Saito, Hirokazu Kameoka, Takuya Nishimoto, Shigeki Sagayama, "Specmurt Analysis of Multi-Pitch Music Signals with Adaptive Estimation of Common Har-

monic Structure," In *Proc. International Conference on Music Information Retrieval (ISMIR2005)*, pp. 84–91, 2005.

[C11] Hirokazu Kameoka, Jonathan Le Roux, Nobutaka Ono, Shigeki Sagayama, "Speech Analyzer Using a Joint Estimation Model of Spectral Envelope and Fine Structure," In *Proc. Interspeech2006 International Conference on Spoken Language Processing (ICSLP2006)*, in CD-ROM, 2006.

[C12] Nobutaka Ono, Shoichiro Saito, Hirokazu Kameoka, Shigeki Sagayama, "Inverse Filter Analysis of Common Harmonic Structure on Specmurt Using Riemann's Zeta Function," In *Proc. 4th Joint meeting of ASA and ASJ*, in CD-ROM, 2006.

[C13] Jonathan Le Roux, Hirokazu Kameoka, Nobutaka Ono, Shigeki Sagayama, "Parametric Spectrogram Modeling of Single and Concurrent Speech with Spline Pitch Contour," In *Proc. 4th Joint meeting of ASA and ASJ*, in CD-ROM, 2006.

[C14] Yuichiro Yonebayashi, Hirokazu Kameoka, Shigeki Sagayama, "Automatic Determination of Piano Fingering Based on a Hidden Markov Model," In *Proc. the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, to appear, 2007.

[C15] Kenichi Miyamoto, Hirokazu Kameoka, Haruto Takeda, Takuya Nishimoto, Shigeki Sagayama, "Probabilistic Approach to Automatic Music Transcription from Audio Signals," In *Proc. IEEE, International Conference on Acoustics, Speech and Signal Processing (ICASSP2007)*, to appear, 2007.

[C16] Jonathan Le Roux, Hirokazu Kameoka, Nobutaka Ono, Alain de Cheveigné, Shigeki Sagayama, "Harmonic-Temporal Clustering of Speech for Single and Multiple F0 Contour Estimation in Noisy Environments," In *Proc. IEEE, International Conference on Acoustics, Speech and Signal Processing (ICASSP2007)*, to appear, 2007.

Domestic Conferences (in Japanese)

[D1] Hirokazu Kameoka, Koichi Shinoda, Shigeki Sagayama, "Multi-Pitch Estimation of Natural Instruments Performance by using DP Matching in Frequency Domain," In *Proc. IPSJ SIG Technical Report*, 2002-MUS-46, pp. 17–22, 2002.

[D2] Hirokazu Kameoka, Koichi Shinoda, Shigeki Sagayama, "Multi-Pitch Analysis of Natural Instruments Performances using DP Matching in Spectrum Domain," In *Proc. ASJ Autumn Meeting*, 1-1-2, pp. 639–640, 2002.

[D3] Hirokazu Kameoka, Takuya Nishimoto, Koichi Shinoda, Shigeki Sagayama, "Multipitch Estimation using Harmonic Clustering," In *Proc. ASJ Spring Meeting*, 3-7-3, pp. 837–838, 2003.

[D4] Hirokazu Kameoka, Takuya Nishimoto, Koichi Shinoda, Shigeki Sagayama, "Multi-Pitch Estimation Using Harmonic Clustering," In *Proc. IPSJ SIG Technical Report*, 2003-MUS-50, pp. 27–32, 2003.

[D5] Hirokazu Kameoka, Takuya Nishimoto, Shigeki Sagayama, "Estimation of Number of Sound Sources and Octave Position in Multi-Pitch Extraction Using Harmonic Clustering," In *Proc. IPSJ SIG Technical Report*, 2003-MUS-51, pp. 29–34, 2003.

[D6] Hirokazu Kameoka, Takuya Nishimoto, Shigeki Sagayama, "Estimation of Number of Sound sources and Octave Positions in Multipitch Extration using Harmonic Clustering," In *Proc. ASJ Autumn Meeting*, 1-1-2, pp. 639–640, 2003.

[D7] Hirokazu Kameoka, Takuya Nishimoto, Shigeki Sagayama, "F0 Contours Detection Based on Constrained Gaussian Mixture Model and Akaike Information Criterion for Simultaneous Utterances," In *Proc. IPSJ SIG Technical Report*, 2003-SLP-49, pp. 229–234, 2003.

