Article Submitted to Computer Speech and Language

Spectral Stability Based Event Localizing Temporal Decomposition

A.C.R. Nandasena, P.C. Nguyen and M. Akagi

Japan Advanced Institute of Science and Technology
1-1 Asahidai, Tatsunokuchi, Nomi, Ishikawa 923-1292, Japan

Abstract

In this paper a new approach to Temporal Decomposition (TD) of speech, called Spectral Stability Based Event Localizing Temporal Decomposition (S^2BEL-TD), is presented. The original method of TD proposed by Atal (1983) is known to have the drawbacks of high computational cost, and the high parameter sensitivity of the number and locations of events. In S^2BEL-TD, the event localization is performed based on a maximum spectral stability criterion. This overcomes the high parameter sensitivity of events of Atal's method. Also, S^2BEL-TD avoids the use of the computationally costly singular value decomposition routine used in the Atal's method, thus resulting in a computationally simpler algorithm for TD. Simulation results show that an average spectral distortion of about 1.5 dB can be achieved with line spectral frequencies as the spectral parameter. It is shown that the temporal pattern of the speech excitation parameters can also be well described using the S^2BEL-TD technique.

1. Introduction

In articulatory phonetics, speech production is considered as a sequence of overlapping articulatory gestures, each of which may be thought of as a movement towards and away from an ideal, but often not reached, articulatory target. The sound produced by such an articulatory movement corresponds to a phoneme or a sub-phoneme in speech. In other words, each gesture produces an acoustic event that should approximate a phonetic target. Adjacent gestures overlap one another resulting in the characteristic transitions between phonemes that can be observed in almost any parametric representation of the acoustic speech signal. Due to co-articulation and reduction in fluent speech, a target may not be reached before articulation towards the next phonetic target begins. It has long been a difficult task to determine such targets and their temporal evolutionary patterns from the acoustic signal alone.

The so-called *temporal decomposition* method for analyzing speech achieves the objective of decomposing speech into targets and their temporal evolutionary patterns, without any recourse to any explicit phonetic knowledge. This model of speech takes into account the above articulatory considerations and results in a description of speech in terms of event targets describing the ideal articulatory configurations of the successive acoustic events in speech, and event functions describing their temporal evolutionary patterns. Therefore, it tries to achieve an optimal transformation from the multidimensional spectral parameter space to the phonetic space which can be considered for many applications to be a powerful speech analysis technique.

Suppose that a given utterance has been produced by a sequence of K movements aimed at realizing K acoustic targets. Let us denote the speech parameters corresponding to the kth target by $\mathbf{a}(k)$, and the temporal evolution of this event by a function, $\phi_k(n)$. The frame number n varies between 1 and N. In temporal decomposition of speech, the observed speech parameters, $\mathbf{y}(n)$, are approximated by $\hat{\mathbf{y}}(n)$, a linear combination of event targets as follows.

$$\hat{\mathbf{y}}(n) = \sum_{k=1}^{K} \mathbf{a}_k \phi_k(n), \quad 1 \leq n \leq N \tag{1}$$

In matrix notation the Equation (1) can be written as;

$$\hat{\mathbf{Y}} = \mathbf{A}\mathbf{\Phi} \qquad \hat{\mathbf{Y}} \in R^{P\times N}, \mathbf{A} \in R^{P\times K}, \mathbf{\Phi} \in R^{K\times N}$$

where, P is the dimension of the spectral parameters. In Equation (1), both the event targets and event functions are unknown and the temporal decomposition analysis involves the determination of them once the speech parameter sequence of an utterance is given.

Each acoustic event in speech starts, gradually grows in magnitude and vanishes with a certain degree of overlapping between them. Therefore, the event functions which are representative of the temporal evolutionary patterns of these events should be; (i) *time-limited* to describe explicitly, the start and end points in time and the duration of each event, (ii) *non-negative* to describe the magnitude of the events during their existence, and (iii) *smooth* to describe the gradualness of growth and decay of the events resembling the gradualness of movement of the articulators in speech production. In the temporal decomposition analysis point of view, these properties of the event functions can be used as mathematical constraints in determining the event functions.

The concept of temporal decomposition of speech has attracted many researchers in the recent years, specially in application areas such as speech coding, recognition and segmentation. The fact that temporal decomposition decomposes the speech parameters into two elementary components, which occur at a lower rate than the original speech parameters, gives a means of coding speech efficiently at a lower bit rate (Cheng & O'Shaughnessy, 1991; Shiraki & Honda,

1993; Ghaemmaghami & Deriche, 1996; Lemma *et al.*, 1997). The strong relationship between the temporal decomposition representation of speech and the speech production mechanism has provided the necessary motivation to investigate its application in speech recognition (Bimbot *et al.*, 1988; Dijk-Kappers & Marcus, 1989). Its usefulness in speech segmentation has also been investigated (Dix & Bloothooft, 1994).

2. Atal's Method of Temporal Decomposition

Temporal decomposition of speech was first proposed (Atal, 1983) as a method for efficient coding of LPC parameters. Although the original implementation of temporal decomposition of speech (Atal, 1983) was mathematically solid, it is known to have the following two major drawbacks. (i) The method is computationally costly, making it impractical. (ii) High parameter sensitivity of the number and locations of the events. In other words, they are very sensitive to some trivial changes in the analysis parameters.

Atal's temporal decomposition method involves the following procedure. For a detailed mathematical treatment, the reader is referred to Atal (1983). First, the spectral parameter matrix of a windowed speech segment of about 200-300 ms is decomposed into two orthogonal matrices and a diagonal matrix of eigenvalues, using the so-called singular value decomposition.

$$Y^T = UDV^T$$

where, Y^T is the $N \times P$ spectral parameter matrix, U is a $N \times P$ orthogonal matrix, V is a $P \times P$ orthogonal matrix, and D is a diagonal matrix of eigenvalues. N and P are the number of frames in the windowed speech segment and the order of the spectral parameters, respectively. This allows the event functions to be expressed as a linear combination of a set of orthogonal functions, and also allows the number of events, M, to be fixed in the windowed speech segment under analysis, by taking into account only the number of significant eigenvalues. Normally, a window of about 200-300 ms gives $M = 5$.

$$\phi_k(n) = \sum_{i=1}^{M} b_{ki} u_i(n)$$

where, $u_i(n)$ is the element (n, i) of the matrix U and b_{ki} are a set of coefficients. Next, the nearest event function, $\phi(n)$, to the center of the windowed speech segment, $n = n_c$, is evaluated by considering the minimization of a distance measure, $\theta(n_c)$.

$$\theta(n_c) = \sqrt{\frac{\sum_{n=1}^{N}(n - n_c)^2 \phi^2(n)}{\sum_{n=1}^{N} \phi^2(n)}}$$

Minimization of $\ln(\theta(n_c))$, with respect to the coefficients b_i leads to an eigenvector problem of a matrix $R \in R^{K \times K}$.

$$R\mathbf{b} = \lambda\mathbf{b}$$

where the element (i, r) of the matrix R is given by,

$$R_{ir} = \sum_{n=1}^{N} (n - n_c)^2 u_i(n) u_r(n),$$

and $\mathbf{b}$ is the vector of coefficients b_i. The solution corresponding to the smallest eigenvalue λ provides the optimum $\mathbf{b}$.

In order to analyze a complete utterance the above procedure should be repeated with windows located at intervals through out the utterance. Atal's method requires the window to be shifted by a small interval, i.e. by a frame interval, to ensure that no event function is missed. Therefore, if the total number of windows is L, SVD and eigenvector solving should be performed L times. SVD is a highly involved computational procedure and this is known to be the major reason for the high computational complexity of the Atal's method.

Since the window is shifted at each time by a small interval, the same event function is generally found for several adjacent windows. In order to find the locations of the event functions, and to reduce the total set of event functions, a reduction algorithm based on a zero crossing criterion of a timing function, $\nu(l)$, is incorporated.

$$\nu(l) = \frac{\sum_{n=1}^{N} (n - l)\phi^2(n)}{\sum_{n=1}^{N} \phi^2(n)}$$

The function $\nu(l)$ crosses the $\nu(l) = 0$ axis from positive to negative at each location l which equals the location of one of the $\phi_k(n)$ for some k.

The spectral targets, $\mathbf{a}_k$, are determined by considering the minimization of the squared error between reconstructed and original spectral parameters, E_i, with respect to a_{ik}'s.

$$E_i = \sum_{n=1}^{N} \left(y_i(n) - \sum_{k=1}^{K} a_{ik} \phi_k(n) \right)^2, \quad 1 \leq i \leq P$$

where N and K are the total number of frames and events in the entire utterance. Finally, an iterative refinement procedure is used to improve the event function shapes and to reduce the reconstruction error. The refined set of event functions are evaluated by minimizing the reconstruction error, E_n, of spectral vectors.

$$E_n = \sum_{i=1}^{P} \left(y_i(n) - \sum_{k=1}^{K} a_{ik} \phi_k(n) \right)^2, \quad 1 \leq n \leq N$$

The resultant $\phi_k(n)$'s are used to obtain an even better estimates of the targets, $\mathbf{a}_k$'s. The procedure is repeated until both $\phi_k(n)$'s and $\mathbf{a}_k$'s converge to a set of stable values.

