SPECTRAL STABILITY BASED EVENT LOCALIZING TEMPORAL DECOMPOSITION

A.C.R. Nandasena, Masato Akagi

Graduate School of Information Science,
Japan Advanced Institute of Science and Technology, Japan
nanda@jaist.ac.jp, akagi@jaist.ac.jp

ABSTRACT

In this paper a new approach to temporal decomposition (TD) of speech, called *"Spectral Stability Based Event Localizing Temporal Decomposition"*, abbreviated S^{2}BEL-TD, is presented. The original method of TD proposed by Atal is known to have the drawbacks of high computational cost, and the instability of the number and locations of events [1]. In S^{2}BEL-TD, the event localization is performed based on a maximum spectral stability criterion. This overcomes the instability problem of events of the Atal's method. Also, S^{2}BEL-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 of TD. Simulation results show that an average spectral distortion of about 1.5 dB can be achieved with LSF as the spectral parameter. Also, we have shown that the temporal pattern of the speech excitation parameters can also be well described using the S^{2}BEL-TD technique.

1. INTRODUCTION

Temporal decomposition was first proposed as a method for efficient coding of LPC parameters [1]. TD involves the decomposition of spectral parameters into a sequence of overlapping event functions and an associated sequence of event targets, as given in Eq. (1).

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

where, $\mathbf{a}_k$ and $\phi_k(n)$ are the k^{th} event target, i.e. spectral target, and the k^{th} event function, respectively. $\hat{\mathbf{y}}(n)$ is the approximation of $\mathbf{y}(n)$, the n^{th} spectral parameter vector, produced by the TD model. n represents the discrete time index. Eq. (1) can be written in matrix form as;

$$\hat{\mathbf{Y}} = \mathbf{A}\mathbf{\Phi} \qquad \hat{\mathbf{Y}} \in R^{P\times N}, \mathbf{A} \in R^{P\times K}, \mathbf{\Phi} \in R^{K\times N} \tag{2}$$

where P, N and K are the order of the spectral parameters, the number of frames in the speech segment and the number of event functions, respectively.

Although the original implementation of temporal decomposition of speech by Atal was mathematically solid, it is known to have the following two major drawbacks [1]. (i) The method is computationally costly, making it impractical. (ii) Instability of the number and locations of the events. In other words, they are very sensitive to some trivial changes in analysis parameters, i.e analysis window size etc. The high computational time of the Atal's method has been mainly attributed to the use of the computationally involved singular value decomposition (SVD), and the repeated evaluation of the event functions at small time intervals before screening out the redundant event functions using a reduction algorithm. Marcus & Lieshout investigated the possible validity of TD as a method of determining phonetically plausible events in speech, but came out with the instability problem of the original method with respect to the number and locations of the event functions [2]. Dijk-Kapper & Marcus improved the TD method to make events more stable, but the computational time has more or less remained the same because the time consuming SVD was still involved [3].

We intend 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, which avoids the necessity of redundant evaluation of event functions.

2. S^{2}BEL-TD OF SPEECH

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

[STEP 1] Determination of the *event targets.*
(First approximation)

$$\mathbf{A}^{(0)} = \left[\ \mathbf{a}_k^{(0)}\ \right]_{1\le k\le K} \tag{3}$$

[STEP 2] Determination of the *event functions.*
(First approximation)

$$\mathbf{\Phi}^{(0)} = \left[\ \phi_k(n)^{(0)}\ \right]_{1\le k\le K,\ 1\le n\le N} \tag{4}$$

[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 2.1, 2.2, and 2.3, respectively.

