TR-IT-0055

Modelling speaker-specific intonation characteristics

Hélène Valbret

1994.3

ABSTRACT
This report describes work carried out at ATR during my post-doctoral year, to model the speaker-specific characteristics of intonation using information extracted from the fundamental frequency of the speech waveform. It also describes an implementation of pitch-synchronous signal processing techniques for realisation of intonation in the CHATR speech synthesis system, and the tools developed for the extraction, analysis, and resynthesis of fundamental frequency information.

Chapter 1

Introduction

As computers become able to produce more and more high-quality synthesized speech, more studies have been devoted to speaker individuality and speaking styles. The naturalness of synthesized speech should be improved by introducing some individual characteristics, emotional effects, accents ... The generation of the CHATR speech synthesizer system ([Black & Taylor 94]) led us to study the speaker's characteristics.

Let us first briefly review the sources of speaker variability. Schematically, it can be attributed to 2 main factors.

- the speaking style :

  In learning speech mechanisms, each speaker, influenced by his dialect and social environment, has developed his own articulatory strategy. The choice of the vocabulary, the syntactic structure as well as articulatory and prosodic habits may vary greatly from one speaker to another. The choice of specific articulatory gesture will lead to different types of voice (palatalized, nasalized ...[Laver 80]) ; Prosodic habits include speaking rate, phonemic durations, intonation (accents), amplitudes ...

- the physiological characteristics :

  Roughly speaking, the production of speech involves 3 main structures. The respiratory system, the larynx and the vocal tract which involved both oral and nasal cavities. These 3 structures vary greatly from one speaker to another (dimensions, tissus, elasticity [Stevens 72]). The variations show up in the fundamental frequency (mean and range), the glottal excitation (regularity, harmonicity) - which produces different voice quality such as modal, laryngalized, breathy ([Karlsson 90], [Laver 80]) - as well as the spectral characteristics (formant dispersion ([Fant 60],[Fant 66])).

Whereas a few studies have now been dedicated to speaker spectral characteristics, as far I know, very few dealt with the prosodic individual characteritics. This study is thus concerned with this last topic and focusses on fundamental frequency.

Prosodic features are involving many different levels factored to language, style, speaker, etc. Each language has particular ways of structuring a sentence : phrasing

as well as accent positions are partly language dependant. However, phrase and accent characteristics are probably speaker dependant : timing (location of accent event (rise or fall) in the syllable), amplitude, accent shape ... We propose two different approaches to separating the different components.

- We first used a pitch model. This allowed us to represent a pitch contour with only a few parameters related to the utterance phonological description. The models we choose to examine are described in chapter 3. Chapter 5 will present an evaluation of them.

- We then tried a second approach comparing the observed pitch contour of the same sentence uttered by different speakers. The difference between both contours is then language independant. However speakers intentions, while uttering the sentence, may still be different. This approach will be presented in chapter5

This document will briefly describe some techniques and algorithms we developped and used for this work. The first two chapters will concern the analysis techniques whereas the third one presents the analysis-synthesis system based on TD_PSOLA (Time Domain Pitch Synchronous OverLap and Add [Moulines & Charpentier 90]) which can be used for speech synthesis as well as perceptual evaluation. the following chapter will describe to the experiments we carried out to evaluate the approaches we proposed.

Chapter 2

Pitch Tracking and Pitch-Synchronous Waveform Marking

The first step of any study of prosodic features is the extraction of the fundamental frequency. We will describe in this chapter the procedure we implemented to deal with this task. As said earlier, we not only want to analyse the different intonation patterns but also would like to modify them. Therefore, we implemented the PSOLA (Pitch-Synchronous-Overlap-and-Add [Moulines & Charpentier 90]) approach which allows to perform pitch and time scaling without voice quality degradation. Such a technique needs not only the detection of the fundamental frequency but also the computation of a sequence of marks aligned pitch-synchronously on the signal waveform. The second part of this chapter will be devoted to this "localisation". In the third part we will briefly indicate how to use our program and we will show some examples of results in the 4th part as well as some weakness of our algorithm.

2.1 Pitch Tracking

Pitch-tracking is not a trivial task and many approaches have been reported and are still studied ([Hess 83], [Medan *et al* 91], [Bagshaw *et al* 93],...).

Our pitch-tracking procedure is based on three different criteria which are unfortunately dependant upon different thresholds. Two thresholds limit the range of the pitch values : the range is usually set to [50,250] Hz for male speakers and [150,400] Hz for female.

- First, the zero-crossing rate is determined : this rate is usually higher for non-voiced signals than for voiced one. The speech signal is then low-passed filtered in order to attenuate higher-harmonic components of the signal.
- Silence and speech are descriminated using an energy criterium. A level threshold is thus needed and may depend on the record conditions.

- The short-term autocorrelation coefficient is computed according to the following formula :

$$A_m(d) = \frac{\sum_{i=m}^{m=N-1} x(i)x(i-d)}{\sqrt{\sum_{i=m}^{m+N-1}[x(i)]^2 \sum_{i=m}^{m+N-1}[x(i-d)]^2}}$$

  It measures the similitude between two segments of signal spaced d samples apart. Since voiced-speech has a quasi-periodic structure, this coefficient will present a maximum for a decay d equal to the pitch period. Therefore, the autocorrelation coefficient is computed for each value d of the pitch-period range. The 5 highest local maxima are then detected and kept in memory as a potential candidate. However, a threshold is used to reject the "too" small maxima (threshold as a percentage of the higher maximum).

We then elaborated a "probability" measure from the different values computed above. We defined it as follows :

$$\begin{aligned}
P_z(s) &= \frac{zero_thres}{zero_crossing(s)} \\
P_e(s) &= (1 + e^{-0.3*(energy(s)-ener_thres)})^{-1} \\
P_a(s) &= autocorr(s) \\
P_{voice}(s) &= P_z(s) * P_e(s) * P_a(s)
\end{aligned}$$

A speech segment is then detected as voiced for voicing probability above a threshold (typically 0.5), as unvoiced elsewhere. For the time being, we just used the first "autocorrelation candidate" to define the probability measure. Later, the other candidates could be used to eliminate pitch-tracking error.

A smoothing is then applied on the extracted pitch contour.

2.2 Pitch-Synchronous Marking

The signal marking uses the results of the pitch detection algorithm. On non-voiced portion the marks are apposed uniformly on the waveform, typically every 10 ms. On voiced portions, the marks are pitch-synchronous. The procedure can be decomposed into 3 steps for each voiced portion :

- The voiced segment is delimited (begin and end) according to the voiced/non-voiced decision coefficient.

- The major waveform peak in the middle of the voiced segment is then detected. Negative or positive peak can be choosen according to the will of the user ; ideally the marks should correspond to the glottal closure. The first mark is apposed on this peak.

- All the other marks of the voiced segment will then be derived recursively from the first one. Both the pitch information and a peak detection algorithm are used to determine the location of each new mark.

2.3 Manual

The program, called "tracker", allows both pitch-tracking and signal marking. Different options have to be specified :

- -i : input signal file ; the signal file is a raw file (binary short).
- -o : output file(s) ; tracker creates 2 output files : the pitch contour and the mark files. There are 2 different potential formats : binary and text. Both binary and text files contain a short header.
  - In case of mark files, the header indicates the name of the input file and the signal sampling frequency (in kHz). The file contains 2 fields. The first one is the sample location of the mark : the difference between two consecutive marks locations is the local pitch value. The second field contains the voicing decision : 0 means non-voiced ; 1 is voiced.
  - In case of pitch files, the name of the input file is writen as well as the pitch computation rate (in ms). The pitch value in Hz is then following this header.
- -p : indicates that the user is just interested in pitch contour : the marking procedure (which is time-consumming) is not run.
- -m : only the mark output file is created. Note that in that case, the pitch tracking is performed anyway.
- -c : constant file. This constant file contains the different thresholds used in the tracking and marking procedures ; the input signal sampling frequency as well as the desired output format. However, this file is not necessary : if it is not found, some default values are used. Let us describe in detail which informations the constant files may contain.
  - Waveform sampling rate in kHz.
    SAMPLE_RATE
  - Output files type with 2 possibilities : "binary" or "ascii".
    F0_FILE_TYPE
    MK_FILE_TYPE
  - F0 tracking options.
    * F0 sample rate : frame interval between 2 successive pitch values (ms).
      FRAME_INTERVAL
    * Definition of pitch range : minimum and maximum pitch values are given (Hz).
      MIN_PITCH
      MAX_PITCH

    * Energy Threshold (to distinguish between silence and speech).
      ENERGY_THRES
    * Zero Crossing Threshold as a percentage of the length of the window
      ZERO_THRES
    * Smoothing Option : a double smoothing can be performed on the tracked pitch values. The number of points for each smoothing procedure can be fixed.
      · SMOOTHING (yes = 0; no = 1)
      · FIRST
      · SECOND

- Pitch Synchronous Marking options.
    * Marking period on non-voiced portion (ms).
      NONSYNCH_WIN
    * Pitch-synchronous marking on maxima or minima (cf 2.2).
      POSITION
    * Thresholds used in the recursive marking procedure of voiced segment.
      · Definition of a "variation interval" to take into account the imprecision of the pitch value : $f_0 \in [f_0(1-\epsilon_1), f_0(1+\epsilon_1]$.
        MIN_RANGE $(1-\epsilon_1)$
        MAX_RANGE $(1+\epsilon_1)$
      · 2 thresholds used in the search of local maxima in order to (try to) take into account the local perturbations of waveform.
        THRES_AMP
        THRES_DUR
      · 1 threshold limits the size of local peaks : too small peaks (at voiced segment boundaries) are neglected. "Small" is defined as a percentage of the central peak amplitude (cf 2.2).
        THRES_PEAK

Here follows an example of a standard constant file : the values given are the defaults. Note that each value has to be preceded by its specification (as follows) so that the file can be incomplete or randomly organized.

```
SAMPLE_RATE 12

F0_FILE_TYPE  ascii
MK_FILE_TYPE  ascii

FRAME_INTERVAL 5
MAX_PITCH  400
MIN_PITCH  70
```

```
ENERGY_THRES  600
SMOOTHING  0
FIRST  7
SECOND  9
ENERGY_THRES    50
ZERO_THRES      25

NONSYNCH_WIN  10
POSITION  0
THRES_PEAK  0.04
MIN_RANGE  0.5
MAX_RANGE  1.5
THRES_AMP  0.85
THRES_DUR  10.0
```

2.4 Some examples and Errors

This chapter presents a few examples of marked speech signals. The algorithm has been performed on Japanese and English, female and male voices. The first picture (2.1) illustrates the case of good performances whereas the two following show pitch tracking errors which have been found to happen much more often on one of our speaker (female english : Sally from CSTR database).

Let us review the typical errors.

- Pitch Tracking Errors

  In silent portion, a very hich pitch period is sometimes detected. That seems to happen when the energy threshold is not high enough. Note that in these cases, the normalized autocorrelation function presents a lot of local maxima whereas in voiced segment the number of maxima is usually limited to two (corresponding to the pitch and the double pitch values).

  At the beginning and end of voiced segment, fundamental frequency is usually over-estimated.

  Very few doubling pitch have been found. They are usually erased by the smoothing procedure. A kind of dynamic programming technique may replace avantageously the smoothing. In that case, the different pitch candidates provided by the autocorrelation criterium should be processed to find the "optimal pitch contour".

- Marking Errors

  The most common error is illustrated on the last figure (2.2), although the pitch value has been correctly determined. These errors can be partly recovered by limiting strongly the range of peak search. However, we didn't succeed in completely eliminate this kind of problemes.