As described above the high computational cost of Atal's method (Atal, 1983) can be mainly attributed to the use of the computationally involved SVD, and

the repeated evaluation of the event functions at short time intervals before screening out the redundant event functions using a reduction algorithm. Marcus & Lieshout (1984) investigated the possible validity of TD as a method of determining phonetically plausible events in speech, but came out with the parameter sensitivity problem of the original method with respect to the number and locations of the event functions. In other words, they are very sensitive to some trivial changes in analysis parameters, i.e analysis window size, number of parameters retained after singular value decomposition, etc. Dijk-Kappers & Marcus (1989) improved the TD method to make events more stable, i.e. less parameter sensitive, but the computational cost has more or less remained the same because the time consuming SVD was still involved.

3. S^{2}BEL-TD of Spectral Parameters

The proposed new approach to temporal decomposition of speech, called Spectral Stability Based Event Localizing Temporal Decomposition (S^{2}BEL-TD), intends to overcome the drawbacks of the original method of Atal by implementing it in a mathematically simpler way, i.e. by avoiding SVD, while adopting a spectral stability criterion to determine the number and locations of the events. Given these number and locations, the subsequent computation of refined event targets and event functions is much less demanding than the traditional TD method. Also, this makes the number and locations of the events more parameter independent.

The S^{2}BEL-TD of Speech involves the following three computational steps.

STEP 1: Determination of the *event targets* (first approximation).

$$\mathbf{A}^{(0)} = \left[\ \mathbf{a}_k^{(0)} \ \right]_{1 \le k \le K}$$

STEP 2: Determination of the *event functions* (first approximation).

$$\mathbf{\Phi}^{(0)} = \left[\ \phi_k(n)^{(0)} \ \right]_{1 \le k \le K,\ 1 \le n \le N}$$

STEP 3: Iterative refinement of *event targets* & *event functions*.

$$(\mathbf{A}^{(0)}, \mathbf{\Phi}^{(0)}) \Rightarrow (\mathbf{A}^{(1)}, \mathbf{\Phi}^{(1)}) \Rightarrow \cdots (\mathbf{A}^{(S)}, \mathbf{\Phi}^{(S)})$$

The superscript notation indicates the iteration step number. The details of the Steps 1, 2, and 3 are given in the Sections 3.1, 3.2, and 3.3, respectively.

3.1. Determination of Event Targets

The determination of the first approximation of the event targets is based on a maximum spectral stability criterion. The spectrally stable points in speech are used as a hint for the locations where speech events exist. It is assumed that each acoustic event that exists in speech gives rise to a spectrally stable point in

its neighborhood. Therefore, the locations of the spectrally stable points and the corresponding spectral parameter sets can be used as a good approximation to the event locations and event targets, respectively. Because of this use of points of maximum spectral stability for event detection, the new approach is termed *spectral stability based event localizing* temporal decomposition.

The transition rate of the ith spectral parameter, $y_i(n)$, at the time point n is calculated as the gradient of the best fitting straight line, i.e. regression line, within the time window $[n-M, n+M]$, as given in Equation (2). The squared sum of these transition rates of individual spectral parameters, $y_i(n)$, where $1 \leq i \leq P$, is defined as the Spectral Feature Transition Rate (SFTR) at the time point n, and is given by Equation (3).

$$c_i(n) = \frac{\sum_{m=-M}^{M} m y_i(n+m)}{\sum_{m=-M}^{M} m^2}, \quad 1 \leq i \leq P \tag{2}$$

$$\text{SFTR}: \quad s(n) = \sum_{i=1}^{P} c_i(n)^2, \quad 1 \leq n \leq N \tag{3}$$

The local minima of $s(n)$ indicate the frames with maximum local spectral stability in speech, and these points are considered as the approximate locations of the events, and the corresponding spectral parameter vectors as the initial approximation of the event targets. Therefore, if the local minima of $s(n)$ are at $n_1, n_2, .., n_K$, where $n_1 < n_2 < ... < n_K$, the initial approximation of the event target matrix, $\mathbf{A}^{(0)}$, can be formed as;

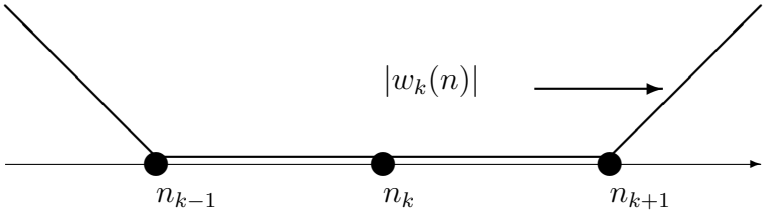
$$\begin{aligned} \mathbf{A}^{(0)} &= \left[\begin{array}{cccc} \mathbf{a}_1^{(0)} & \mathbf{a}_2^{(0)} & \cdots & \mathbf{a}_K^{(0)} \end{array} \right] \\ &= \left[\begin{array}{cccc} \mathbf{y}(n_1) & \mathbf{y}(n_2) & \cdots & \mathbf{y}(n_K) \end{array} \right] \end{aligned}$$

The number of events, K, and their locations, $n_1 < n_2 < ... < n_K$, are determined through the SFTR analysis. Therefore, the window size, $2M$, of SFTR analysis is the only parameter that effects the number and locations of the events in the S^{2}BEL-TD algorithm.

3.2. Determination of Event Functions

Since the speech events exist only for a limited time duration in continuous speech, event functions should be time limited. This makes it necessary to add a constraint to this effect, when evaluating them. This is achieved using a weighting function, $w_k(n)$, corresponding to each event function, $\phi_k(n)$. The weighting function $w_k(n)$ for the kth event function is defined as follows.

$$w_k(n) = \begin{cases} n_{k-1} - n, & \text{if } 1 \leq n < n_{k-1} \\ 0, & \text{if } n_{k-1} \leq n \leq n_{k+1} \\ n - n_{k+1}, & \text{if } n_{k+1} < n \leq N \end{cases}$$



$$\mathbf{w}_k = \left[\begin{array}{cccc} w_k(1) & w_k(2) & \cdots & w_k(N) \end{array} \right]$$

$$\mathbf{W} = \left(\begin{array}{c} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \cdots \\ \mathbf{w}_K \end{array} \right) \in R^{K \times N}$$

where, $\mathbf{W}$ is called the weighting function matrix. The reason for the above definition of the weighting function can be justified as follows. It is known that event function $\phi_k(n)$ exists around its center n_k. But, little is known about its length, i.e. the duration of its existence, at this stage. Since spectral properties would be less governed by the kth event in the region beyond the centers of the adjacent events, n_{k-1} and n_{k+1}, $\phi_k(n)$ should be fairly small in amplitude and gradually decreasing in this region. Therefore, $w_k(n)$ is set to zero in between the adjacent event centers, and is linearly increased beyond those points. This provides the event function total freedom to show its temporal behavior between the points of adjacent event centers, n_{k-1} and n_{k+1}, but a decreasing degree of freedom beyond those points.

Although, n_{k-1} and n_{k+1} may not be the best limits for the event function $\phi_k(n)$, they are used at this stage to evaluate the first approximation of the event functions. In Section 3.3 the use of adaptive weighting functions with adaptive limits for the events is described as a part of the refinement process. By considering the columns of the matrix $\mathbf{W}$, diagonal matrices are formed as;

$$\mathbf{W}_n = \text{diag} \left[\begin{array}{cccc} w_1(n) & w_2(n) \cdots & w_K(n) \end{array} \right] \in R^{K \times K}$$

The functional $J(\boldsymbol{\phi}_n, \lambda)$ is formulated by taking into account the sum of the squared error between the original and the reconstructed spectral parameters, and a constraint to limit the spreading of event functions in time, as given in Equation (4).

$$J(\boldsymbol{\phi}(n), \lambda) = \sum_{i=1}^{P} (y_i(n) - \hat{y}_i(n))^2 + \lambda \sum_{k=1}^{K} w_k(n)^2 \phi_k(n)^2, \quad 1 \leq n \leq N \qquad (4)$$

where λ is a constant weighting factor and,

$$\boldsymbol{\phi}(n) = \left[\begin{array}{cccc} \phi_1(n) & \phi_2(n) & \cdots & \phi_K(n) \end{array} \right]^T, \quad 1 \leq n \leq N$$

$y_i(n)$ and $\hat{y}_i(n)$ are the i^{th} element of the spectral vectors $\mathbf{y}(n)$ and $\hat{\mathbf{y}}(n)$, respectively.