2.1. Determination of Event Targets

The transition rate of the i^{th} 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 Eq. (5). The squared sum of these transition rates of individual spectral parameters, $y_i(n)$, where $1 \le i \le P$, is defined as the Spectral Feature Transition Rate (SFTR) at the time point n, and is given by Eq. (6).

$$c_i(n) = \frac{\sum_{m=-M}^{M} m y_i(n+m)}{\sum_{m=-M}^{M} m^2}, \qquad 1 \le i \le P \tag{5}$$

$$\text{SFTR}: \quad s(n) = \sum_{i=1}^{P} c_i(n)^2, \qquad 1 \le n \le N \tag{6}$$

The local minima of $s(n)$ indicate the frames with maximum local spectral stability in speech, and we take these points as the 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;

$$\begin{aligned} \mathbf{A}^{(0)} &= \left[\begin{array}{cccc} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_K \end{array} \right] \\ &= \left[\begin{array}{cccc} \mathbf{y}(n_1) & \mathbf{y}(n_2) & \cdots & \mathbf{y}(n_K) \end{array} \right] \end{aligned} \tag{7}$$

2.2. 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. We achieve this by using a weighting function, $w_k(n)$, corresponding to each event function, $\phi_k(n)$.

Weighting function $w_k(n)$ for the k^{th} event function, $\phi_k(n)$, is defined as follows.

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

$$\mathbf{w}_k = \left[\begin{array}{cccc} w_k(1) & w_k(2) & \cdots & w_k(N) \end{array} \right] \tag{8}$$

$$\mathbf{W} = \left(\begin{array}{c} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \vdots \\ \mathbf{w}_K \end{array} \right) \in R^{K \times N} \tag{9}$$

Here, the intention is to allow a free evolution of an event function in the region between adjacent event locations, and to limit its behavior outside that region. By considering the columns of the matrix $\mathbf{W}$, diagonal matrices are formed as;

$$\mathbf{W}_n = diag \left[\begin{array}{cccc} w_1(n) & w_2(n) \cdots & w_K(n) \end{array} \right] \in R^{K \times K} \tag{10}$$

The functional $J\left(\vec{\phi}(n), \lambda\right)$ 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 Eq. (11).

$$J\left(\vec{\phi}(n), \lambda\right) = \sum_{i=1}^{P} \left(y_i(n) - \hat{y}_i(n)\right)^2 + \lambda \sum_{k=1}^{K} w_k(n)^2 \phi_k(n)^2 \tag{11}$$

where λ is a constant weighting factor and,

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

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

Minimization of the functional $J\left(\vec{\phi}(n), \lambda\right)$ with respect to $\vec{\phi}(n)$ results in the Eq. (12), using which the initial approximation of the event function matrix, $\mathbf{\Phi}^{(0)}$, could be formulated as given in Eq. (13).

$$\vec{\phi}(n) = \left(\mathbf{A}^T\mathbf{A} + \lambda {\mathbf{W}_n}^T \mathbf{W}_n\right)^{-1} \mathbf{A}^T \mathbf{y}_n \tag{12}$$

$$\text{where,} \quad 1 \le n \le N$$

$$\mathbf{\Phi}^{(0)} = \left(\begin{array}{cccc} \vec{\phi}(1) & \vec{\phi}(2) & \cdots & \vec{\phi}(N) \end{array} \right) \tag{13}$$

2.3. Iterative Refinement Procedure

We adopt an iterative procedure to improve the shapes of the event functions and the TD model accuracy, and to refine the event targets. The initial event functions show undesirable minor-lobes, i.e. negative projections, apart from the desirable major-lobes as shown in Fig. 1. The iterative refinement procedure effectively smooth-outs the minor-lobes while allowing the major-lobes to evolve freely. It also improves the TD model accuracy and refines the event targets. This involves the recursive performance of the procedures described in the Sections 2.3.1 and 2.3.2. Generally, 4 to 5 iterations are required to shape up the event functions.