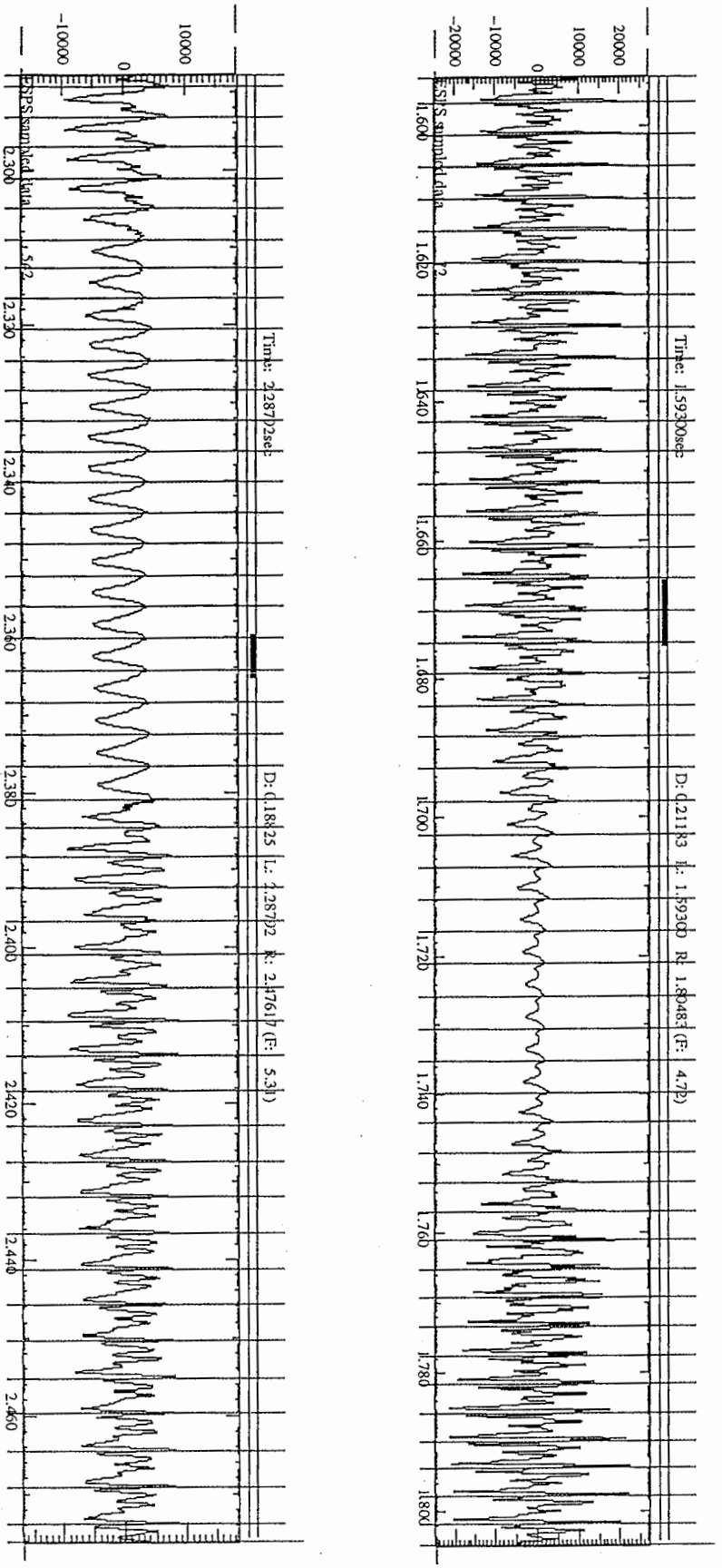

Figure 2.1: Pitch marking in case of success :2 english female speakers.

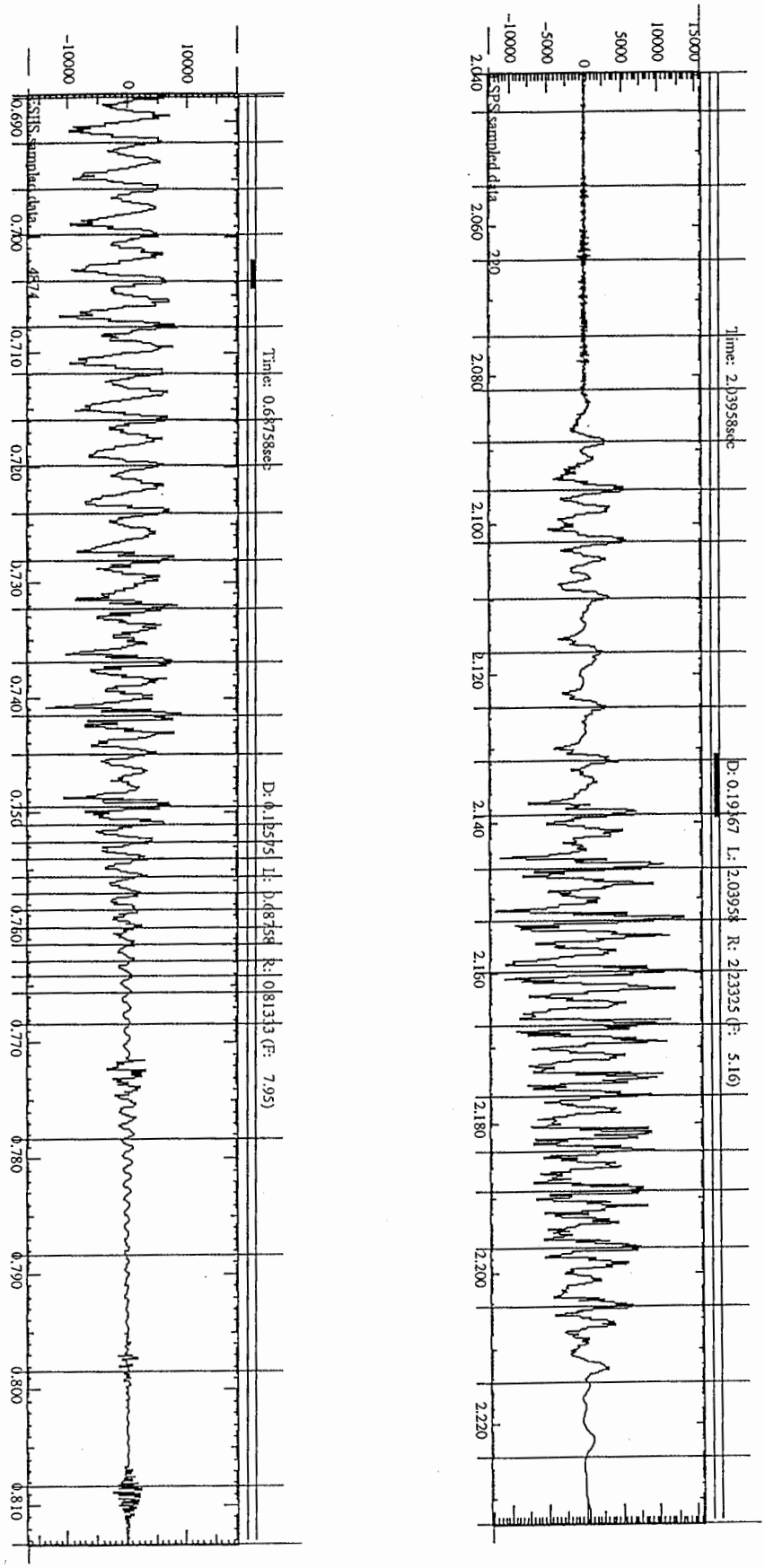

Figure 2.2: Pitch marking in case of failure :
upper part : male japanese speaker (wrong peak marking) ; lower part : english female speaker (wrong pitch detection).

Chapter 3

Pitch Models

The procedure described in the previous chapter provides us with a rough pitch-contour. These contours include the speaker characteristics as well as the phonogical characteristics. We would like to filter out the language specificity. In that order, our first idea consisted in using some kind of pitch model, which would provide a phonological description of the sentence under analysis while allowing an accurate synthesis of the pitch contour. The litterature provides us with many different pitch models but most of them concentrate in one aspect only (phonology [Pierrehumbert 80] ; accurate representation of pitch contour [Hirst *et al* 91]). Two models seem particularly well adapted to our purpose : the Fujisaki Model ([Fujisaki 91]) and the RFC Model ([Taylor 93]). Let us briefly describe them.

3.1 Fujisaki Model

3.1.1 Formulation

Fujisaki model ([Fujisaki & Hirose 84], [Fujisaki 91]) considers the pitch-contour as a superposition of a sequence of phrase components and a sequence of accent components. More precisely, phrase components are the response of a critically damped second-order linear system to a set of impulses whereas accent components are the response of another critically-damped second order linear system to a set of stepwise functions. These two components are described by the following set of equations :

$$G_{p_i}(t) = \begin{cases} \alpha_i^2 t e^{-\alpha_i t} & \text{if } t >= 0 \\ 0 & \text{otherwise} \end{cases}$$

$$G_{a_j}(t) = \begin{cases} Min(1-(1+\beta_j t)e^{-\beta_j t}, \theta) & \text{if } t >= 0 \\ 0 & \text{otherwise} \end{cases}$$

where α_i and β_j are the natural angular frequency of the phrase control mechanism to the i_{th} phrase command and the accent control mechanism to the j_{th} accent command respectively and θ the ceiling level of the accent component.

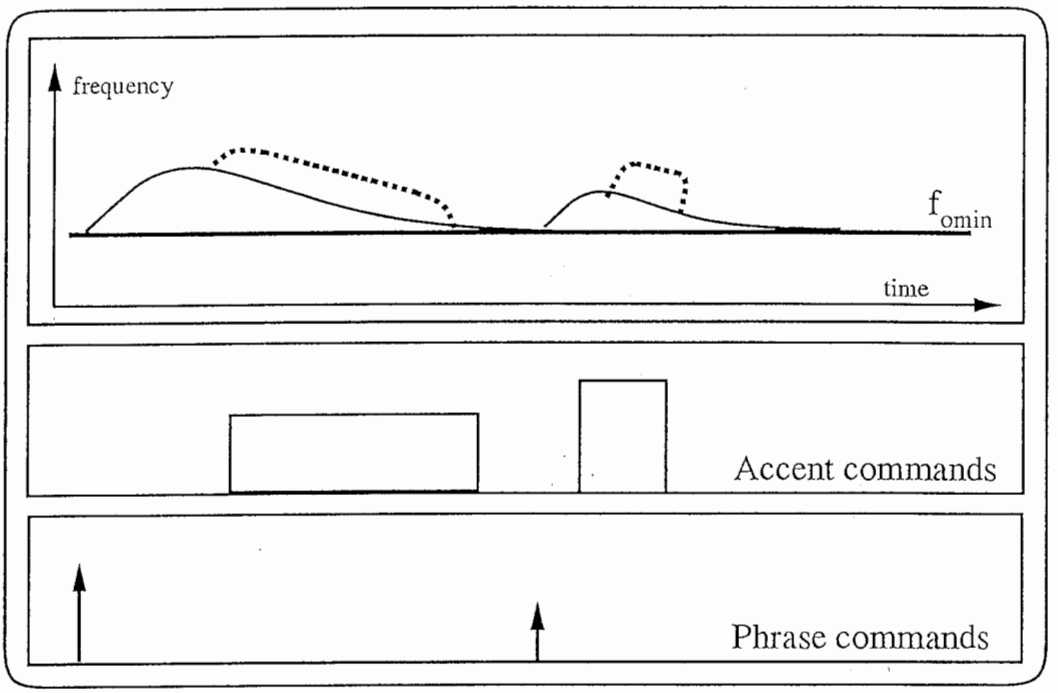

Figure 3.1: Fujisaki Model
pitch contour viewed as the superposition of phrase and accent components

The pitch-contour is then derived from these two components by the relation :

$$ln(F_0(t)) = ln(F_{min}) + \sum_{i=1}^{P} A_{p_i} G_{p_i}(t - T_{o_i}) + \sum_{i=1}^{A} A_{a_i}[G_{a_i}(t - T_{1_i}) - G_{a_i}(t - T_{2_i})]$$

F_{min} is the asymptotic value of the fundamental frequency in the absence of accent components, P is the number of phrase commands and A the number of accent commands, A_{p_i} A_{a_i} are the amplitude of the i_{th} phrase command and the i_{th} accent command respectively ; T_{o_i} is the timing of the i_{th} phrase component and T_{1_i} and T_{2_i} are respectively the onset and offset of the i_{th} accent component.
This is illustrated in figure (3.1).

3.1.2 Advantages and Drawbacks

Such a model is very attractive for two main reasons.

- First, the model can decompose a whole sentence in a few parameters only which are more or less linked to a physiological theory ([Fujisaki 91]).
- The second reason is connected to our purpose. The model allows to control independantly accent prominences and pitch range. By increasing the amplitude of the phrase component, one can raise the pitch range ; by varying the amplitude of the i_{th} step input, one can change the pitch of the i_{th} accent peak only.

Unfortunately, the model has also some drawbacks.

- The analysis of the contour in only a few parameters assumes some very strong constraints on the accent patterns which are not always corresponding to the real pitch-contour.

- Moreover, the estimation of accent and phrase parameters are not yet fully automatic and needs usually, as a first step, the detection of the different events (accent step functions or phrase impulses). Some attempts are under study to perform this analysis. Geoffrois's work ([Geoffrois 93]) is based on least squares parameter optimization and on prosodic event detection. Hirai's technique ([Hirai *et al* 93] uses the linguistic and synctactic information to derive the different events ; accent commands are classified according to their type (flat or declining) and the previous accent type (6 classes) ; phrase commands are divided into 3 groups by syntactic structure. Amplitudes of the quantized commands are then optimized using a large speech database for training data.