$\boldsymbol{\phi}(n)$, where $1 \leq n \leq N$, is determined by considering the minimization of the functional $J(\boldsymbol{\phi}(n), \lambda)$ with respect to $\boldsymbol{\phi}(n)$ as follows.

$$\begin{aligned} \frac{\partial J(\boldsymbol{\phi}(n), \lambda)}{\partial \phi_r(n)} &= \sum_{i=1}^{P} 2 \left(\sum_{k=1}^{K} a_{ik}\phi_k(n) - y_i(n) \right) a_{ir} + 2\lambda w_r(n)^2 \phi_r(n) \\ &= 0 \end{aligned}$$

$$\sum_{i=1}^{P} a_{ir} \left(\sum_{k=1}^{K} a_{ik}\phi_k(n) \right) + \lambda w_r(n)^2 \phi_r(n) = \sum_{i=1}^{P} a_{ir} y_i(n), \qquad 1 \leq r \leq K \tag{5}$$

Conversion of Equation (5) into matrix notation results in;

$$\mathbf{A}^T\mathbf{A}\boldsymbol{\phi}(n) + \lambda \mathbf{W}_n^T \mathbf{W}_n \boldsymbol{\phi}(n) = \mathbf{A}^T \mathbf{y}(n)$$

$$\boldsymbol{\phi}(n) = \left(\mathbf{A}^T\mathbf{A} + \lambda {\mathbf{W}_n}^T \mathbf{W}_n \right)^{-1} \mathbf{A}^T \mathbf{y}_n, \quad 1 \leq n \leq N \tag{6}$$

Therefore, the first approximation of the event function matrix, $\boldsymbol{\Phi}^{(0)}$, can be formed as;

$$\boldsymbol{\Phi}^{(0)} = \begin{pmatrix} \boldsymbol{\phi}(1) & \boldsymbol{\phi}(2) & \cdots & \boldsymbol{\phi}(N) \end{pmatrix} \tag{7}$$

The weighting factor λ in the functional $J(\boldsymbol{\phi}(n), \lambda)$ determines the relative weighting between the two error terms involved. A suitable value for λ is to be selected, based on simulation results. This value of λ used to determine the first approximation of the event functions is referred to as $\lambda^{(0)}$ in the Sections followed.

3.3. Iterative Refinement Procedure

An iterative refinement procedure is adopted to improve the shapes of the event functions and the reconstruction accuracy of TD, and to refine the event targets. The initial event functions show undesirable minor lobes, i.e. negative ripples, apart from the desirable major lobes as shown in Fig. 1. This violates the non-negativity property imposed on the event functions. The iterative refinement procedure effectively smooths-out the minor lobes while allowing the major lobes to evolve freely. It also improves the reconstruction accuracy of TD and refines the event targets. This involves the recursive performance of the procedures described in the Sections 3.3.1 and 3.3.2. Generally, 4 to 5 iterations are required to shape up the event functions.

3.3.1. Refinement of Event Functions

Event functions are recalculated using the procedure of Section 3.2, but with an adaptive weighting function and the quantitative balancing of the two error-terms of the functional $J(\boldsymbol{\phi}(n), \lambda)$, as described below.

$$(\mathbf{A}^{(l-1)}, \boldsymbol{\Phi}^{(l-1)}) \rightarrow \boldsymbol{\Phi}^{(l)}, \quad 1 \leq l \leq S$$

where, l and S are the iteration step number and total number of iterations, respectively.

Adaptive Weighting function:
An adaptive weighting function is defined as given in Equation (8). It is adaptive to the major-lobe limits of the event functions.

$$w_k^{(l)}(n) = \begin{cases} l_k^{(l-1)} - n, & \text{if } 1 \leq n < l_k^{(l-1)} \\ 0, & \text{if } l_k^{(l-1)} \leq n \leq r_k^{(l-1)} \\ n - r_k^{(l-1)}, & \text{if } r_k^{(l-1)} < n \leq N \end{cases} \tag{8}$$

Where, $l_k^{(l-1)}$ and $r_k^{(l-1)}$ are the left and right limits of the major lobe of the event function $\phi_k(n)^{(l-1)}$. This definition of adaptive weighting function restricts the minor-lobes while allowing the major-lobe to evolve freely. Therefore, it gives rise to major-lobe expansion and contraction, with a simultaneous minor-lobe reduction, when the iterations are performed.

Quantitative Balancing of the functional $J(\boldsymbol{\phi}(n), \lambda)$:
Weighting factor $\lambda^{(l)}$ at the iteration step l is selected so as to balance the two error terms of the functional $J(\boldsymbol{\phi}(n), \lambda)$ using the results obtained at the iteration step $(l-1)$, i.e. $\boldsymbol{\Phi}^{(l-1)}$ and $\mathbf{A}^{(l-1)}$, as given below.

$$\lambda^{(l)} = \sigma \times \left(\frac{\sum_{n=1}^{N} \sum_{i=1}^{P} \left(y_i(n) - \hat{y}_i{}^{(l-1)}(n) \right)^2}{\sum_{n=1}^{N} \sum_{k=1}^{K} w_k^{(l)}(n)^2 \phi_k^{(l-1)}(n)^2} \right)$$

where, $\hat{y}_i{}^{(l-1)}(n) = \sum_{k=1}^{K} a_{ik}^{(l-1)} \phi_k^{(l-1)}(n)$, and σ is the constant balancing ratio. The event functions matrix, $\boldsymbol{\Phi}^{(l)}$, at the iteration step l is calculated as follows, similar to the Equations (6) and (7).

$$\boldsymbol{\phi}(n)^{(l)} = \left(\mathbf{A}^{(l-1)^T} \mathbf{A}^{(l-1)} + \lambda^{(l)} \mathbf{W}_n^{(l)^T} \mathbf{W}_n^{(l)} \right)^{-1} \mathbf{A}^{(l-1)^T} \mathbf{y}_n, \quad 1 \leq n \leq N$$

where,

$$\mathbf{W}_n^{(l)} = \text{diag} \left[\begin{array}{cccc} w_1^{(l)}(n) & w_2^{(l)}(n) & \cdots & w_K^{(l)}(n) \end{array} \right]$$

Hence,

$$\boldsymbol{\Phi}^{(l)} = \left(\begin{array}{cccc} \boldsymbol{\phi}(1)^{(l)} & \boldsymbol{\phi}(2)^{(l)} & \cdots & \boldsymbol{\phi}(N)^{(l)} \end{array} \right)$$

3.3.2. Refinement of Event Targets

Refinement of event targets involves the recalculation of them by minimizing the squared error between the original and the reconstructed spectral parameters, with respect to the target vectors. Event targets at the lth iteration are calculated from the event functions at the lth iteration, as described below.

$$\boldsymbol{\Phi}^{(l)} \rightarrow \mathbf{A}^{(l)}, \quad 1 \leq l \leq S$$

The squared error between the original and reconstructed ith spectral parameter at the iteration step l can be expressed as follows.

$$E_i^{(l)} = \sum_{n=1}^{N} \left(y_i(n) - \sum_{k=1}^{K} a_{ik}^{(l)} \phi_k^{(l)}(n) \right)^2, \quad 1 \le i \le P$$

By setting the partial derivative of $E_i^{(l)}$ with respect to a_{ir}, to zero;

$$\begin{aligned} \frac{\partial E_i^{(l)}}{\partial a_{ir}} &= \sum_{n=1}^{N} \left(y_i(n) - \sum_{k=1}^{K} a_{ik}^{(l)} \phi_k^{(l)}(n) \right) \left(-2\phi_r(n)^{(l)} \right) \\ &= 0 \\ \sum_{k=1}^{K} a_{ik}^{(l)} \sum_{n=1}^{N} \phi_k^{(l)}(n) \phi_r^{(l)}(n) &= \sum_{n=1}^{N} y_i(n) \phi_r^{(l)}(n) \end{aligned} \tag{9}$$

where, $1 \le r \le K, \quad 1 \le i \le P$

Equation (9) gives P sets of K variable simultaneous equations, using which $a_{ik}^{(l)}$, where $1 \le k \le K$ and $1 \le i \le P$, could be evaluated. Therefore, the event target matrix at the iteration step l can be formed as follows.

$$\mathbf{A}^{(l)} = \left[\; a_{ik}^{(l)} \; \right]_{1 \le i \le P, \quad 1 \le k \le K}$$

3.3.3. Termination and Convergence of Iterations

The two steps (3.3.1) and (3.3.2) are repeatedly performed until the minor lobe content, $MLC^{(l)}$, drops below a certain predetermined threshold level, e.g.1%. Minor lobe content, $MLC^{(l)}$,at the lth iteration step is defined as follows.

$$MLC^{(l)} = \sqrt{\frac{\sum_{k=1}^{K} \sum_{n=1}^{N} \phi_k^{(l)}(n)^2 c_k^{(l)}(n)}{\sum_{k=1}^{K} \sum_{n=1}^{N} \phi_k^{(l)}(n)^2}} \times 100\%$$

where,

$$c_k^{(l)}(n) = \begin{cases} 0, & \text{if } l_k^{(l)} \le n \le r_k^{(l)} \\ 1, & \text{otherwise} \end{cases}$$

where, $l_k^{(l)}$ and $r_k^{(l)}$ are the left and right limits of the major lobe of the kth event function, at the lth iteration step. Also, we define the root-mean-squared-error between the original and reconstructed spectral parameters, at the lth iteration step as follows.

$$E_{rms}^{(l)} = \sqrt{\frac{1}{NP} \sum_{n=1}^{N} \sum_{i=1}^{P} \left(y_i(n) - \hat{y}_i{}^{(l)}(n) \right)^2}$$

Convergence of $MLC^{(l)}$ and $E_{rms}^{(l)}$ with the iteration step number l is an important property for the iterative refinement procedure. Simulation results show that good convergence can be achieved by properly selecting the parameters $\lambda^{(0)}$ and σ.

3.4. Segmental S^{2}BEL-TD

The present algorithm of S^{2}BEL-TD analysis takes the total length of the input speech as a block for the TD analysis. Although there is no problem with this for word utterances and short sentence utterances, for relatively long utterances with more than about 500 frames, taking the whole utterance as a single segment for TD analysis proves time consuming. This can be simply attributed to the large dimension of the matrices involved in the computational procedure. This makes it necessary to develop the TD analysis algorithm so that it will work on short speech blocks, or segments, when analyzing a long utterance of input speech. This is termed segmental S^{2}BEL-TD analysis. On the other hand, if S^{2}BEL-TD is to be used in any kind of real time analysis, segmental analysis becomes inevitable.