[D8] Hirokazu Kameoka, Takuya Nishimoto, Shigeki Sagayama, "F0Tracking Simultaneous Speech Based on Maximum A Posteriori Estimation for Constrained Gaussian Mixture Model," In *Proc. ASJ Spring Meeting*, 2-7-6, pp. 275–276, 2004.

[D9] Seiya Oda, Hirokazu Kameoka, Takuya Nishimoto, Shigeki Sagayama, "Spectral Detection of Inharmonic Sound Sources in Mixed Sound based on Extended Harmonic Clustering using Maximum A Posteriori Estimation," In *Proc. ASJ Spring Meeting*, 2-9-3, pp. 689–690, 2004.

[D10] Hirokazu Kameoka, Takuya Nishimoto, Shigeki Sagayama, "Fundamental Frequency Detection and Spectral Separation of Mixed Sound Based on Harmonic-GMM and Information Criterion," In *Proc. the 66th National Convention of IPSJ*, 3ZA-1, pp. 2-427–2-428, 2004.

[D11] Hirokazu Kameoka, Shoichiro Saito, Takuya Nishimoto, Shigeki Sagayama, "Recursive Estimation of Quasi-Optimal Common Harmonic Structure Pattern for Specmurt Anasylis: Piano-Roll-Display Visiualization and MIDI Conversion of Polyphonic Music Signal" In *Proc. IPSJ SIG Technical Report*, 2004-MUS-56, pp. 41–48, 2004.

[D12] Hirokazu Kameoka, Shoichiro Saito, Takuya Nishimoto, Shigeki Sagayama, "Automatic Determination of the Common Harmonic Structure Pattern in Specmurt Method for Pitch Visualization of Music Signals," In *Proc. ASJ Autumn Meeting*, 2-6-15, pp. 803–804, 2004.

[D13] Shigeki Sagayama, Haruto Takeda, Hirokazu Kameoka, Takuya Nishimoto, "Music Information Processing Viewed from Speech Recognition," In *Proc. ASJ Autumn Meeting*, 2-6-9, pp. 785–788, 2004.

[D14] Hirokazu Kameoka, Takuya Nishimoto, Shigeki Sagayama, "Time-Space Clustering for Multi-pitch Spectral Segregation Using Kernel Audio Stream Model," In *Proc. ASJ Spring Meeting*, 3-7-19, pp. 601–602, 2005.

[D15] Hirokazu Kameoka, Shigeki Sagayama, "Tutorial: EM Algorithm and Its Application to Multipitch Analysis," In *Proc. ASJ Technical Report on Musical Acoustics*, 2005.

[D16] Hirokazu Kameoka, Takuya Nishimoto, Shigeki Sagayama, "Harmonic-Temporal-structured Clustering (HTC) for Simultaneous Estimation of Audio Features in Music," In *Proc. IPSJ SIG Technical Report*, 2005-MUS-61-12, pp. 71–78, 2005.

[D17] Hirokazu Kameoka, Nobutaka Ono, Shigeki Sagayama, "Composite Function Model of Spectral Envelope and Harmonic Structure for Speech Analysis," In *Proc. ASJ Autumn Meeting*, 2-6-4, pp. 265–266, 2005.

[D18] Hirokazu Kameoka, Takuya Nishimoto, Shigeki Sagayama, "Music Signal Analysis based on Harmonic-Temporal-structured Clustering using Deterministic Annealing EM Algorithm," In *Proc. ASJ Autumn Meeting*, 3-10-16, pp. 769–770, 2005.

[D19] Jonathan Le Roux, Hirokazu Kameoka, Nobutaka Ono, Shigeki Sagayama, "Harmonic Temporal Clustering of Speech Spectrum," In *Proc. ASJ Spring Meeting*, 2-11-3, pp. 307–308, 2006.

[D20] Nobutaka Ono, Shoichiro Saito, Hirokazu Kameoka, Shigeki Sagayama, "Inverse Filter Analysis of Common Harmonic Structure on Specmurt by Using Riemann's ζ function," In *Proc. ASJ Autumn Meeting*, 1-5-25, pp. 555–556, 2006.

[D21] Shoichiro Saito, Hirokazu Kameoka, Nobutaka Ono, Shigeki Sagayama, "POCS-based Common Harmonic Structure Estimation Algorithm for Specmurt Analysis and Discussions on Its Covergence," In *Proc. ASJ Autumn Meeting*, 1-5-24, pp. 553–554, 2006.

[D23] Yuichiro Yonebayashi, Hirokazu Kameoka, Shigeki Sagayama, "Automatic Determination of Piano Fingering Using HMM," In *Proc. ASJ Autumn Meeting*, 3-2-6, pp. 727–728, 2006.