3.1.3 Algorithm

We developed an algorithm to derive Fujisaki's parameters. This work is related to Hirai's study. We assumed that events location s are determined by the linguistic and syntactic information. Note that the Japanese prosody has been well studied in the literature ([Sagisaka & Sato 86], [Pierrehumbert & Beckman 88], [Kubozono 93] and seems to be strongly defined by this information.

The determination of Fujisaki's parameters is an optimization problem. The criterium chosen here is the mean-squared error between the observed pitch-contour and the estimated one. This criterium is given by the following equation :

$$E = \sum_n [ln(F_0(n)) - ln(\hat{F}_0(n))]^2$$

where $F_0(n)$ and $\hat{F}_0(n)$ are respectively the observed sampled pitch value and the estimated pitch at time n.

- A first attempt : fully optimized amplitudes

  The partial derivations of this equation provides us with a set of equations linear in amplitude and non-linear otherwise (the linear equations are derived in Appendix). Our first idea consisted in using a gradient method to optimize the non-linear parameters (timing and angular frequencies) and to compute the optimized amplitudes via singular value decomposition. The input to the optimization algorithm is then the sampled pitch-contour and the number of phrases and accents.

  The synthesized pitch thus obtained is a very good approximation of the pitch contour. Unfortunately the accents and phrases timing and amplitudes are no more meaningful : phrase impulse command in the middle of an accent, negative amplitudes for both accents and phrases with very high $\hat{F}_{0_{min}}$,

very variable angular natural frequencies. It then appears necessary to add constraints on every parameters to maintain them "physically meaningful".

- A second attempt : optimization under constraints.

  The second algorithm is less powerful. Not only the accents and phrases numbers are needed but also their location in time. The natural angular frequencies are set to fixed values ($\alpha = 3$ and $\beta = 20$). The timing parameters are affined through an exhaustive research on a small interval centered around the input value. As phrases and accents magnitudes as well as $f_{0_{min}}$ are linked and may compensate one another, we choose to add some constraints on $f_{0_{min}}$ and phrases amplitudes : both should be lower than the observed pitch-contour in most cases (90 % for example).

  This second algorithm has the advantage to relate the different parameters to the "phonetic-syntaxic" input as only "meaningful" accents and phrases are considered. The main drawback is the necessity of an pre-processing which will allocates phrases and accents. An exhaustive research on the whol time domain under constraints would be time and cpu consumming (a few months for one sentence !?). We didn't focus on the pre-processing part and didn't deepen this approach but some work in ATR ([Hirai *et al* 93]) is still continuing.

3.1.4 Discussion

The figures (3.2) and (3.3) give some examples of results that may be obtained in the best and worst cases . The second example is particularly affected by the micro-prosody which is already attenuated by median smoothing. Because of our inability to derive a fully automatic Fujisaki model, we didn't pursue this approach. However, as mentioned above, some research is still dedicated to that approach. The last result was a mean estimation error (mean-squared error) of 30 Hz on our database (500 sentences uttered by a male japonese). The standard deviation of this error is quite important and the difference between the estimated pitch value and the observed pitch value may be as high as 80 Hz. We believe that this may be explained by the strong constraints applied on accent ad phrase shapes. However, Geoffrois' approach has been shown to be successful and fully automatic. Some experiments has been done on words. But as far as I know, it has not been yet proved to be "phonologically-meaningful" on sentences.

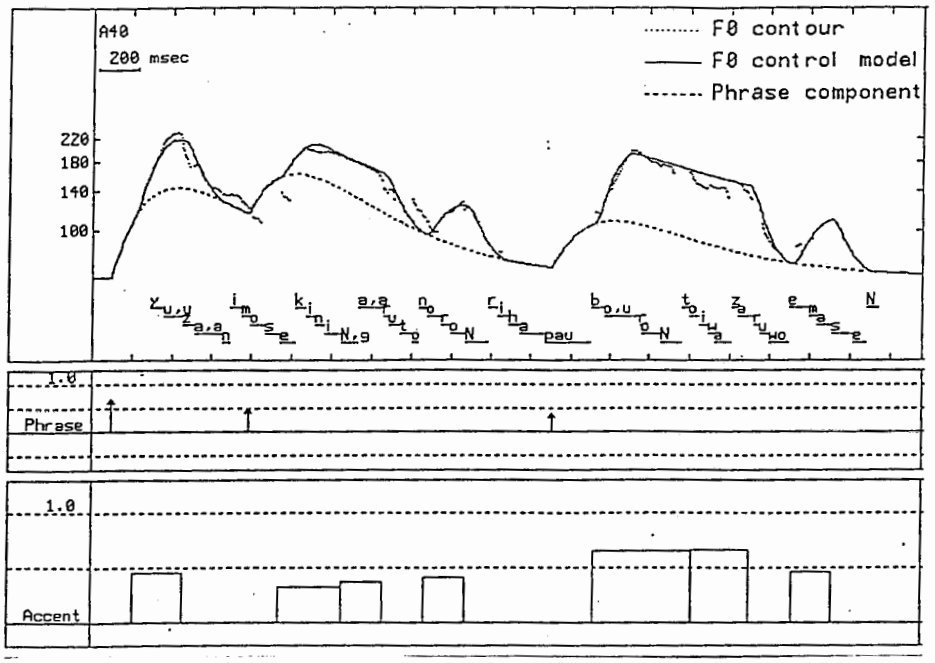

Figure 3.2: The Fujisaki model in case of good matching.

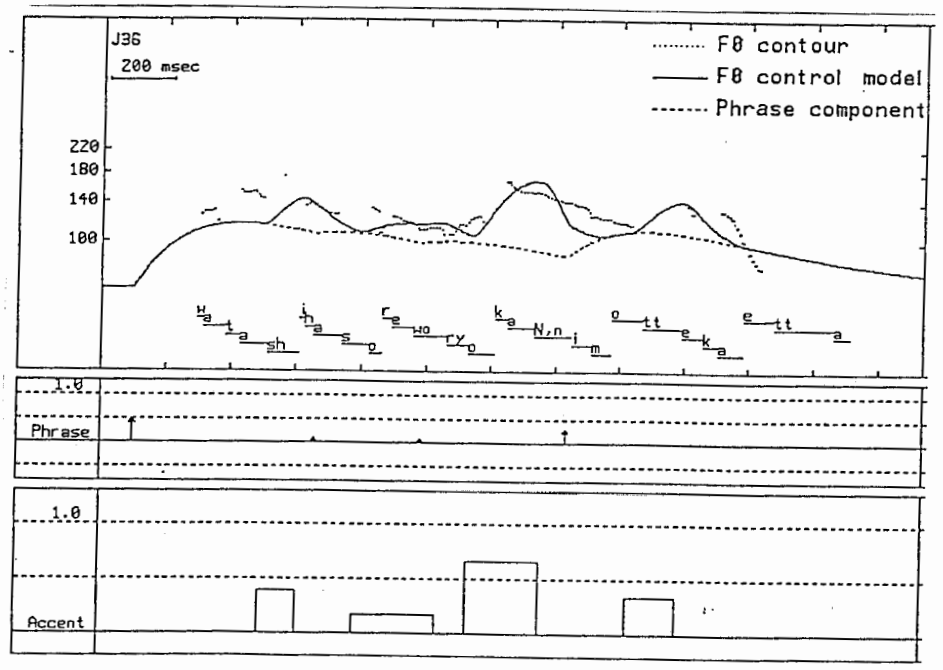

Figure 3.3: The Fujisaki model in case of failed matching.

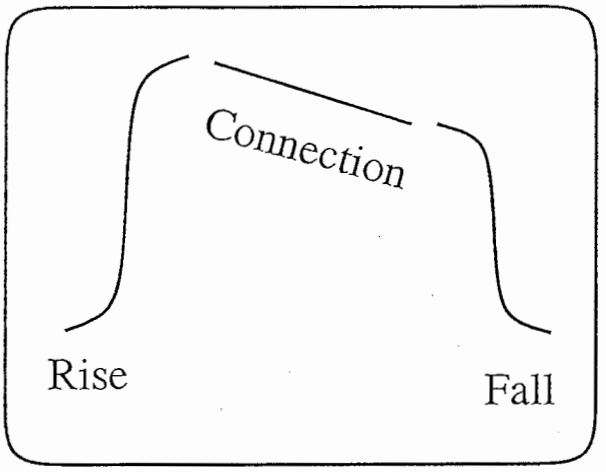

Figure 3.4: Rise, Fall and Connection Elements

3.2 Rise-Fall-Connection Model

As written above, the main drawback of the Fujisaki model is its implementation so that phrases and accents paramenters can be derived automatically. We thus investigated another model : the Rise-Fall-Connection model, proposed by P.Taylor ([Taylor 93]) which he implemented. We will briefly describe here his theory and his algorithm. More details can be found in ([Taylor 92]).

3.2.1 Formulation

The RFC model (which stands for Rise/Fall/Connection Model) decomposes the pitch contour into a sequence of rises, falls and connection elements. Rises and falls are described by a quadratic function according to the following equation :

$$f_0 = \begin{cases} A - 2A(t/D)^2 & 0 < t < D/2 \\ 2A(1 - t/D)^2 & D/2 < t < D \end{cases}$$

where A is the amplitude and D the duration of the rise or the fall element. A and D may have any value.

Every pitch contour can thus be encoded by an alternating succession of such elements linked by straight lines, i.e. the connection elements. The RFC elements are shown figure 3.4.

The underlying theory assumes that rises and falls occur only in case of pitch accent or boudary rise. Thus, the RFC description should allow the definition of the phonological contour of the sentence : each rise or fall is part of a phonological event.

3.2.2 Algorithm

Let us describe very briefly the algorithm developed by Taylor to estimate the RFC parameters. The algorithm involved two main stages :

- The first stage consists of detecting the main movements (rises and falls) of the pitch contour. This labelling module is dependant upon two thresholds ; a rise gradient threshold and a fall gradient one. Once the pitch contour has been re-sampled at 50 ms intervals, the pitch value of each frame is compared to the pitch value of the preceding one. If the difference exceeds the rise gradient threshold, the frame is classified as a rise ; if it is below the fall gradient, the frame is labelled as a fall ; elsewhere, the frame is left unlabelled. Adjacent frames described by the same label are then grouped together.

  This procedure usually identifies correctly the main movements of the pitch contour. However, in a few cases due to some segmental effects (sudden rise or drop in the pitch contour), some frames are mislabelled leading to a short-duration rise in the middle or a fall movement and reciprocally. Therefore, a post-processing has been added : it erases the "spurious" sections whose durations falls below a certain threshold.

- The second stage determines the durations and amplitudes of the rise/fall elements defined in section 3.2.1 which best approximate the movements detected during the first stage. This procedure first delimits a "search region" at the boundaries of each movement in which the rise/fall elements may start or end. A new set of threshold is used to define the size and position of these regions. Then every potential element which starts or ends in such areas are synthesized and compared to the original contour. The best shape, according to the mean-squared criterium, is then selected.

3.2.3 Advantages and Drawbacks

The RFC model is very simple and flexible. As amplitudes and durations of the different components may vary greatly, a large panel of different accent types can be approximated. Moreover, the model has been implemented to perform a fully automatic analysis.

However, as the model is very flexible and very loose, one has to be aware of the tendancy to perfectly match the pitch contour by adding "extra-components" which do not correspond to any phonological element. For example, obstruent perturbations, which causes sudden spikes and glitches, may be represented by short rises and falls. These kinds of errors are partly removed by the pre-processing (already mentioned above) of the raw pitch-contour as well as a training procedure (described in [Taylor 92]) which tunes the different thresholds defined above. However, this training procedure is not automatic and is very difficult to carry on.

The same examples presented for the Fujisaki model are shown in figure (3.5, 3.6).