The implementation of segmental analysis is based on a mutually non-interacting events criterion. Let E_i and E_j be two events with event functions $\phi_i(n)$ and $\phi_j(n)$. The indices i and j describe the chronological order of the two events E_i and E_j. The two event E_i and E_j are called mutually non-interacting if the following condition is satisfied.

$$\sum_{n=1}^{N} \phi_i(n)\phi_j(n) = 0$$

$$\text{i.e} \quad \phi_i(n)\phi_j(n) = 0, \quad 1 \leq n \leq N$$

This means that either $\phi_i(n)$ or $\phi_j(n)$ is zero at all time points n. This situation can be easily visualized as two non-overlapping event functions. Obviously, if the events E_i and E_j are separated in time by a sufficient number of intermediate events they would be mutually non-interacting. We are interested in the minimum l, let this be L, such that,

$$\sum_{n=1}^{N} \phi_i(n)\phi_j(n) = 0, \quad \text{if } |i-j| > L$$

By simple observation of TD results over a large set of speech data it was confirmed that $L = 3$. This means that two event functions with at least 3 intermediate events, do not overlap. Therefore, an event could be accurately evaluated without any unaccounted mutual effects, if the speech segment contains at least 3 adjacent events to both sides. In speech production point of view this may mean that the feed-forward and feed-back co-articulation do not occur over more than 3 acoustic events.

Using the above result an algorithm for the segmental TD analysis is developed as follows. Input speech is segmented with at least $2L$ events in the overlapping region between two adjacent segments. In each segment we neglect the first and last L events as inaccurate due to unaccounted mutual effects, except for the first and last segment of the input speech. In the first segment, only the last L events are neglected, and in the last segment, only the first L events are neglected. The segment size is kept fixed around 100 frames.

4. Simulation Results

The ATR Japanese and the TIMIT English speech database were used for the speech data. Both Log Area Ratio (LAR) parameters and Line Spectral Frequency (LSF) parameters were considered as a candidate spectral parameter for the S^{2}BEL-TD. LAR parameters have given better results, i.e. better reconstruction accuracy, in temporal decomposition (Dijk-Kappers, 1989) over the other LPC related spectral parameters. LSF parameters have been known to have the best interpolation properties (Paliwal, 1995; Choi *et al.*, 1995), i.e. linear combination-ability. S^{2}BEL-TD was implemented on both LAR and LSF parameters and their reconstruction accuracies are compared as a part of the performance evaluation of the method. 10th order LAR and LSF parameters were calculated using a LPC analysis window of $2M = 40$ ms at 10 ms frame intervals, from 8 kHz sampling speech files.

The male Japanese word utterance "*aikawarazu*" was used with a SFTR analysis window size of 40 ms to investigate the convergence properties of S^{2}BEL-TD algorithm. The spectral parameter is LAR and simulations were performed for $\lambda^{(0)}$ values of 10, 5, 1, 0.2 and 0.1. The initial minor lobe content, $MLC^{(0)}$, and the initial RMS error between reconstructed and original spectral parameters, $E^{(0)}_{rms}$, obtained for different values of $\lambda^{(0)}$ are shown in the Fig. 2. A high value for $\lambda^{(0)}$ causes a high reconstruction error and a relatively low $MLC^{(0)}$, while a low value for $\lambda^{(0)}$ causes a relatively low reconstruction error and a high $MLC^{(0)}$. Fig. 3 shows the typical shape of initial event functions, $\phi_k(n)^{(0)}$ for some k, for different values of the initial weighting factor $\lambda^{(0)}$.

The iterative refinement of the event functions and the targets was performed according to the procedure described in Section 3.3. The initial weighting factor $\lambda^{(0)}$ and the balancing ratio σ are constant to be set appropriately according to the simulation results. Simulation was performed for $\lambda^{(0)}$ values of 10, 1, 0.2 and for σ values of 5, 1, 0.2 while maintaining $\sigma = 1$ and $\lambda^{(0)} = 0.2$, respectively. The convergence patterns of the reconstruction error ($E^{(l)}_{rms}$ against l) are shown in the Fig. 4 and Fig. 5. Reconstruction error decreases and reaches a certain minimum after a few iterations. Fig. 6 shows the effect of the iterative refinement on the event function shapes. Minor lobe content decreases and becomes almost negligible after a few iterations. The minor lobe smoothing and major lobe reshaping can be observed as desirable effects of the refinement procedure.

In Fig. 7, a plot of SFTR and the final event functions are shown for the female English sentence utterance "*we always thought we would die with our boots on*". The spectral parameter is LSF, which has the same tendency as LAR but the magnitude of $\lambda^{(0)}$ is different. Here $\lambda^{(0)} = 0.005$, $\sigma = 1$ were selected as appropriate values for the initial weighting factor and balancing ratio, respectively. SFTR window size of $2M = 40$ ms was selected resulting in an average event rate of about 20 events/sec. The speech waveform of the utterance is also shown together with the phonetic transcription for reference. The window size, $2M$, of SFTR analysis is the only parameter that effects the number and locations of the

events in the S^{2}BEL-TD algorithm. It controls the event rate, and can be appropriately selected to achieve the optimal performance of S^{2}BEL-TD for different applications. In speech coding point of view, window size, $2M$, can be selected so as to obtain a certain optimal tradeoff between reconstruction accuracy of the spectral parameters (spectral distortion) and the bit rate. In speech decoding, it can be selected to optimize the correlation between phonemes/sub-phonemes and events.

5. Performance Evaluation

In this section, the performance of S^{2}BEL-TD in terms of interpolation property, computational complexity, and stability of the number and locations of the events were evaluated.

Spectral Distortion (SD) is a commonly used measure in evaluating the performance of LPC quantization (Shiraki & Honda, 1993) and interpolation (Paliwal, 1995). SD measure is also used for evaluating the interpolation performance of the proposed S^{2}BEL-TD algorithm. The spectral distortion evaluated is that between the original spectral parameters, $\mathbf{y}(n)$, and the reconstructed, or synthesized, spectral parameters, $\hat{\mathbf{y}}(n)$.

The results are provided in terms of spectral distortion histograms, average spectral distortion and percentage outliers having spectral distortion greater than 2 dB. The outliers are divided into the following two types. Type 1: consists of outliers in the range 2-4 dB, and Type 2: consists of outliers having spectral distortion greater than 4 dB. Spectral distortion, D_n, for the nth frame is defined (in dB) as follows.

$$D_n^2 = \frac{1}{F_s}\int_0^{F_s}[10log_{10}(P_n(f)) - 10log_{10}(\hat{P}_n(f))]^2 df$$

where F_s is the sampling frequency, and $P_n(f)$ and $\hat{P}_n(f)$ are the LPC power spectra corresponding to the nth frame of the original spectral parameters, $\mathbf{y}(n)$, and the reconstructed spectral parameters, $\hat{\mathbf{y}}(n)$, respectively.

A set of 250 sentence utterances of the ATR Japanese speech database and another set of 192 sentence utterances of the TIMIT English speech database were selected for spectral distortion evaluation. The Japanese speech data set consists of about 20 minutes of speech from 10 speakers (5 male & 5 female). Meanwhile, the English speech data set contains 24 speakers, 2 male and 1 female from each of 8 dialect regions. Each speaker read a different set of 5 phonetically-compact sentences (the SX sentences) and 3 phonetically-diverse sentences (the SI sentences). Both LAR and LSF parameters were calculated, and S^{2}BEL-TD analyzed. SD was calculated on a frame-by-frame basis.

Table I & Table II give the summary of the spectral distortion results obtained for the above sets of utterances with LAR and LSF as the spectral parameter. The distribution of the spectral distortion in the form of histograms are shown

in Fig. 8 and Fig. 9, each for both cases of LAR and LSF parameters concerning with one speech data set. Results indicate slightly better performance in the case of LSF parameters over LAR parameters.

(Table I & Table II here)

Since the S^{2}BEL-TD aims at overcoming the two drawbacks of high computational cost, and the high parameter sensitivity of the number and locations of the events imposed on the Atal's method, it is necessary to evaluate the performance of S^{2}BEL-TD on these aspects. With respect to computational complexity the S^{2}BEL-TD shows a significant improvement over the original method by Atal. This can be mainly attributed to the fact that the SVD is the most time consuming part of the Atal's method (Dijk-Kappers & Marcus, 1989) and the SVD is not required for S^{2}BEL-TD. Moreover, the S^{2}BEL-TD was implemented in a mathematically simpler way than that of Atal's method. The instability problem of the number and locations of the events with respect to TD analysis window size and the number of parameters retained after SVD, has also been overcome in S^{2}BEL-TD. It has been emphasized in Section 3.1 that the window size of SFTR analysis is the only parameter that effects the number and locations of the events in the S^{2}BEL-TD method. Since SFTR is a local measure, the TD analysis window size makes no difference in the number and locations of the event functions found. But this is not the case in the original method by Atal, where even a trivial change in window size or number of parameters retained after SVD leads to a dramatic changes in the number and locations of the event functions. Investigation of Atal's method by Marcus & Lieshout (1984) has revealed this fact.