[D24] Shoichiro Saito, Hirokazu Kameoka, Nobutaka Ono, Shigeki Sagayama, "POCS-based Common Harmonic Structure Estimation for Specmurt Analysis," In *Proc. IPSJ SIG Technical Report*, 2006-MUS-65, pp. 13–18, 2006.

[D25] Yuichiro Yonebayashi, Hirokazu Kameoka, Shigeki Sagayama, "Automatic Determination of Piano Fingering based on Hidden Markov Model," In *Proc. IPSJ Technical Report*, 2006-MUS-65 , pp. 7–12, 2006.

[D26] Hirokazu Kameoka, Masataka Goto, Shigeki Sagayama, "Selective Amplifier of Periodic and Non-periodic Components in Concurrent Audio Signals with Spectral Control Envelopes," In *Proc. IPSJ SIG Technical Report*, 2006-MUS-66, pp. 77–84, 2006.

[D27] Shoichiro Saito, Hirokazu Kameoka, Nobutaka Ono, Shigeki Sagayama, "Iterative Multipitch Estimation Algorithm for MAP Specmurt Analysis," In *Proc. IPSJ SIG Technical Report*, 2006-MUS-66, pp. 85–92, 2006.

[D28] Ken-ichi Miyamoto, Hirokazu Kameoka, Haruto Takeda, Takuya Nishimoto, Shigeki Sagayama, "Automatic Music Transcription Combining HTC Multipitch Analysis and HMM-based Rhythm and Tempo Estimation," In *Proc. ASJ Autumn Meeting*, 2-7-3, pp. 583–584, 2006.

[D29] Shoichiro Saito, Hirokazu Kameoka, Nobutaka Ono, Shigeki Sagayama, "Multipitch Estimation of Music Audio Signals through MAP Specmurt Analysis," In *Proc. ASJ Autumn Meeting*, 2-7-2, pp. 581–582, 2006.

[D30] Hirokazu Kameoka, Le Roux Jonathan, Nobutaka Ono, Shigeki Sagayama, "Harmonic Temporal Structured Clustering: A New Approach to CASA," In *Proc. ASJ Technical Report on Psychological and Physiological Acoustics*, Vol. 36, No. 7, H-2006-103, pp. 575–580, 2006.

[D31] Yosuke Izumi, Hirokazu Kameoka, Nobutaka Ono, Shigeki Sagayama, "Applying EM Algorithm to 2ch BSS Based on Sparseness of Speech," In *Proc. IEICE Technical Report on Electroacoustics*, Vol. 106, EA2006-96, pp. 43–48, 2006.

[D32] Hirokazu Kameoka, Nobutaka Ono, Shigeki Sagayama, "Parameter Optimization Algorithm for Sinusoidal Signal Model," In *Proc. IEICE Technical Report on Electroacoustics*, Vol. 106, EA2006-97, pp. 49–54, 2006.

[D33] Hirokazu Kameoka, Nobutaka Ono, Shigeki Sagayama, "A Parameter Optimization Algorithm for Sinusoidal Model and Its Application to 1ch Blind Source Separation," In *Proc. ASJ Spring Meeting*, to appear, 2007.

[D34] Nobutaka Ono, Yosuke Izumi, Hirokazu Kameoka, Shigeki Sagayama, "2ch BSS under Noisy Environments by ML Time-Frequency Masking with EM Algorithm," In *Proc. IEICE Annual Conference*, to appear, 2007.

Appendix B

Awards Received

1. Best Presentation Award at IPSJ SIGMUS Annual Symposium 2003
2. Best Presentation Award at IPSJ SIGMUS Annual Symposium 2004
3. Best Student Paper Award Finalist at IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2005
4. 20th Telecom System Technology Student Award from the Telecomunications Advancement Foundation (TAF)
5. Yamashita Memorial Research Award from IPSJ
6. Best Presentation Award at IPSJ SIGMUS Annual Symposium 2005
7. Itakura Innovative Research Encouragement Award from ASJ

Awards Received (in Japanese)

1. 情報処理学会 SIGMUS『夏のシンポジウム 2003』 ベストプレゼンテーション賞
2. 情報処理学会 SIGMUS『夏のシンポジウム 2004』 ベストプレゼンテーション賞
3. Best Student Paper Award Finalist at IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2005)
4. 第 20 回 電気普及財団テレコムシステム技術学生賞
5. 情報処理学会 平成 17 年度 山下記念研究賞
6. 情報処理学会 SIGMUS『夏のシンポジウム 2005』 ベストプレゼンテーション賞
7. 日本音響学会 第 2 回 独創研究奨励賞 板倉記念