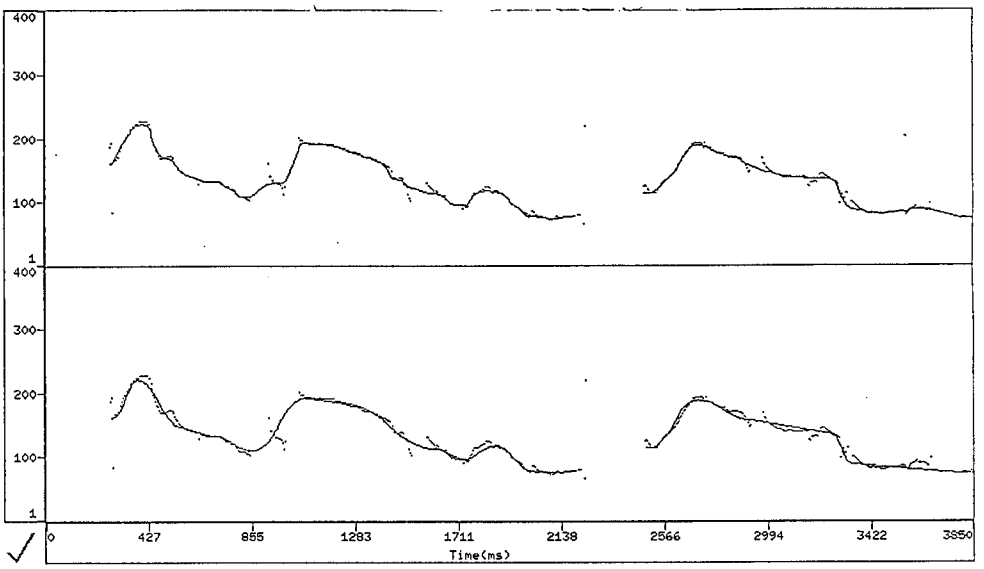

Figure 3.5: RFC modelisation for sentence : "Yuuzaa ni mo sekiniN ga aruto no roNri ha bouroN to iwazaru wo emasen".
upper part : observed pitch contour (dash) and smoothed contour (solid); bottom part : observed pitch contour (dash) and RFC synthesized pitch contour (solid)

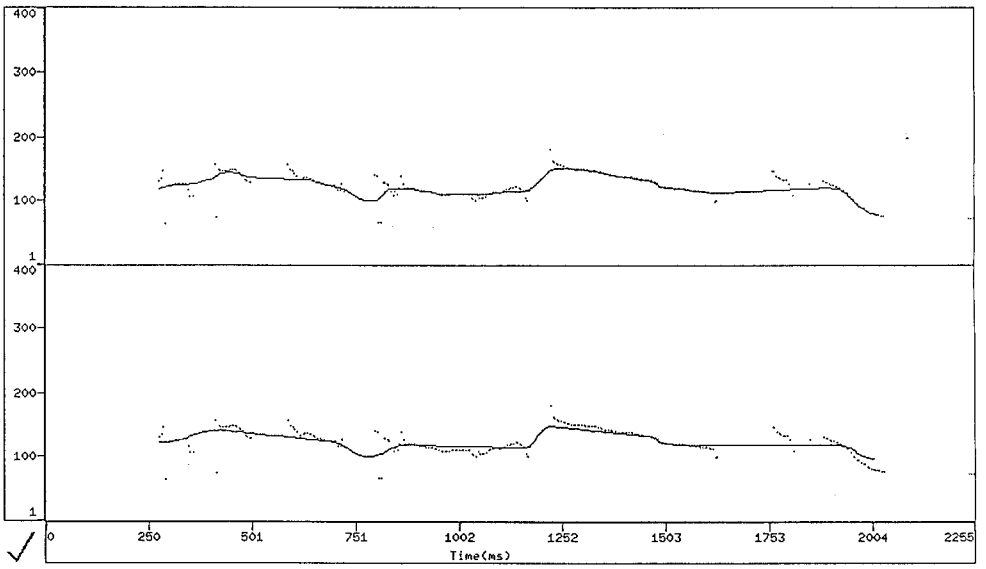

Figure 3.6: RFC modelisation for sentence : "Watashi ha sore wo ryokaN ni motte kaetta".
upper part : observed pitch contour (dash) and smoothed contour (solid); bottom part : observed pitch contour (dash) and RFC synthesized pitch contour (solid)

Some other experiments have been conducted on an English database. The results were not so good due to the fact that the pitch contour was very discontinuous : voiced portions were so short that the algorithm was not able to detect the different movements.

3.3 Appendix

Fujisaki's Problem : Parameters Optimization.

The problem consists in finding the parameters - phrases and accents timing T_{0_i}, T_{1_i} and T_{2_i}, natural angular frequency α_i and β_i, and amplitudes A_{p_i} A_{a_i} which minimize the following mean-square criterium :

$$E = \sum_n (F_0(n) - \hat{F}_0(n))^2$$

where
$ln(\hat{F}_0(t)) = ln(\hat{F}_{0_{min}}) + \sum_{i=1}^{P} A_{p_i} G_{p_i}(t - T_{o_i}) + \sum_{i=1}^{A} A_{a_i} H_{a_i}(t, T_{1_i}, T_{2_i})$ and

$$\begin{aligned} G_{p_i}(t) &= \begin{cases} \alpha_i^2 e^{-\alpha_i t} & \text{if } t >= 0 \\ 0 & \text{otherwise} \end{cases} \\ H_{a_i}(t, T_{1_i}, T_{2_i}) &= G_{a_i}(t - T_{1_i}) - G_{a_i}(t - T_{2_i}) \\ G_{a_i}(t) &= \begin{cases} Min(1 - (1 + \beta_i t)e^{-\beta_i t}, \theta) & \text{if } t >= 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

The solutions of this problem are the zeros of the partial derivatives of E and are thus solutions of the following set of equations. The time and angular parameters are the solutions of non-linear equations whereas the amplitudes and $\hat{F}_{0_{min}}$ may be computed from the following linear set of equations :

Partial Derivative according to $F_{0_{min}}$:

$$\sum_{n=1}^{N} ln(F_0(n)) =$$

$$N ln(F_{0_{min}}) + \sum_{i=1}^{P} A_{p_i} \sum_{n=1}^{N} G_{p_i}(n - T_{o_i}) +$$

$$\sum_{i=1}^{A} A_{a_i} \sum_{n=1}^{N} H_{a_i}(t, T_{1_i}, T_{2_i})$$

Partial Derivative according to A_{p_i} :

$$\sum_{n=1}^{N} ln(F_0(n)) G_{p_k}(n - T_{o_k}) =$$

$$ln(F_{0_{min}}) \sum_{n=1}^{N} G_{p_k}(n - T_{o_k}) +$$

$$\sum_{i=1}^{P} A_{p_i} \sum_{n=1}^{N} G_{p_i}(n - T_{o_i}) G_{p_k}(n - T_{o_k}) +$$

$$\sum_{i=1}^{A} A_{a_i} \sum_{n=1}^{N} H_{a_i}(t, T_{1_i}, T_{2_i}) G_{p_k}(n - T_{o_k}) k = 1 ... P$$

Partial Derivative according to A_{a_i} :

$$\sum_{n=1}^{N} ln(F_0(n)) H_{a_k}(n, T_{1_k}, T_{2_k}) =$$

$$ln(F_{0_{min}}) \sum_{n=1}^{N} H_{a_k}(n, T_{1_k}, T_{2_k}) +$$

$$\sum_{i=1}^{P} A_{p_i} \sum_{n=1}^{N} G_{p_i}(n - T_{o_i}) H_{a_k}(n, T_{1_k}, T_{2_k})) +$$

$$\sum_{i=1}^{A} A_{a_i} \sum_{n=1}^{N} H_{a_i}(t, T_{1_i}, T_{2_i}) H_{a_k}(n, T_{1_k}, T_{2_k})) k = 1 ... P$$

which can be described by a matrix equation :

$$A\ \mathrm{X} = B$$

where X and B are two vectors of dimension (A+P+1) (number of accents + number of phrases + 1) and A is a (A+P+1) square matrix.

$$X = (ln(F_{0_{min}}), A_{p_1}, A_{p_2}, ..., A_{p_P}, A_{a_1}, A_{a_2}, ..., A_{a_A})^{-T}$$

$$B = \begin{pmatrix} \sum_{n=1}^{N} ln(F_0(n)) \\ \sum_{n=1}^{N} ln(F_0(n)) G_{p_1}(n - T_{o_1}) \\ \cdot \\ \cdot \\ \sum_{n=1}^{N} ln(F_0(n)) G_{p_P}(n - T_{o_P}) \\ \sum_{n=1}^{N} ln(F_0(n)) H_{a_1}(n, T_{1_1}, T_{2_1}) \\ \cdot \\ \cdot \\ \sum_{n=1}^{N} ln(F_0(n)) H_{a_A}(n, T_{1_A}, T_{2_A}) \end{pmatrix}$$

A is defined by the set of equations :
$A = a_{i,j}$
where

$$
\begin{array}{lcll}
a_{1,1} & = & N & \\
a_{1,j+1} & = & \sum_{n=1}^{N} G_{p_j}(n - T_{o_j}) & \text{j=1...P} \\
a_{1,j+P+2} & = & \sum_{n=1}^{N} H_{a_j}(n, T_{1_j}, T_{2_j}) & \text{j=1...A} \\
a_{i+1,1} & = & \sum_{n=1}^{N} G_{p_i}(n - T_{o_i}) & \text{i=1...P} \\
a_{i+1,j+1} & = & \sum_{n=1}^{N} G_{p_i}(n - T_{o_i}) G_{p_j}(n - T_{o_j}) & \text{i=1...P,j=1...P} \\
a_{i+1,j+P+2} & = & \sum_{n=1}^{N} G_{p_i}(n - T_{o_i}) H_{a_j}(n, T_{1_j}, T_{2_j}) & \text{i=1...P,j=1...A} \\
a_{i+P+2,1} & = & \sum_{n=1}^{N} H_{a_i}(n, T_{1_i}, T_{2_i}) & \text{i=1...A} \\
a_{i+P+2,j+1} & = & \sum_{n=1}^{N} H_{a_i}(n, T_{1_i}, T_{2_i}) G_{p_j}(n - T_{o_j}) & \text{i=1...A,j=1...P} \\
a_{i+P+2,j+P+2} & = & \sum_{n=1}^{N} H_{a_i}(n, T_{1_i}, T_{2_i}) H_{a_j}(n, T_{1_j}, T_{2_j}) & \text{i=1...A,j=1...A}
\end{array}
$$

Such an equation is easily solved by means of the traditionnal Singular Value Decomposition which provides us with the pseudo-inverse of A.

Chapter 4

Prosodic Transformations

This chapter describes the prosodic transformations for synthesis purpose. Our module is based on TD-PSOLA, developped in CNET by Charpentier and Moulines [Moulines & Charpentier 90], which we will describe here in the first section. In the following sections, we will present how we can exactly match a target prosody. This technique allows high quality synthesis : it aims at modifying the prosodic information without changing the voice quality (keeping the spectra unchanged which may not always be a good point (e.g. female voice has a first formant higher)). The tool we developped may be very useful for synthesis system based on unit concatenation (see CHATR) and also for perceptual evaluation. In this last direction, we have been involved in a study carried out in ATR by Ofuka ([Ofuka 93]) on politeness.

4.1 Time Domain Pitch Synchronous OverLap and Add (TD-PSOLA)

Here we will describe a technique which has been described more fully in [Moulines & Charpentier 9(The technique is based on the Short-Time Fourier Analysis and Synthesis approach ([Allen & Rabiner 77]) which we won't detail either. The TD-PSOLA procedure involves three steps.

- The original speech signal is decomposed into a sequence of pitch synchronous short-term signals. These short-term signals are computed by multiplying the input signal by a bench of windows (usually Hamming windows) centered around the pitch marks. The figure 4.1 illustrates this first step.
- The second step consists in creating a sequence of synthesized short-term signals from the analysed ones. This stage can be decomposed into two steps :
  - a series of synthesis pitch-marks are generated and an analysis-synthesis mapping is determined : each synthesis pitch-mark is linked to two analysis pitch-marks. The next section is devoted to this procedure.

- the synthesis short-term signals are obtained by linear combination of the mapped analysis short-term signal.

This second stage is illustrated in figure 4.2.

- The last step of the procedure corresponds to the OverLap-Add synthesis which is described by the folowing formula :

$$\hat{s}(n) = \frac{\sum_m \hat{s}_m(n)}{\sum_m w(n-\hat{t}_m)}$$

where w denotes the synthesis window, $\hat{t}_m$ the synthesis pitch-marks and $\hat{s}_m$ the synthesis short-term signals computed previously.

4.2 Source/Target Pitch Mapping

4.2.1 Target Pitch Marking

As mentioned above, we aim at modifying a source pitch contour into a target pitch contour. We may imagine two different possibilities.

- Copy of the prosody of a natural utterance : In that case, the pitch-marking is used on the target utterance providing us with a set of pitch-marks. Only the mapping will have to be performed.

- Copy of a synthesized pitch contour : The procedure determines a set of pitch marks given a sampled pitch contour (which can be synthesized via any pitch model). Pitch-Marks are obtained recursively. Let us suppose, the current pitch-mark has been determined, the following one will be set such that the two marks are distant from the current pitch period.

An example is drawn on figure 4.3.