In addition, the S^{2}BEL-TD was used for analyzing a considerable number of speech utterances spoken by different speakers (male & female) in different speech conditions and worked satisfactorily. All speech utterances were well S^{2}BEL-TD analyzed using the same parameters, i.e. the initial weighting factor $\lambda^{(0)}$ and the balancing ratio σ.

6. S^{2}BEL-TD of Excitation Parameters

In this section, the application of S^{2}BEL-TD technique to speech excitation parameters and some simulation results are presented.

6.1. Determination of Excitation Targets

The S^{2}BEL-TD technique is employed to describe the temporal characteristics of the speech excitation parameters, i.e gain, pitch and voicing. The same event functions evaluated for the spectral parameters are used to describe the temporal pattern of the gain, pitch and voicing parameters also. The speech production mechanism is assumed to be a synchronously controlled process with respect to the movement of different articulators, i.e. jaws, tongue, larynx, glottis etc., and

therefore the temporal evolutionary patterns of different properties of speech, i.e. spectrum, pitch, gain and voicing, can be described by a common set of event functions.

Let $b(n)$ be an excitation parameter, i.e. gain, pitch or voicing. Then $b(n)$ is approximated by $\hat{b}(n)$, the reconstructed excitation parameter for the nth frame, as follows in terms of excitation targets, b_k's, and the event functions, $\phi_k(n)$'s.

$$\hat{b}(n) = \sum_{k=1}^{K} b_k \phi_k(n), \quad 1 \leq n \leq N \tag{10}$$

In matrix notation, Equation (10) can be written as;

$$\hat{B} = A_b \mathbf{\Phi}$$

where $\hat{B}$ and A_b are the reconstructed excitation parameter vector and excitation target vector, respectively.

In Equation (10), the event functions, $\phi_k(n)$'s, are known and therefore the excitation targets, b_k's, are determined by minimizing the squared error between the original excitation parameters and the reconstructed excitation parameters as follows.

$$E_b = \sum_{n=1}^{N} \left(b(n) - \sum_{k=1}^{K} b_k \phi_k(n) \right)^2$$

By setting the partial derivative of E_b with respect to b_r, to zero;

$$\begin{aligned} \frac{\partial E_b}{\partial b_r} &= \sum_{n=1}^{N} \left(b(n) - \sum_{k=1}^{K} b_k \phi_k(n) \right) (-2\phi_r(n)) \\ &= 0, \end{aligned}$$

$$\sum_{k=1}^{K} b_k \sum_{n=1}^{N} \phi_k(n)\phi_r(n) = \sum_{n=1}^{N} b(n)\phi_r(n), \quad 1 \leq r \leq K \tag{11}$$

Equation (11) gives a set of K variable simultaneous equations, using which b_k, where $1 \leq k \leq K$, could be evaluated.

In the case of pitch parameters, linear interpolation was used within the unvoiced segments to form a continuous pitch contour. In the case of voicing parameters, a hard limiter with a threshold value of 0.5 was used to determine the reconstructed binary voicing parameters and binary voicing targets, from the non-binary results of Equation (10) and (11), respectively.

6.2. Simulation Results

The gain, pitch and voicing parameters, hereafter indicated by $g(n)$, $p(n)$, and $v(n)$, respectively, were calculated at 10 ms frame intervals with a 40 ms analysis window, for the sentence utterance "*kantan na shiryou wo ookuri shimasu node,*

shibaraku omachi kudasai ", of the ATR Japanese speech database. Each parameter contour was S^{2}BEL-TD analyzed according to the procedure described in the Section 6.1 with the event functions obtained from S^{2}BEL-TD analysis of LSF parameters.

Fig. 10 shows the plots of original and reconstructed gain parameters and the plot of frame-wise gain error, $e_g(n)$, where $e_g(n) = \hat{g}(n) - g(n)$. The RMS gain error, $\sqrt{E_g}$, where $E_g = \frac{1}{N}\sum_{n=1}^{N} {e_g}^2(n)$, was found to be about 4 dB. Fig. 11 shows the plots of original and reconstructed pitch frequency parameters and the plot of frame-wise pitch frequency error, $e_p(n)$, where $e_p(n) = \hat{p}(n) - p(n)$. The RMS pitch error, $\sqrt{E_p}$, where $E_p = \frac{1}{N}\sum_{n=1}^{N} {e_p}^2(n)$, was found to be about 2.3 Hz. In the case of binary voicing parameters, the voicing error, $e_v(n)$, where $e_v(n) = \hat{v}(n) - v(n)$, appeared only at, but not all, voiced/unvoiced boundaries as error spikes of mostly 1 frame. The percentage number of frames with voicing errors was found to be about 4%.

Moreover, the performance of S^{2}BEL-TD in terms of excitation parameters has also been evaluated over the set of 250 Japanese sentence utterances and the set of 192 English sentence utterances used in Section 5. The RMS gain error, RMS pitch error and percentage number of frames with voicing errors were found about 4 dB, 6 Hz and 5%, respectively. It was observed that the RMS gain error and RMS pitch error can be mainly attributed to some discrete time points, where the corresponding frame-wise gain error and pitch error obtained very high values. Meanwhile, no voicing errors were observed during continuous voiced and unvoiced segments, except for the points of voicing transitions.

The significant match between the original and reconstructed excitation parameters results in the fact that a common set of event functions can be used to describe the temporal patterns of both spectral and excitation parameters.

7. Conclusions

This paper presents a new approach to temporal decomposition of speech. The spectral stability criterion used in event localizing, and the use of adaptive weighting functions in determining the event functions, can be highlighted as the main features of the proposed S^{2}BEL algorithm for TD. The former makes the event localization more parameter independent eventually overcoming the instability problem of the Atal's method. The latter gives a great degree of freedom to the event functions to evolve through iterations. Also, the S^{2}BEL algorithm which makes no use of SVD algorithm and the redundant calculation of event functions, can be considered as a significant improvement in terms of computational cost compared to the original method by Atal. On continuous speech S^{2}BEL-TD can be performed on a segmental basis. The representation of speech excitation parameters also in terms of excitation targets and event functions makes S^{2}BEL-TD a complete higher-level parametric model of speech. With these improvements, S^{2}BEL-TD has the potential to become a strong tool

in analyzing speech, from which researchers working on speech coding, recognition and synthesis may profit.

References

Atal, B.S., Efficient coding of LPC parameters by temporal decomposition, *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (1983). 81–84.

Bimbot, F., Chollet, G., Deleglise, P. & Montacie, C., Temporal decomposition and acoustic-phonetic decoding of speech, *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (1988). 445–448.

Cheng, Y.M. & O'Shaughnessy, D., Short-term temporal decomposition and its properties for speech compression, *IEEE Transactions on Signal Processing*, (1991). **39**, 1282–1290.

Choi, H.B., Wong, W.T.K., Cheetham, B.M.G. & Goodyear, C.C., Interpolation of Spectral Information for Low Bit Rate Speech Coding, *Proceedings of the European Conference on Speech Communication and Technology*, (1995). 1033–1036.

Dijk-Kappers, A.M.L.V. & Marcus, S.M., Temporal decomposition of speech, *Speech Communications*, (1989). **8**, 125–135.

Dijk-Kappers, A.M.L.V., Comparison of parameter sets for temporal decomposition, *Speech Communications*, (1989). **8**, 203–220.

Dix, P.J. & Bloothooft, G., A breakpoint analysis procedure based on temporal decomposition, *IEEE Transactions on Speech and Audio Processing*, (1994). **2**, 9–17.

Ghaemmaghami, S. & Deriche, M., A new approach to very low-rate speech coding using temporal decomposition, *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (1996). 224–227.

Lemma, A.N., Kleijn, W.B. & Deprettere, E.F., LPC Quantization Using Wavelet Based Temporal Decomposition of the LSF, *Proceedings of the European Conference on Speech Communication and Technology*, (1997). 1259–1262.

Marcus, S.M. & Van-Lieshout, R.A.J.M., Temporal decomposition of speech, *IPO Annual Progress Report 19*, (1984). 26–31.

Paliwal, K.K., Interpolation properties of linear prediction parametric representations, *Proceedings of the European Conference on Speech Communication and Technology*, (1995). 1029–1032.

Shiraki, Y. & Honda, M., Extraction of temporal pattern of spectral sequence and its quantization performance, *Proceedings of 1993 Spring Meeting of the Acoustical Society of Japan*, (1993). 1-403–404.

Table 1: Average spectral distortion and percentage number of outlier frames for LAR's and LSF's. The speech data set consists of 250 sentence utterances spoken by 10 speakers (5 male & 5 female) of the ATR Japanese speech database

Parameter	Avg. SD (dB)	$\leq$ 2 dB	2-4 dB	$>$ 4 dB
LAR	1.7831	69.0%	26.8%	4.2%
LSF	1.4643	80.6%	18.5%	0.9%

Table 2: Average spectral distortion and percentage number of outlier frames for LAR's and LSF's. The speech data set consists of 192 sentence utterances spoken by 24 speakers (2 male & 1 female from each of 8 dialect regions) of the TIMIT English speech database

Parameter	Avg. SD (dB)	$\leq$ 2 dB	2-4 dB	$>$ 4 dB
LAR	1.6863	72.7%	23.7%	3.6%
LSF	1.4778	79.9%	19.0%	1.1%

Figure 1: Typical shape of an initial event function. Note the presence of undesirable minor lobes, i.e. negative ripples, in addition to the desirable major lobe.

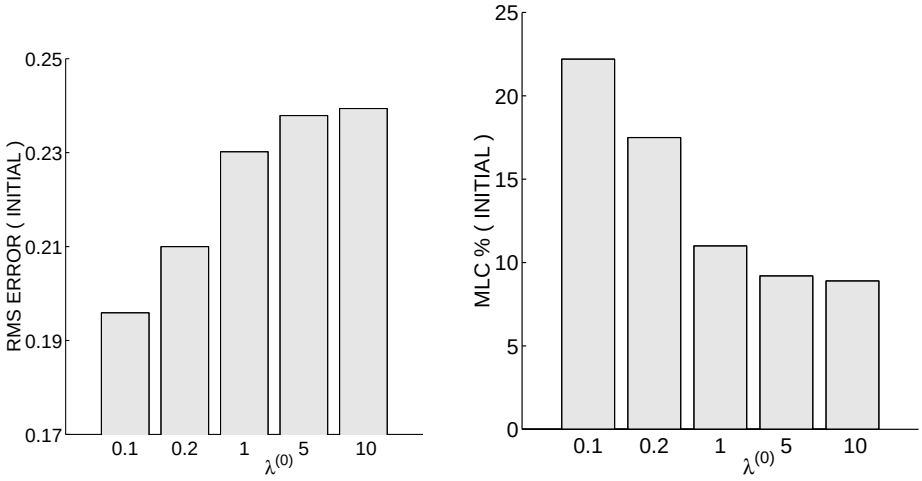


Figure 2: Initial RMS Error between original and reconstructed spectral parameters, $E_{rms}^{(0)}$ (left), and initial minor lobe content, $MLC^{(0)}$ (right), for different values of $\lambda^{(0)}$, as bar plots. Note that $MLC^{(0)}$ decreases, but $E_{rms}^{(0)}$ increases with increasing $\lambda^{(0)}$.

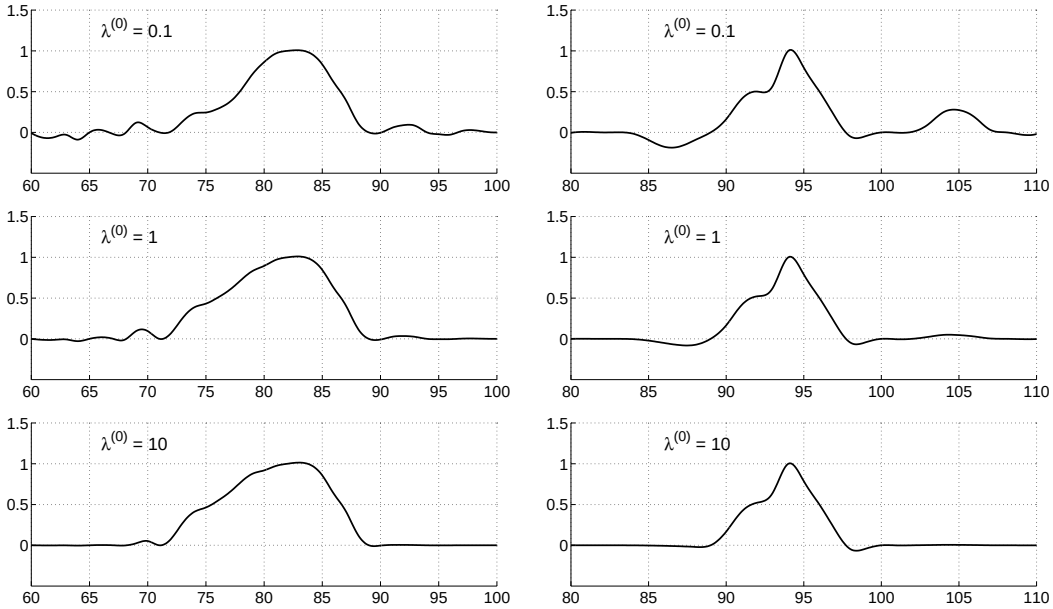


Figure 3: Typical shape of the Initial event functions, $\phi_k(n)^{(0)}$, for some k. Note that $MLC^{(0)}$ increases as $\lambda^{(0)}$ decreases.

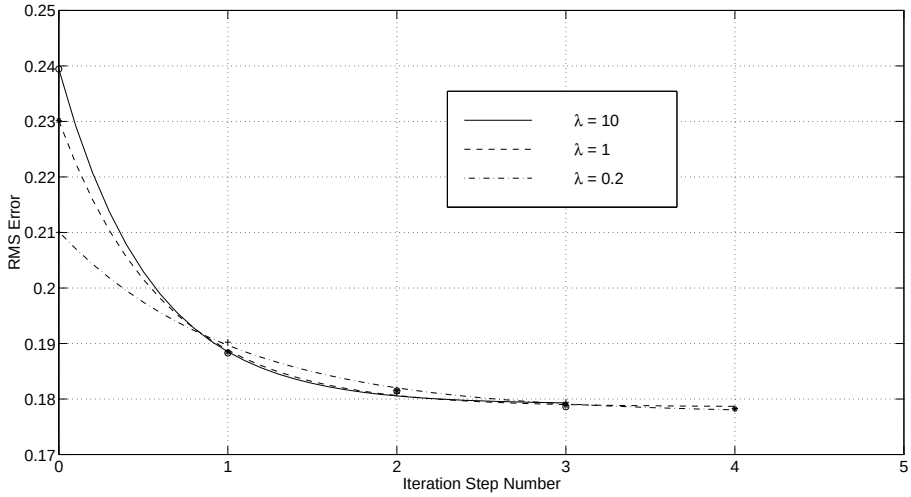


Figure 4: Convergence patterns of the reconstruction error, $E_{rms}^{(l)}$, with iteration step l, for different values of $\lambda^{(0)}$. Balancing ratio is $\sigma = 1$. Note that after few iterations $E_{rms}^{(0)}$ reaches a minimum.

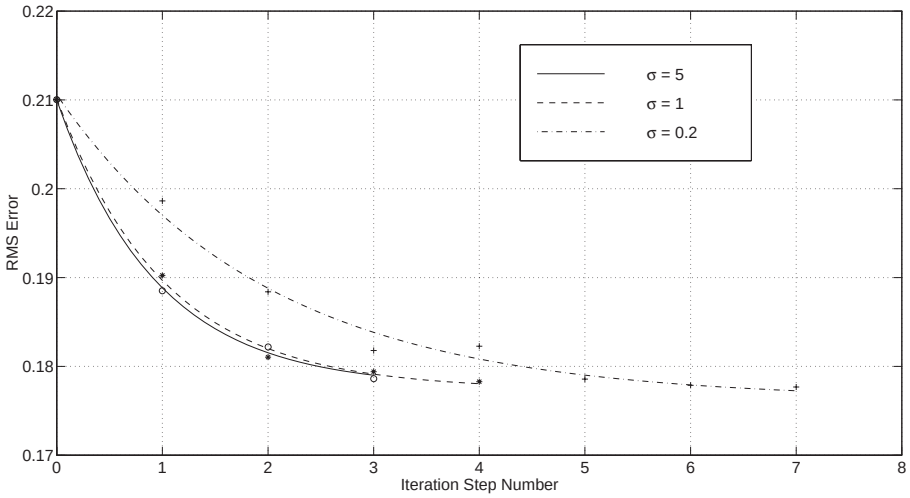


Figure 5: Convergence patterns of the reconstruction error, $E_{rms}^{(l)}$, with iteration step l, for different σ. Initial weighting factor is $\lambda^{(0)} = 0.2$. Note that after few iterations $E_{rms}^{(0)}$ reaches a minimum, and σ acts as an accelerating factor for convergence.