4.2.2 Pitch Mapping

In order to perform the TD-PSOLA synthesis, we need to map the source to the target pitch-marks. We will suppose here that both the source and the target phonetic labellings are known (the phonetic contents have to be similar). We then derive a piece-wise mapping from the labelling : for each label, the mapping function is assumed linear and depend on the respective source and target phoneme durations. This is illustrated on figure 4.4.

Once the pitch-mapping is obtained, the TD-PSOLA approach is performed and the signal is synthesized.

4.3 Manual

We developed two similar programs to perform these transformations. The first one, called "prosody", is interactive and allows the changing of an arbitrary pitch-contour and/or durations by using a graphical user interface. Synthetic or natural pitch-contours can also be input. The second one, called transform, copies a target sampled pitch-contour onto the original signal. Besides, this last algorithm has been integrated into CHATR, a generic synthesizer developed in ATR.

- prosody : It can be called without any argument. But, a constant file called "proso.cst" may be provided. This constant file allows sample-rate modification, pitch-marks and pitch file-format modifications (only ascii and binary are supported : these formats correspond to the pitch-tracking output format) and last the pitch sample-rate. An example follows :

  ```
  ; Waveform sampling rate in Hz
  SAMPLE_RATE      12000

  ; File types
  F0_FILE_TYPE     binary
  MK_FILE_TYPE     binary

  ; F0 tracking options
  FRAME_INTERVAL  0.005
  ```

  The following array appears on the screen. Here follow the corresponding explanations.

- transform :

  Different options have to be specified :

  - -i : input file name with no extension. 4 different files are needed :
    1. the original waveform (binary short).
    2. the phonetic description : list of phonemes preceded by its beginning and end in samples.
    3. the pitch-mark file : obtained trough our pitch-tracking algorithm.
    4. the pitch contour file : obtained trough our pitch-tracking algorithm.

    The extensions of those files can be chosen into the constant file. If not, the defaults are respectively "wav", "ctl", "pmk" and "pit".
  - -o : output file name with no extension. 2 different files are needed :
    1. the phonetic description
    2. the pitch contour file

    The program checks that the phonetic description of both source and target files is similar.

- -f : sampling frequency of original waveform (optional as it is mentioned in the pitch-mark file).
- -c : constant file which contains the following field :

  1. files extension
     WAVE_EXTENSION
     PIT_EXTENSION
     MKA_EXTENSION
     CTL_EXTENSION
  2. files format (binary or ascii)
     F0_FILE_TYPE
     MK_FILE_TYPE
  3. F0 sample rate : frame interval between 2 successive pitch values (ms). This value is also contained in the pitch-contour files.
     FRAME_INTERVAL
  4. Marking period on non-voiced portion (ms).
     NONSYNCH_WIN
  5. Dimension of the analysis/synthesis window as a factor of local pitch-period (typically 2 or 3).
     WINDOW_SIZE
  6. Maximum Size of windows (ans short-term signals).
     FFT_SIZE
  7. Use of swapping on non-voiced portion (0 = no / 1 = yes). Swapping avoid periodicity (noise) on non-voiced portions.
     SWAPPING
     Here follows an example.

```
WAVE_EXTENSION .wav
PIT_EXTENSION .pit
MKA_EXTENSION .mka
CTL_EXTENSION .ctl

F0_FILE_TYPE binary
MK_FILE_TYPE binary

FRAME_INTERVAL 5
FFT_SIZE 1024
WINDOW_SIZE 2
SWAPPING 1
NONSYNCH_WIN 10
```

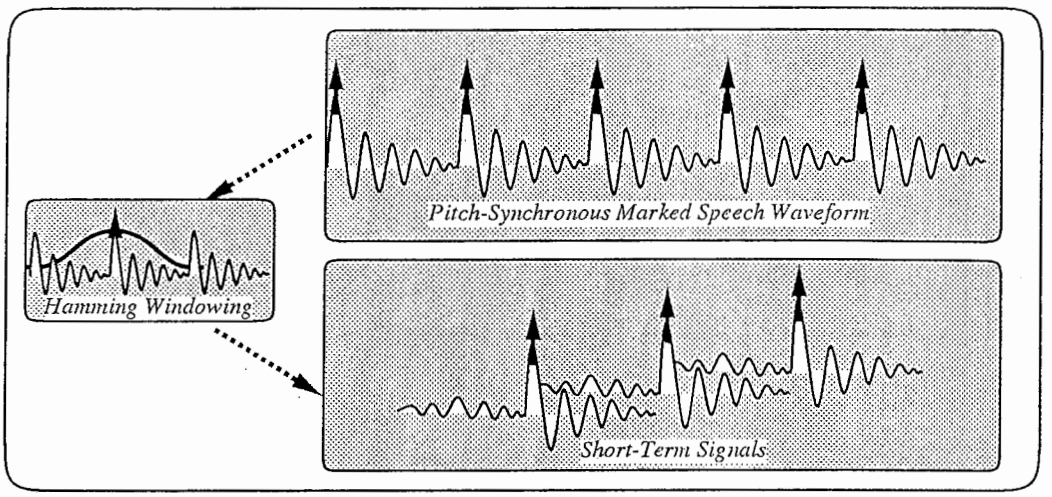

Figure 4.1: Decomposition into Analysis-Short-Term Signals

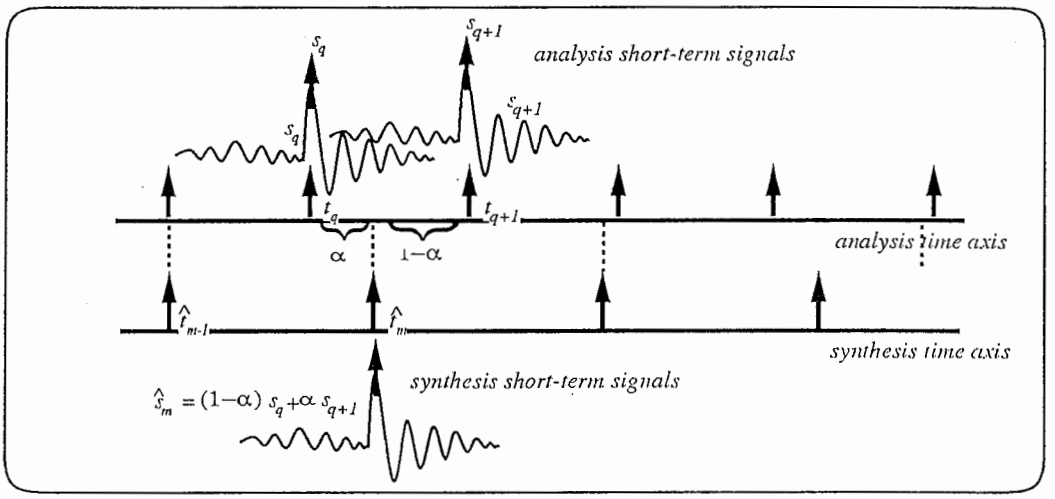

Figure 4.2: Synthesis Short-Term Signals
The synthesis short-term signal $\hat{s_m}$ corresponding to the synthesis pitch mark $\hat{t_m}$ is the linear combination of the 2 analysis short-term signals s_q and s_{q+1}. Note that in this case illustrated above, no time dilatation has been performed.

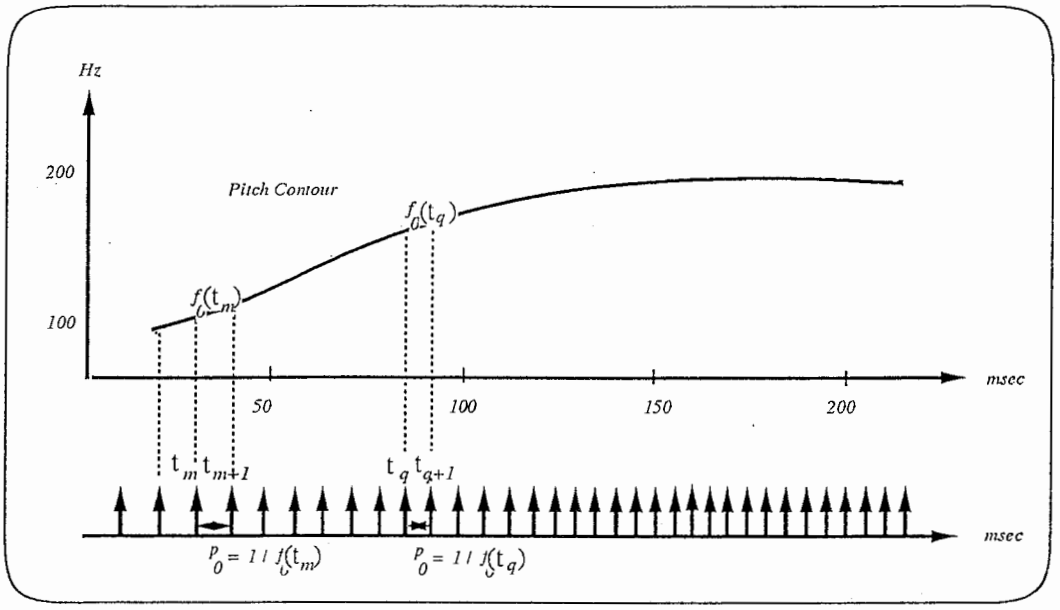

Figure 4.3: From Pitch Contour to Pitch Marking
Pitch-marks are determined recursively. On non-voiced portion, marks are spaced regularly ; On voiced portion, the interval between two successive marks t_m and t_{m+1} is equal to the local pitch period given by the pitch-contour at time t_m

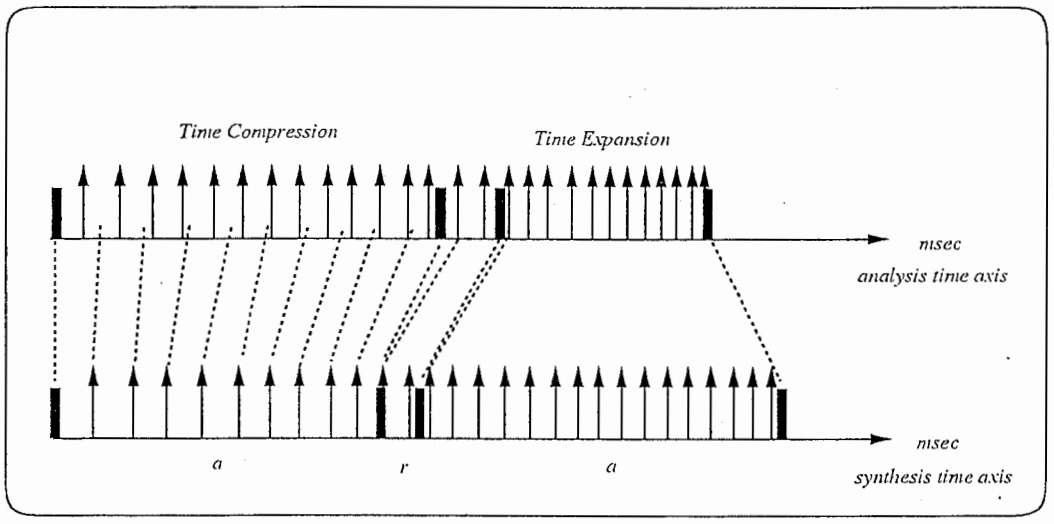

Figure 4.4: Analysis-Synthesis Mapping
The first mark of each phoneme are time-aligned.