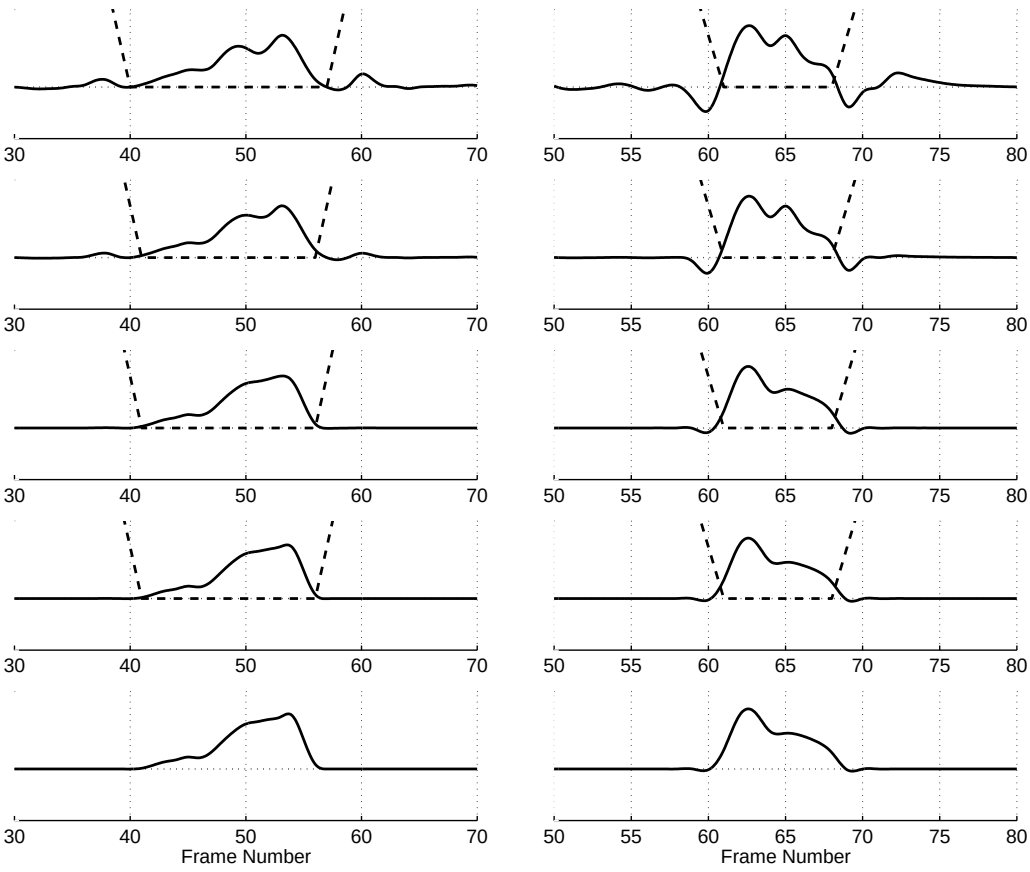
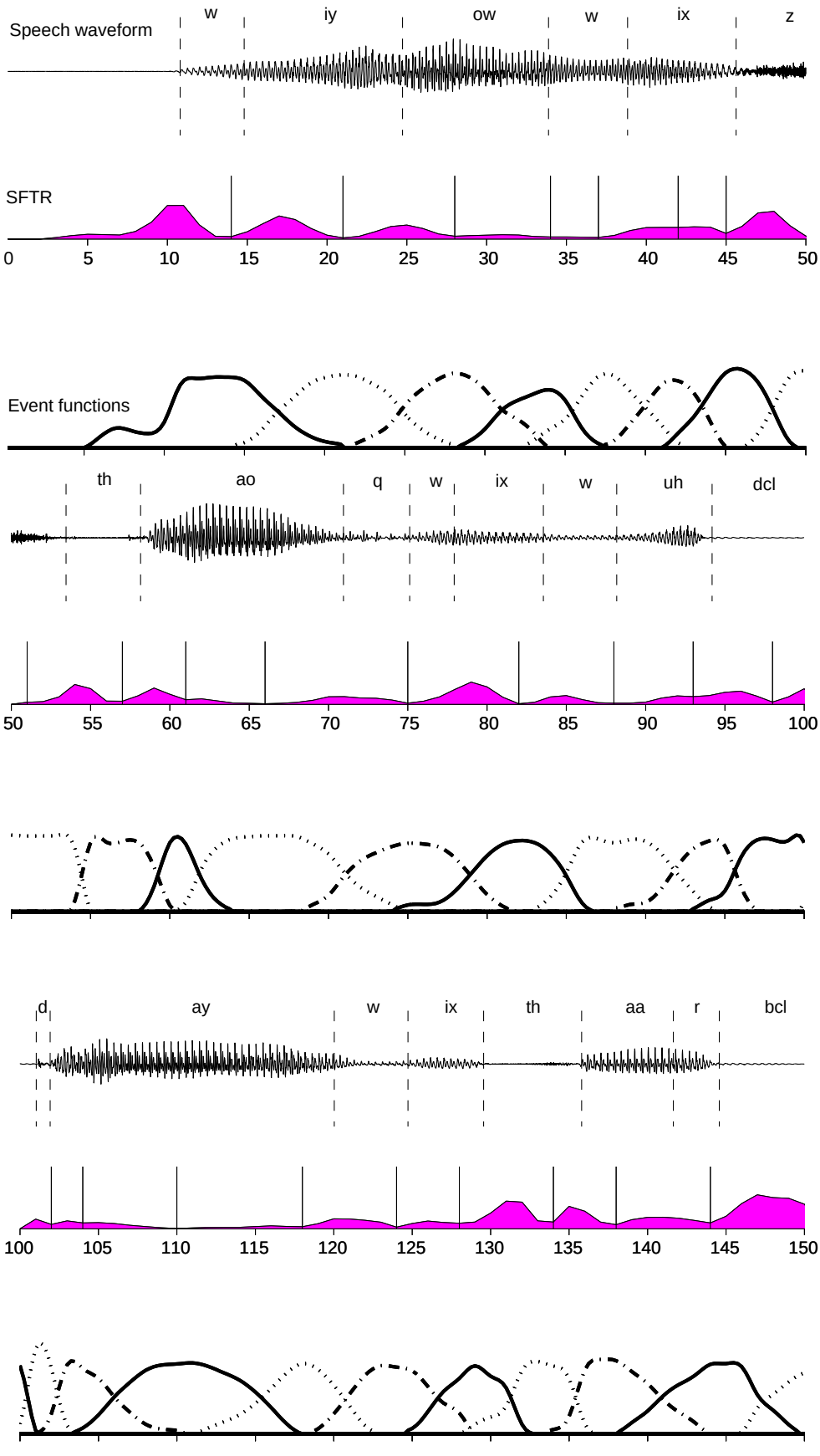


Figure 6: Effect of iterative refinement on event function shapes for some k (Top: initial event functions, $\phi_k(n)^{(0)}$'s, Bottom: final event functions, $\phi_k(n)^{(S)}$'s). Weighting functions, $w_k(n)^{(l)}$'s, are also shown for reference. Note the minor lobe smoothing and major lobe reshaping property which finally results in well-shaped and non-negative event functions.


Speech waveform
w
iy
ow
w
ix
z
SFTR
0
5
10
15
20
25
30
35
40
45
50
Event functions
th
ao
q
w
ix
w
uh
dcl
50
55
60
65
70
75
80
85
90
95
100
d
ay
w
ix
th
aa
r
bcl
100
105
110
115
120
125
130
135
140
145
150

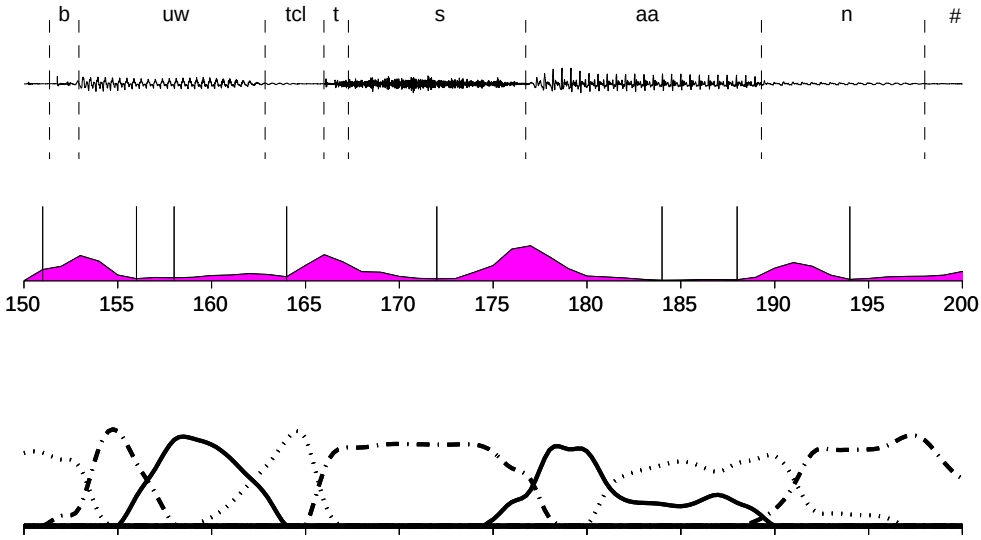


Figure 7: Plot of SFTR and the final event functions for the utterance "*we always thought we would die with our boots on*". S^{2}BEL-TD analysis has been performed on the utterance on a segmental basis. The speech waveform is also shown together with the phonetic transcription for reference. Broken lines in the speech plot show the phoneme boundaries, while the solid lines in the SFTR plot show the spectrally stable frame locations, i.e. local minima of SFTR.

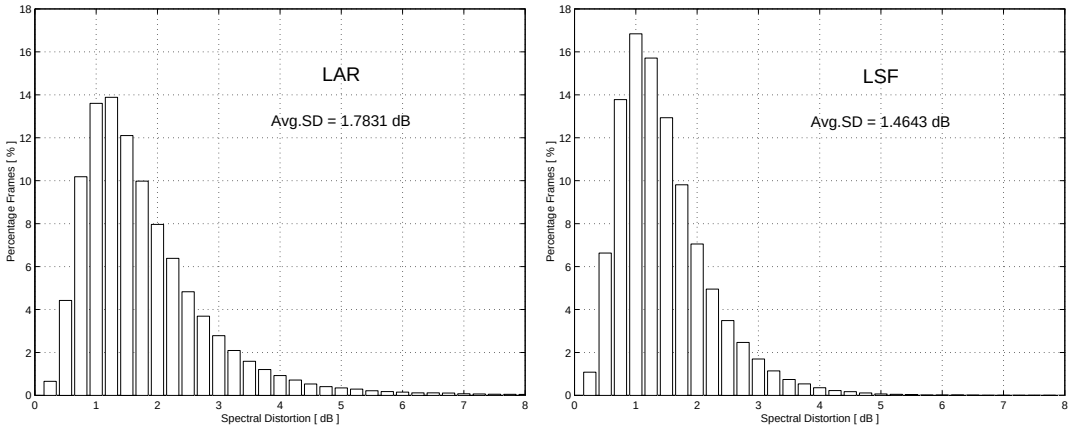


Figure 8: Distribution of the Spectral Distortion (SD) between original and reconstructed spectral parameters in the form of histograms. Left: SD histogram for the LAR parameters, Right: SD histogram for the LSF parameters. Speech data set consists of 250 sentence utterances spoken by 10 speakers (5 male & 5 female) of the ATR Japanese speech database. LSF's show slightly better reconstruction accuracy than LAR's.

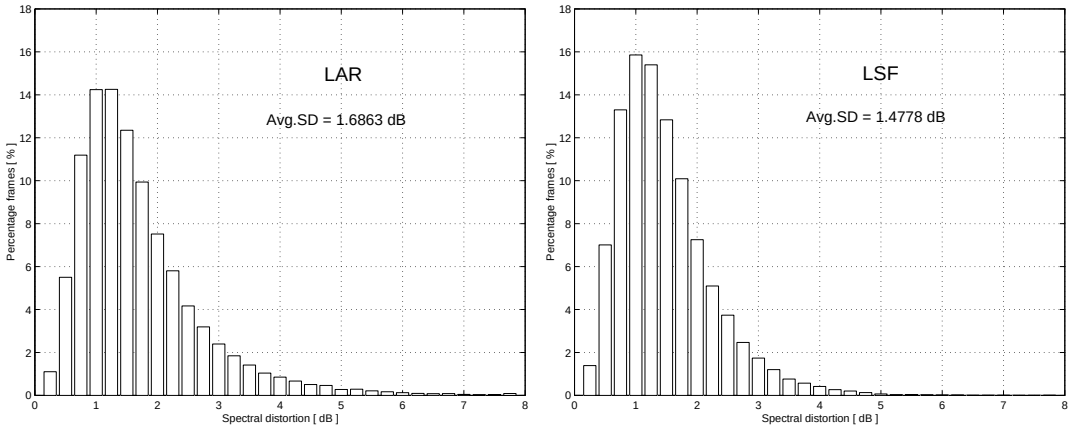


Figure 9: Distribution of the Spectral Distortion (SD) between original and reconstructed spectral parameters in the form of histograms. Left: SD histogram for the LAR parameters, Right: SD histogram for the LSF parameters. Speech data set consists of 192 sentence utterances spoken by 24 speakers (2 male & 1 female from each of 8 dialect regions) of the TIMIT English speech database. LSF's also show slightly better reconstruction accuracy than LAR's.

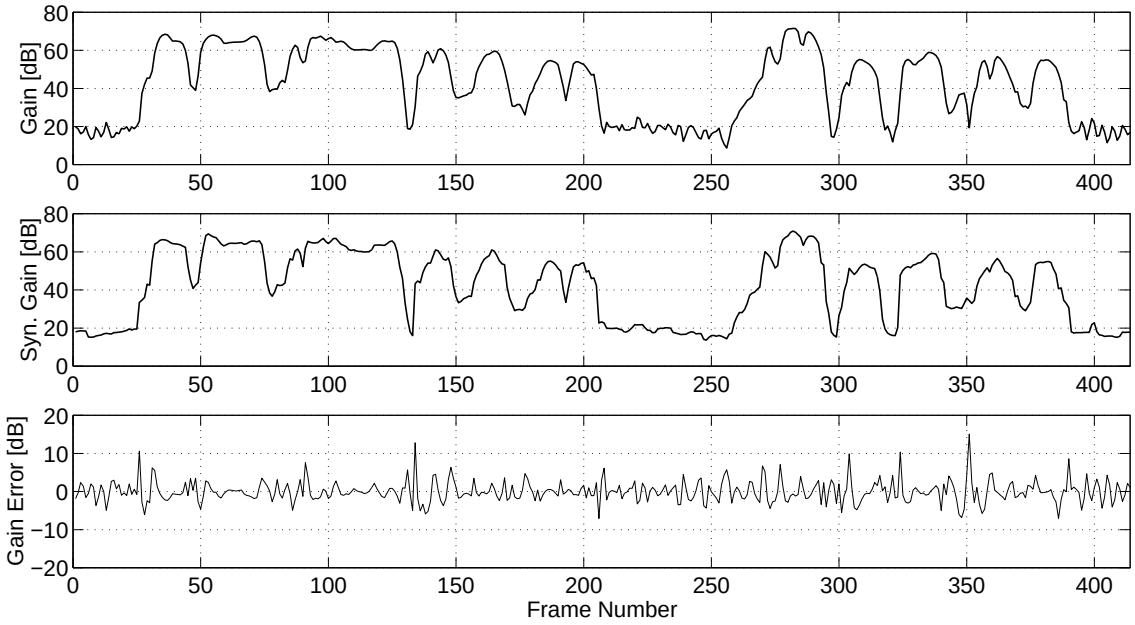


Figure 10: Original gain parameters, $g(n)$, reconstructed gain parameters, $\hat{g}(n)$, and frame-wise gain error, $e_g(n) = \hat{g}(n) - g(n)$, for the utterance "*kantan na shiryou wo ookuri shimasu node, shibaraku omachi kudasai* ", of the ATR Japanese speech database. The RMS gain error is 4.051 dB.

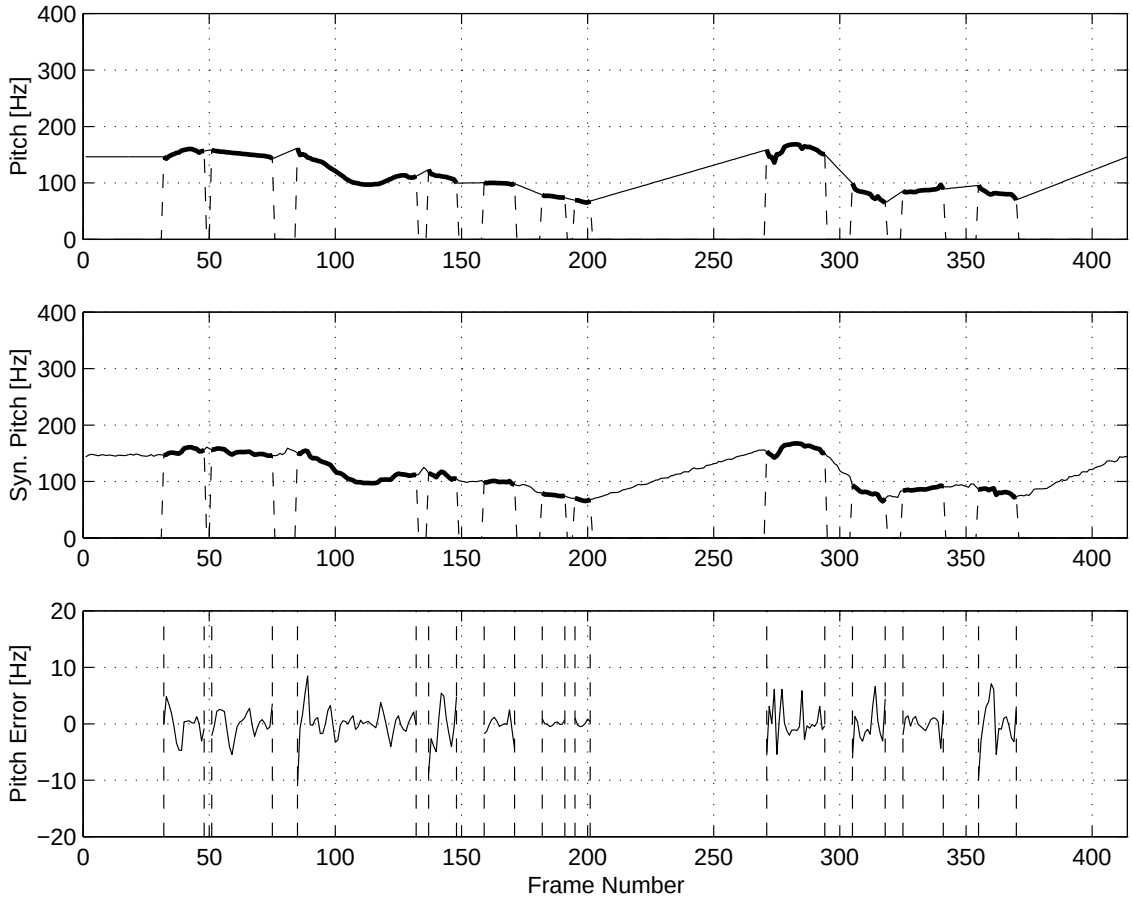


Figure 11: Original pitch parameters, $p(n)$, reconstructed pitch parameters, $\hat{p}(n)$, and frame-wise pitch error, $e_p(n) = \hat{p}(n) - p(n)$, for the utterance "*kantan na shiryou wo ookuri shimasu node, shibaraku omachi kudasai* ", of the ATR Japanese speech database. Pitch error is shown only for the voiced segments of the utterance. The RMS pitch error is 2.2984 Hz.