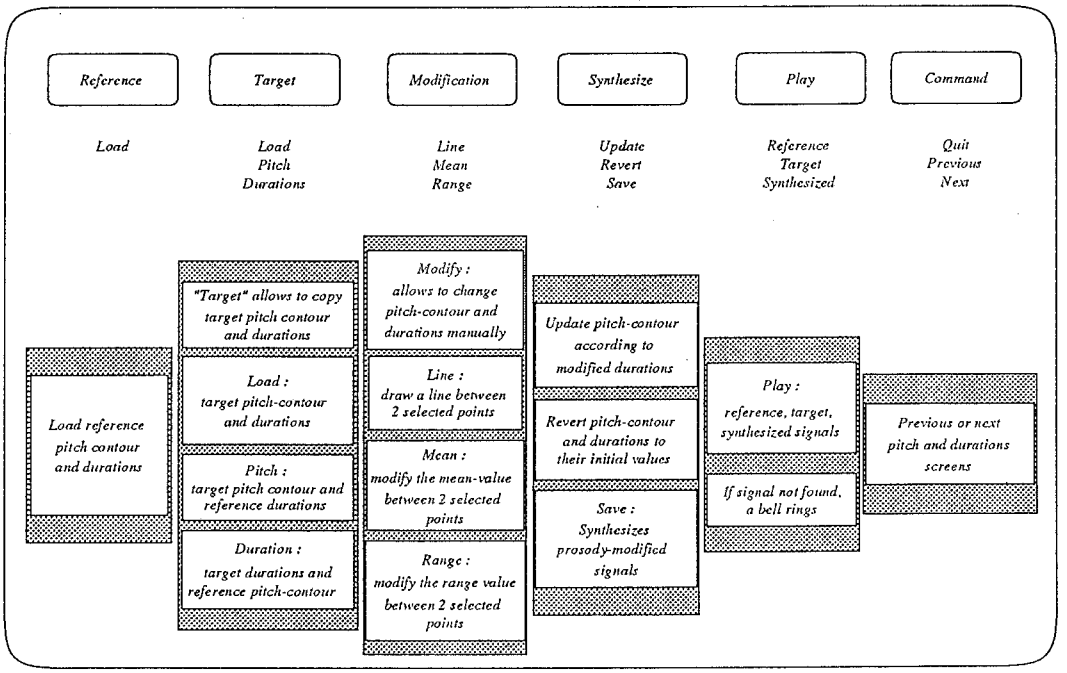

Figure 4.5: Interactive Prosodic Transformation : the graphical user interface

Chapter 5

Results

We will present here some experiments carried out to analyze a speaker's prosodic characteristics. These experiments have not been so far very conclusive and we will try to show the weaknesses and of our approach. The first part of this chapter will evaluate the accuracy of the Fujisaki and the RFC model. The second part will present our experiments carried on rough pitch-contour which led us to use some very simple mapping to perform the pitch modification from one speaker to another.

5.1 Database

The database [Abe *et al* 90] is composed of 500 japanese sentences uttered by 3 native male speakers MHT, MYI and MSH. Only one utterance has been recorded except for speaker MHT. Speakers are professional and the corpus has been read. All speakers are Tokyo dialect native speakers.

The database is labelled in terms of phonetics and phonology : accents, syntax, type of words, assimilation and devoicing ... are defined. The second corpus uttered by MHT is unfortunately not labelled. In order to estimate prosodic intravariability, we used a semi-automatic procedure ([Takami & Sagayama 92]) to segment the datas. The segmentation has then been manually checked. This labeling stage has only be applied on 20 sentences.

5.2 Analysis

5.2.1 Use of a Model

The first set of experiments consisted of analysing the database using one of our models. We then check at the accuracy of the model an analysis-synthesis technique ; for each sentence we measured the mean-squared error between the original pitch-contour and the synthesized one.

We here cite the results obtained in [Hirai *et al* 93]. Four hundred sentences uttered by Speaker MYI has been analysed thanks to Hirai's technique using sta-

tistical analysis. The mean-squared error is computed ; pitch values are measured in a logarithmic scale. The error criterium is then defined as :

$$Error^2 = \frac{1}{N_s}\sum_i^{N_s}\frac{1}{N_{in}}\sum_n^{N_{in}}[ln(\hat{F}_{0_i}(n)) - ln(F_{0_i}(n))]^2$$

where N_s is the number of analysed sentences (400 here), N_{in} is the number of pitch samples obtained for the ith sentence, $\hat{F}_{0_i}(n)$ and $F_{0_i}(n)$ are the n_{th} estimated and observed pitch samples (i_{th} sentence). Standard deviation is also computed.

Speaker	Mean-Squared-Error	Standard Deviation
MYI	0.224	0.0248

which means that for an observed $F_0 = 150Hz$, the error is around 30 Hz.

We analyze the same database with the RFC model and results are summed up by the following array (same error measure) :

Speaker	Mean-Squared-Error	Standard Deviation
MYI	0.061	0.0046

which means that for an observed $F_0 = 150Hz$, the error is around 10 Hz (however, we didn't check the "meaning" of the rise and fall elements on the whole database). The appendix presents the results obtained on MYI and MHT with another error criterium :

$$Error = \frac{1}{N_s}\sum_i^{N_s}\frac{1}{N_{in}}\sum_n^{N_{in}}|\hat{F}_{0_i}(n) - F_{0_i}(n)|.$$

As this second model was more accurate we try to analyse the rise and fall tilts. According to Vaissiere, each speaker tends to use a small amount of different rise and fall gradients. To check this theory, we thus built the histograms of such patterns. The following figures present the results obtained for the two speakers MHT and MYI.

According to these histograms, the rise and fall patterns follow some wide gaussian distributions. Two different interpretations can be done :

- the RFC model is not accurate enough so that rise and fall gradients may vary greatly ; moreover, some rise and fall may still correspond to some microprosody despite a careful analysis.
- Vaissiere's theory is not verified : speakers use as large a range as possible of rise and fall patterns.

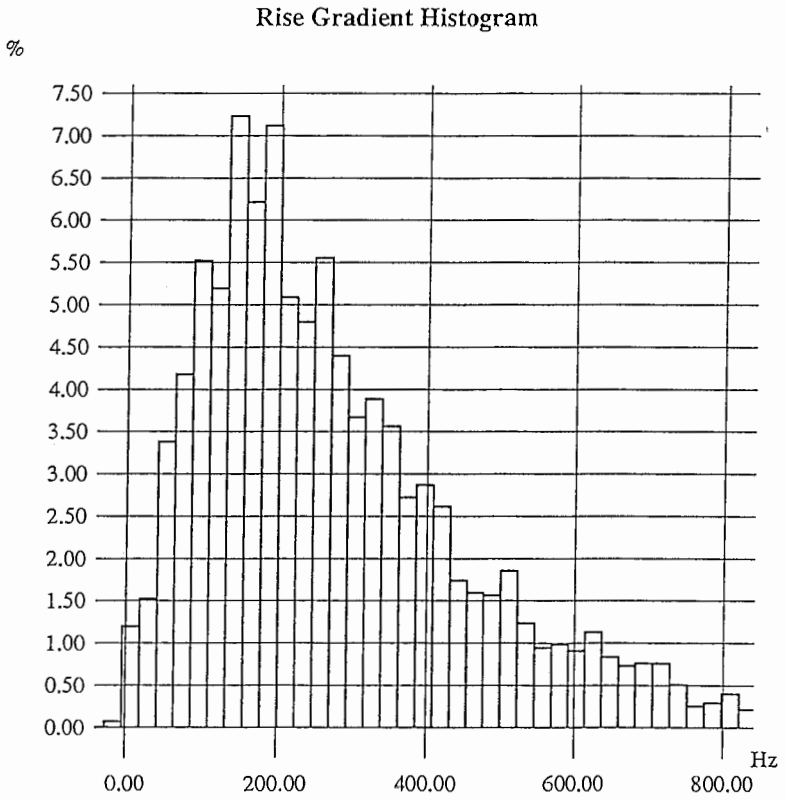

Figure 5.1: Rise Gradient Histogram

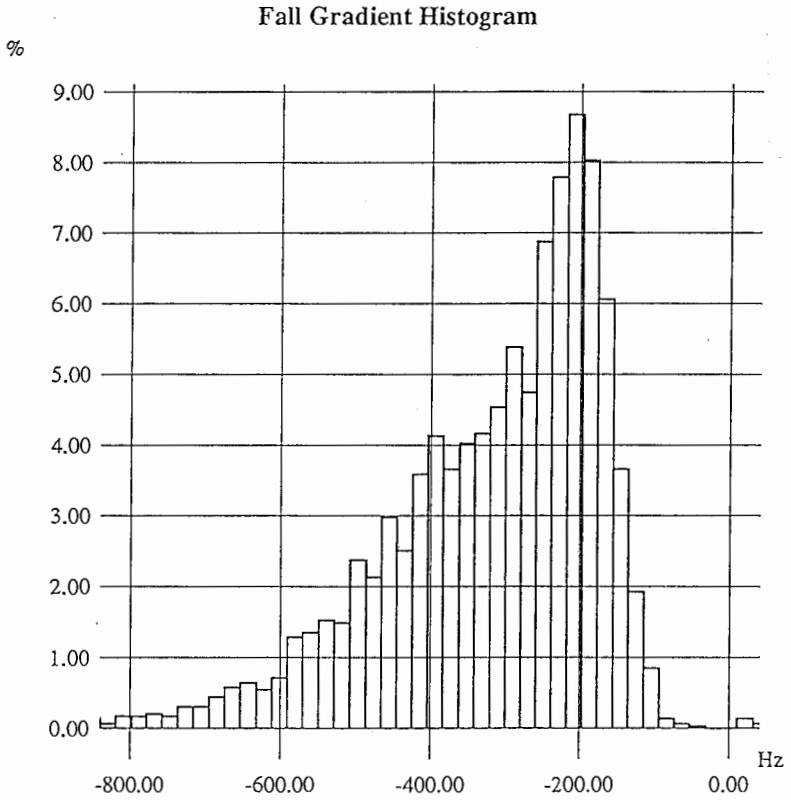

Figure 5.2: Fall Gradient Histogram

5.2.2 Raw Pitch Contour

We also carried out a second set of experiments on the raw pitch contour. We are aware that such an approach may be affected by micro-perturbations. However, we compared raw pitch contour of the same utterance according to the procedure described in the preceding chapter : the pitch-contour is expanded or compressed according to the phoneme durations. The two contours are then compared point by point. We obtained the following measures, each speaker being "normalized" to MHT :

Speaker	Mean-Squared-Error	Standard Deviation
MHT-MHT	8.16	2.37
MYI-MHT	32.64	6.25
MSH-MHT	33.18	4.72

The first one gives an idea of the intra-speaker variability whereas the two others show the inter-speaker differences. The figure 5.3 illustrate this procedure. As one may notice, there is nearly no difference in accent position (start and end mainly happened at the same time). Further more an analysis of the first 20 sentences only shows up a very small number of differences in word grouping (phrase groups).

Next, a graph of the fundamental frequency of one speaker against the fundamental frequency of the other speaker was drawn. MYI's pitch contours have been time-normalized to the corresponding MHT's pitch contours so that each pitch value of MHT is mapped to one value of MYI. For each MHT's pitch value, we then computed the average and standard deviation of the mapped MYI's pitch values. We thus obtained the figures 5.4, 5.5.

It appears that the mapping between the two speakers pitch contour is highly linear except for high frequencies where it seems to be quite random. We may see different reasons which could explain that : first, there is only a few high frequency samples in comparison to the other ranges ; second these high frequencies may be erroneous corresponding to pitch doubling ; third this could correspond to different choices in phrasing : pitch would be reset in one case whereas it won't be in the other case.

We thus decided to compute for each sentence uttered by two speakers a linear function which would map each value of the reference speaker's pitch contour to the corresponding value of the target speaker's normalized pitch contour. The function is simply defined by the following equations :

$$\hat{f}_{targ}(t) = \alpha f_{ref}(\gamma(t)) + \beta$$

where $\gamma(t)$ is the time dilatation/compression function.

The mean pitch value, the range as well as the tilts are thus modified. Figure 5.6 shows some examples obtained that way.

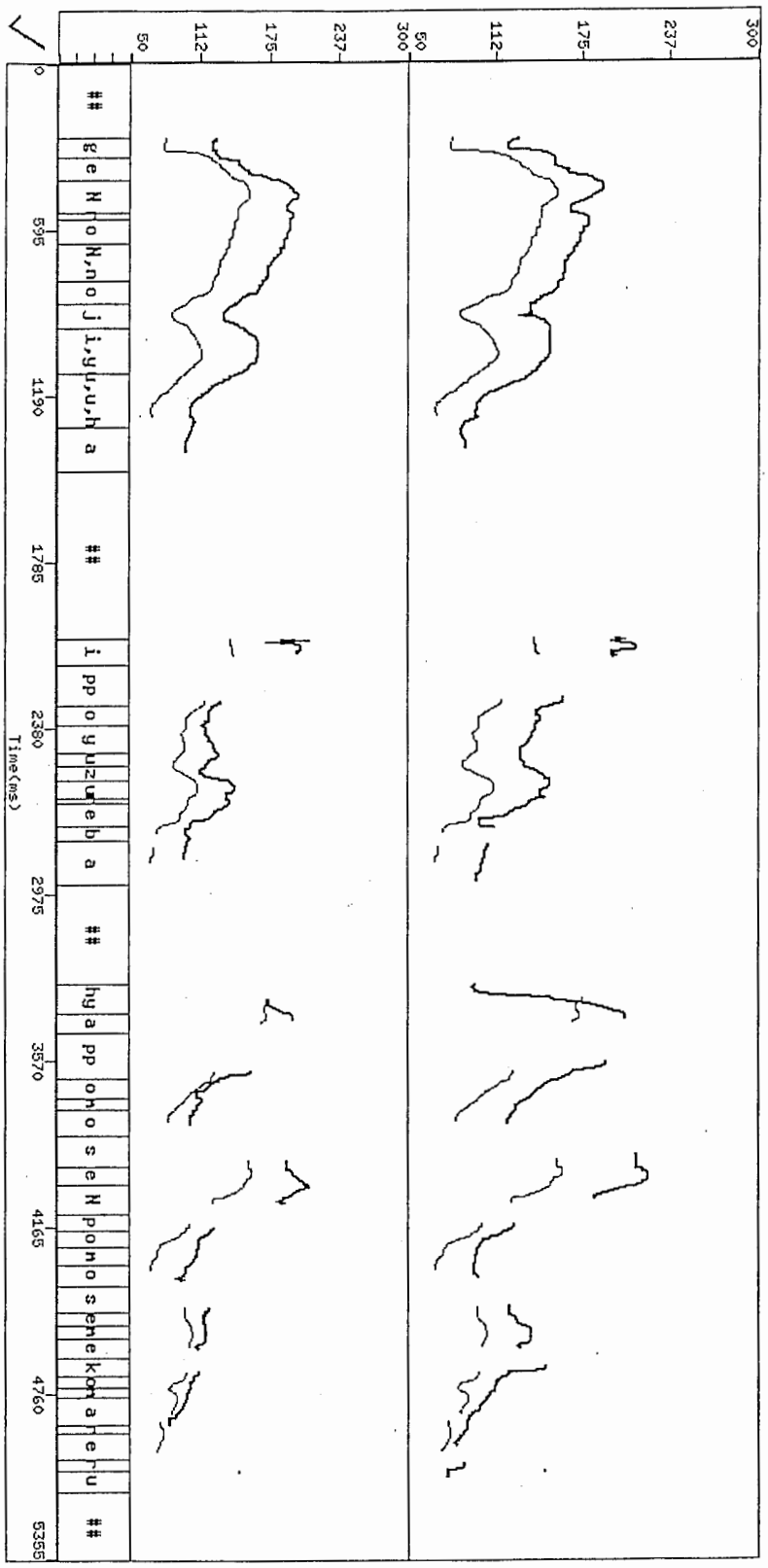

Figure 5.3: Two pairs of pitch contours (2 speakers uttering the same sentences) after time alignment to MHT's phoneme durations
The first pair corresponds to the couple MHT-MYI whereas the second corresponds to the couple MHT-MSH

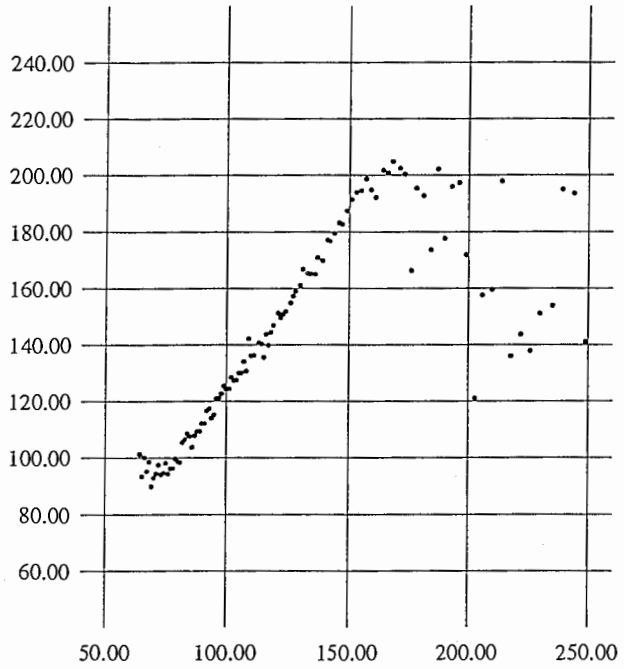

Figure 5.4: MYI's pitch values as a functions of MHT's pitch values : "average curve"

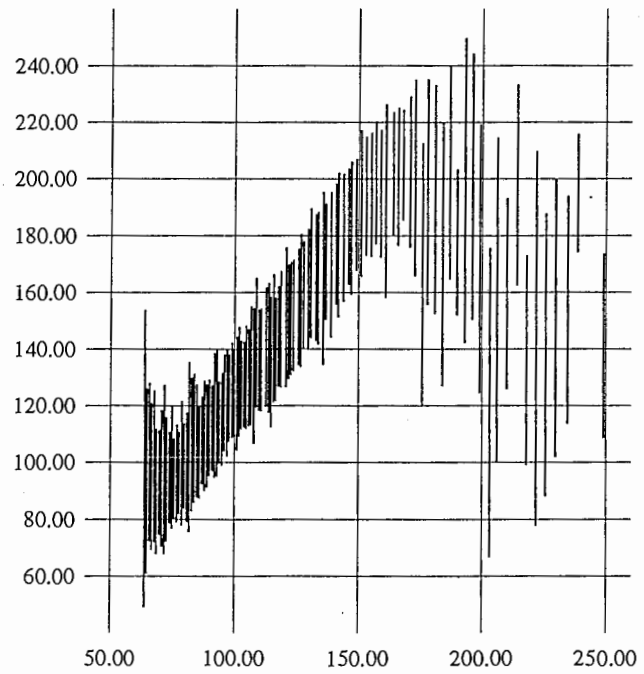

Figure 5.5: MYI's pitch values as a functions of MHT's pitch values : Standard Deviation

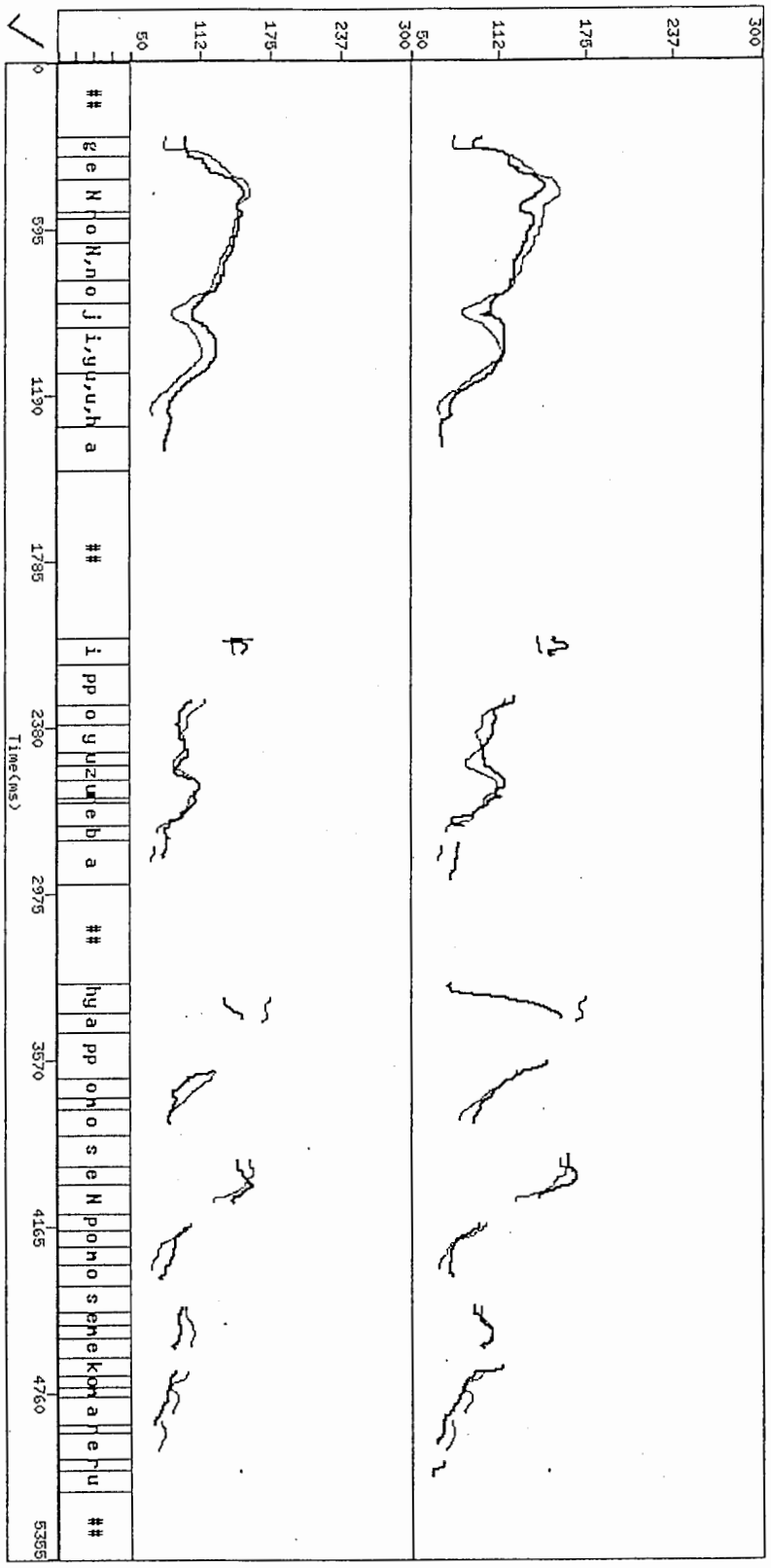

Figure 5.6: Two pairs of pitch contours (2 speakers uttering the same sentences) after time alignment to MHT's phoneme durations and frequency normalisation . The same examples as above are presented here.

We then computed the mean-squared error for each sentence as well as the standard deviation. The following array shows the average value of such error on the whole database.

Speaker	Mean-Squared-Error	Standard Deviation
MHT-MHT	7.22	2.27
MYI-MHT	11.10	2.91
MSH-MHT	12.58	3.38

After such a normalisation the inter-speaker variability is drastically reduced ; Inter-speaker difference and intra-speaker difference become similar.

These results are obtained for our japanese database which has been uttered by some trained speakers : they learn to pronounce it in a similar way : phrasing strategies are usually identical. However, Japanese language seems to be constraint in such a way that in case of the same dialect, word accent positions are stable and very well defined ; the amplitudes only appeared to be variable.

However, we try to compare the coefficients α and β obtained for each sentence and they appear to vary quite a lot.

5.3 Appendix

The RFC model is used to analyse 500 sentences uttered by the 2 speakers MHT and MYI. The error-criterium we choose on this database is the following :

$$Error_1 = \frac{1}{N_s N_{in}} \sum_i^{N_s} \sum_n^{N_{in}} |\hat{F}_{0_i}(n) - F_{0_i}(n)|$$

and

$$Error_2 = \sqrt{\frac{1}{N_s N_{in}} \sum_i^{N_s} \sum_n^{N_{in}} |\hat{F}_{0_i}(n) - F_{0_i}(n)|^2}$$

In this error definition, the pitch value are considered independantly of the sentence (here $N_s N_{in} \approx 328000$). We don't use the logarithmic scale : the Bark scale is linear for frequency values below 500 Hz which is the case here.

Speaker	$Error_1$	$Error_2$
MHT	4.906652	8.916479
MYI	4.605651	8.570259

Here follow the error histograms of each "speaker".
Here follow the error histograms of each "speaker".

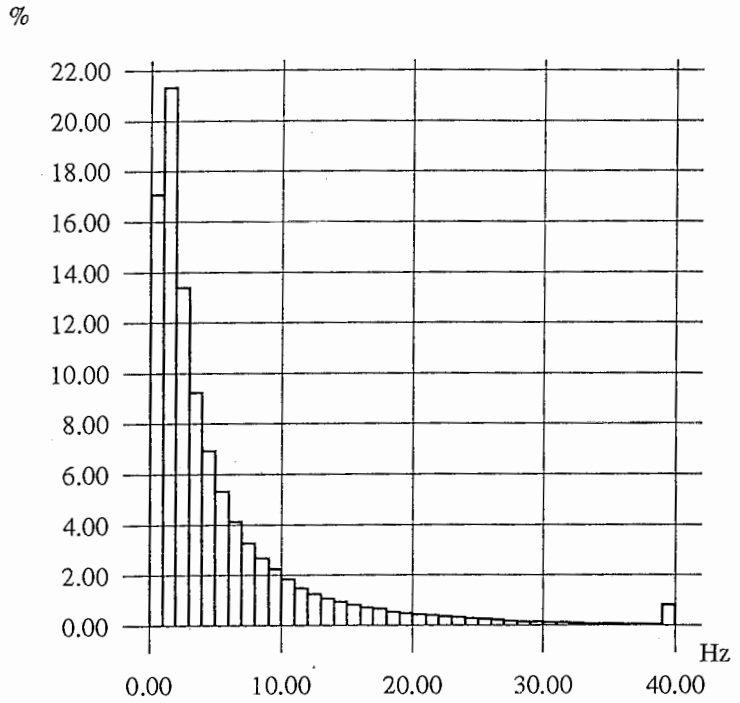

Figure 5.7: Error histogram : speaker MHT

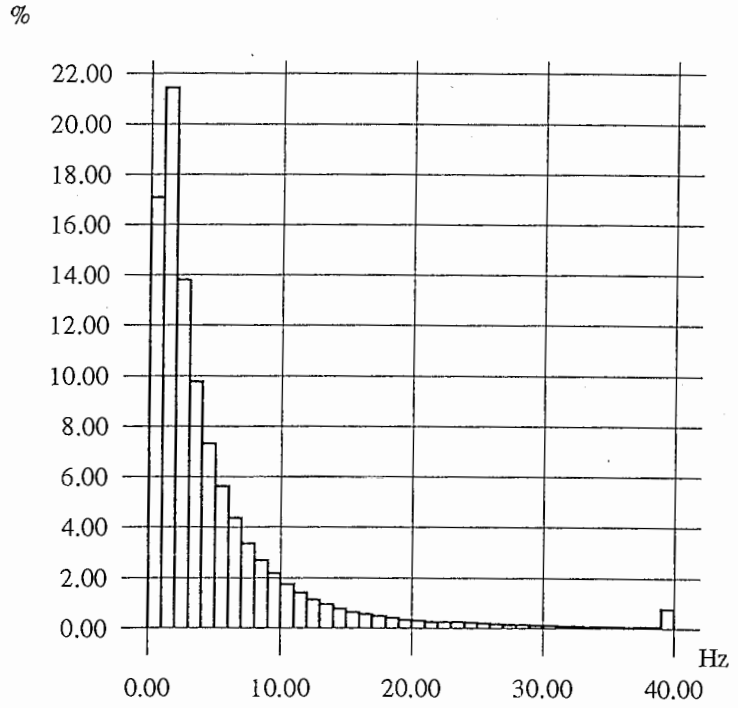

Figure 5.8: Error histogram : speaker MYI

Chapter 6

Conclusions

We described in this report two different methodologies to analyse speaker's prosodic characteristics. The evaluation has been carried out on a Japanese database composed of 3 male speakers uttering 500 sentences.

- The first one is based on pitch modelisation. We compare two different models : the Fujisaki Model (which has been implemented) and the RFC model. The second model has two advantages : it is fully automatic and seems to be much more accurate than the first one. However, the model has to carefully tuned such that rise and fall components are phonologically meaningfull. This training is not yet automatic and quite difficult to do. The use of such models didn't really allow us to conclude on speaker's characteristics as they neither were accurate enough to compare the accents characteristics from one speaker to another (gradient and timing, accents shape is also kept invariant in these models).

- The second method consists in comparing point to point each pair of pitch contours corresponding to the same utterance uttered by two different speakers. It then appears that one pitch contour can be derived from the other by a simple linear transformation. The difference between the frequency normalized pitch contour is nearly reduced to the intra-speaker variability which can be observed when one speaker utters the same sentence twice. However such a technique assumes that there is no timing difference in accent position which seems to be true in our Japanese database but which was not observed in another English database. This accent characteristic can be explained by two reasons : the Japanese database is read by professional speakers who have been trained to speak in a received style ; Japanese language is very constrained by phonological rules and accents location is very restricted. In this case, the use of a linear transformation seems to be sufficient.

Chapter 7

Other task in ATR : spectral analysis

This section describes very briefly the program I developed to analyse speech data. This program performs different traditional analysis : LPC ([Markel & Gray 76], LPC-Cepstrum, Iterative Cepstrum (true envelope) ([Imai & Abe 79] and Delta-Cepstrum ([Sagayama & Itakura 79], [Furui 86]). We will assume that the reader is familiar with these analysis and won't detail them. The interested reader can find some descriptions of those techniques in [Rabiner & Juang 93]. The results of the analysis will be used to select units for speech synthesis. The spectral continuity is necessary to allow high-quality synthesis.

The routine is called "analysis" and needs the following inputs which follows. At least two inputs are required : input signal file and output coefficients file.

- -i <input signal file> : nist and raw formats are supported. In the case of a raw format, the sample-rate is needed.
- The output may be a cepstrum, an lpc or a delta-cepstrum file. In case of lpc-cepstrum file, lpc and cepstrum may be both saved (as well as for delta-cepstrum analysis and cepstrum).
  - -oc <cepstrum file>
  - -ol <lpc file>
  - -od <delta-cepstrum file>

  The output format is either simple (ascii) or headered binary (nist). A header details the analysis. Here follows an example (in case of an ascii file) for a pitch-synchronous cepstrum analysis :

```
utterance_id w0001.wav
sample_rate -12000
window_type hamming
window_size 20  (in ms)
pre_emphasis 0.950000
```

```
data_spacing Pitch_Synchronous
data_type iterative_cepstrum
iteration_order 1
channel_count 16
sample_count 47
```

- 2 other input can be added :
  - -m <mark file> : If the pitch-synchronous-marks file (cf above) can be found, then the analysis is pitch-synchronous else it will be regular.

    In case of pitch-synchronous analysis, analysis windows are centered on the pitch-marks computed previously. The first coefficient of the analysis will then be the position of the pitch-mark in ms.

    In case of regular analysis, the windows are shifted by a fixed period. The first window is centered at instant 0 (half of it is non-zero). Analysis is carried out until the window center cannot be shifted (out of the signal file). We thus obtain (signal_size / window_shift + 1) analysis vectors.
  - -c <constant file> : This constant file allows the use to define specific information concerning the file formats and the analysis details. If such a constant file is not defined, then default values are used. Here is the list of the possible choices.
    * File Formats :
      WAVE_TYPE raw or nist (default is raw)
      MARK_TYPE ascii or xmg (default is ascii)
      CEPSTRUM_TYPE : ascii or nist (default is nist)
      LPC_TYPE : ascii or nist (default is nist)
      DELTA_TYPE : ascii or nist (default is nist)
    * Analysis information : ANALYSIS : cepstrum, lpc, lpc-cepstrum
      CEPS_ORDER : cepstrum analysis order (default is 12)
      LPC_ORDER : lpc analysis order (default is 16)
      FFT_POWER : used in case of "FFT-Cepstrum" analysis (default is 10)
      IMP_CEPS_NB_ITER : number of iterations in case of True-Enveloppe Analysis (improved iterative cepstrum) (default is 20)
      DELTA_CEPSTRUM : 0 in case of no delta-cepstrum analysis or 1 in the contrary (default = no analysis)
      DELTA_WIND : size of the window used for Delta-Analysis (default is 10)
      WINDOW_TYPE : hamming or rectangular (default is hamming)
      WINDOW_DURATION in ms (default is 20 ms)
      WINDOW_SHIFT in ms (used if no pitch-mark file found ; default is 5 ms)
      PRE_ACCENTUATION (default is 0.95)
    * Others :

WAVE_SAMPLE_RATE : in case the input signal file is raw, the sample-rate is need. Default value is fixed to 12000 Hz.
WAVE_ENCODING : format of a raw input signal file : default is "lin16MSB" format which is SUN format.

Bibliography

[Abe *et al* 90] M. Abe, Y. Sagisaka, T. Umeda, and H. Kuwabara. Speech database user's manual. Technical Report TR-I-0166, Advanced Telecommunications Research Institute International, 1990.

[Allen & Rabiner 77] J.B. Allen and L.R. Rabiner. A unified approach to short-time fourier analysis and synthesis. *Proceeding of the IEEE*, 65(11):1558–1564, 1977.

[Bagshaw *et al* 93] P.C. Bagshaw, S.M. Hiller, and M.A. Jack. Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching. In *Eurospeech'93*, pages 1003–1006, 1993.

[Black & Taylor 94] A. W. Black and P. Taylor. CHATR: a generic speech synthesis system. submitted to COLING-94, 1994.

[Fant 60] G. Fant. *Acoustic Theory of Speech Production.* Mouton and Co., s'Gravenhage, 1960.

[Fant 66] G. Fant. A note on vocal tract size factors and non-uniform f-pattern scalings. Technical Report 4/66 (pp.22-30), Royal Institute of technology, Stockholm, Sweden, 1966.

[Fujisaki & Hirose 84] H. Fujisaki and K. Hirose. Analysis of voice fundamental frequency contours for declarative sentences of japanese. *Journal of Acoustical Society of Japan*, 5(4):233–242, 1984.

[Fujisaki 91] H. Fujisaki. Modeling the generation process of f0 contours as manifestation of linguistic and paralinguistic information. In *Proceedings of the 12th Internation Congress of Phonetic Sciences*, pages 1–9, 1991.

[Furui 86] S. Furui. Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Trans. on Signal Processing*, 34(1):52–59, 1986.

[Geoffrois 93] E. Geoffrois. A pitch contour analysis guided by prosodic event detection. In *Eurospeech'93*, pages 793–796, 1993.

[Hess 83] W. Hess. *Pitch Determination of Speech Signals.* Springer-Verlag, Berlin Heidelberg NewYork Tokyo, Berlin Heidelberg, 1983.

[Hirai *et al* 93] T. Hirai, N. Iwahashi, H. Valbret, N. Higuchi, and Y. Sagisaka. Fundamental frequency contour modeling using statistical analysis. In *Fall Meeting of the Acoustical Society of Japan*, pages 225–226, 1993.

[Hirst *et al* 91] D. Hirst, P. Nicolas, and R. Espesser. Coding the f0 of a continuous text in french : an experimental approach. In *Actes du XIIeme congres international des sciences phonetiques*, volume 5, pages 234–237, Aix en Provence, 1991.

[Imai & Abe 79] S. Imai and Y. Abe. Spectral envelope extraction by improved cepstral method. *Transactions on IECE*, J62-A(4):217–223, 1979.

[Karlsson 90] I. Karlsson. *Analysis and Synthesis of Different Voices with Emphasis on Female Speech.* Unpublished PhD thesis, KTH, Department of Speeach Communication and Music Acoustics, Royal Institut of Technology, Stockholm, 1990.

[Kubozono 93] H. Kubozono. *The Organization of Japanese Prosody*, volume 2 of *Studies in Japanese Linguistics.* Kurosio, Tokyo, 1993.

[Laver 80] J. Laver. *The Phonetic Description of Voice Quality.* Cambridge University Press, Cambridge, UK., 1980.

[Markel & Gray 76] J.D. Markel and A.H. Gray. *Linear Prediction of Speech.* Springer-Verlag, Berlin Heidelberg NewYork, 1976.

[Medan *et al* 91] Y. Medan, E. Yair, and D. Chazan. Super resolution pitch determination of speech signals. *IEEE Transactions on Signal Processing*, 39(1):40–48, January 1991.

[Moulines & Charpentier 90] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5/6):453–467, 1990.

[Ofuka 93] E. Ofuka. The role of speaking style, duration and f0 in signalling affect : anger, kindness, and politeness. Technical Report TR-034, Advanced Telecommunications Research Institute International, 1993.

[Pierrehumbert & Beckman 88] J. Pierrehumbert and M. Beckman. *Japanese Tone Structure*. The MIT Press, Cambridge, Massachussetts, London, England, 1988.

[Pierrehumbert 80] J. Pierrehumbert. *The Phonology and Phonetics of English Intonation*. Unpublished PhD thesis, Massachussets Institute of Technology, 1980.

[Rabiner & Juang 93] L. Rabiner and B. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, Englewood Cliffs,NJ, 1993.

[Sagayama & Itakura 79] S. Sagayama and F. Itakura. On individuality in a dynamic measure of speech. In *Spring Meeting of the Acoustical Society of Japan*, pages 589–590, 1979. (in Japanese).

[Sagisaka & Sato 86] Y. Sagisaka and H. Sato. Some accentual characteristics in japanese phrases and long compound. *Journal of Acoustical Society of Japan*, 7(1):65–74, 1986.

[Stevens 72] K.N. Stevens. Sources of inter- and intra-speaker variability in the acoustical properties of speech sounds. In *Proceedings of 7th International Congress of Phonetic Sciences*, pages 206–232, 1972.

[Takami & Sagayama 92] J. Takami and S. Sagayama. A successive state splitting algorithm for efficient allophone modeling. In *ICASSP*, volume 1, pages 573–576, San Francisco, 1992.

[Taylor 92] P. Taylor. *A Phonetic Model of English Intonation.* Unpublished PhD thesis, University of Edinburgh, 1992.

[Taylor 93] P. Taylor. Synthesizing intonation using the RFC model. In *Proc. ESCA Workshop on Prosody, Lund, Sweden*, 1993.