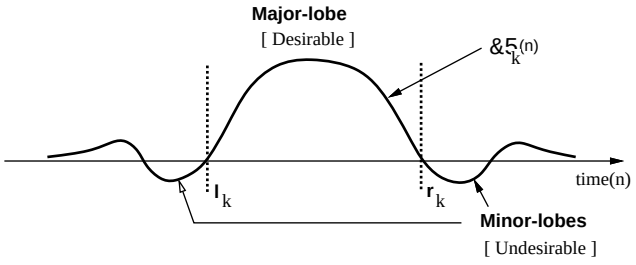

Figure 1. Typical shape of the initial event functions

2.3.1. Refinement of Event Functions

Recalculate the event functions using the procedure of Section 2.2, but use an adaptive weighting function and the quantitative balancing of the two error-terms of the functional $J\left(\vec{\phi}(n)^{(l)}, \lambda^{(l)}\right)$, as described below.

$$(\mathbf{A}^{(l-1)}, \mathbf{\Phi}^{(l-1)}) \rightarrow \mathbf{\Phi}^{(l)}, \qquad 1 \le l \le S \tag{14}$$

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

• *Adaptive Weighting function*

We define an adaptive weighting function as follows. It is adoptive to the major lobe limits of the event functions.

$$w_k^{(l)}(n) = \begin{cases} l_k^{(l-1)} - n, & \text{if } 1 \le n < l_k^{(l-1)} \\ 0, & \text{if } l_k^{(l-1)} \le n \le r_k^{(l-1)} \\ n - r_k^{(l-1)}, & \text{if } r_k^{(l-1)} < n \le N \end{cases} \quad (15)$$

Where, $l_k^{(l-1)}$ and $r_k^{(l-1)}$ are the left and right limits of the major lobe of the event function $\phi_k(n)^{(l-1)}$. The intention here is to restrict the minor-lobes while allowing the major-lobe to evolve freely. Therefore, this gives rise to major-lobe expansion, contraction or shift with a simultaneous minor-lobe reduction, when the iterations are performed.

• *Quantitative Balancing of the functional* $J\left(\vec{\phi}(n), \lambda\right)$

Select the weighting factor $\lambda^{(l)}$ at the iteration step l so as to balance the two error terms of the functional $J\left(\vec{\phi}^{(l)}(n), \lambda^{(l)}\right)$ using the results obtained at the iteration step $(l-1)$, i.e. $\mathbf{\Phi}^{(l-1)}$ and $\mathbf{A}^{(l-1)}$, as given below.

$$\lambda^{(l)} = \sigma \times \left(\frac{\sum_{n=1}^{N} \sum_{i=1}^{P} \left(y_i(n) - \hat{y}_i^{(l-1)}(n)\right)^2}{\sum_{n=1}^{N} \sum_{k=1}^{K} w_k^{(l)}(n)^2 \phi_k^{(l-1)}(n)^2} \right) \quad (16)$$

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.

2.3.2. Refinement of Event Targets

Recalculate the spectral targets by minimizing the squared error between the original and the reconstructed spectral parameters, with respect to the target vectors as follows.

$$\mathbf{\Phi}^{(l)} \rightarrow \mathbf{A}^{(l)}, \qquad 1 \le l \le S \quad (17)$$

$$E_i = \sum_{n=1}^{N} \left(y_i(n) - \sum_{k=1}^{K} a_{ik}^{(l)} \phi_k^{(l)}(n) \right)^2, \qquad 1 \le i \le P \quad (18)$$

By setting the partial derivative of E_i with respect to a_{ir}, to zero we obtain;

$$\sum_{k=1}^{K} a_{ik}^{(l)} \sum_{n=1}^{N} \phi_k^{(l)}(n) \phi_r^{(l)}(n) = \sum_{n=1}^{N} y_i(n) \phi_r^{(l)}(n), \quad 1 \le r \le K \quad (19)$$

This 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.

3. SIMULATION RESULTS

The Female/Japanese utterance "shimekiri ha geNshu desu ka" of the ATR Japanese Speech Database, resampled at 8 kHz, was used as the speech data. 10^{th} order LSF parameters were calculated using a LPC analysis window of 30 ms at 10 ms frame intervals. The plot of SFTR and the final event functions for the above data is shown in Fig. 2. The window size for the SFTR calculation is $2M = 40$ ms. The event rate is about 20 events/sec. $\lambda^{(0)}$=0.005 and σ=1 were selected as appropriate values for the weighting factor and balancing ratio, respectively, based on simulation results.

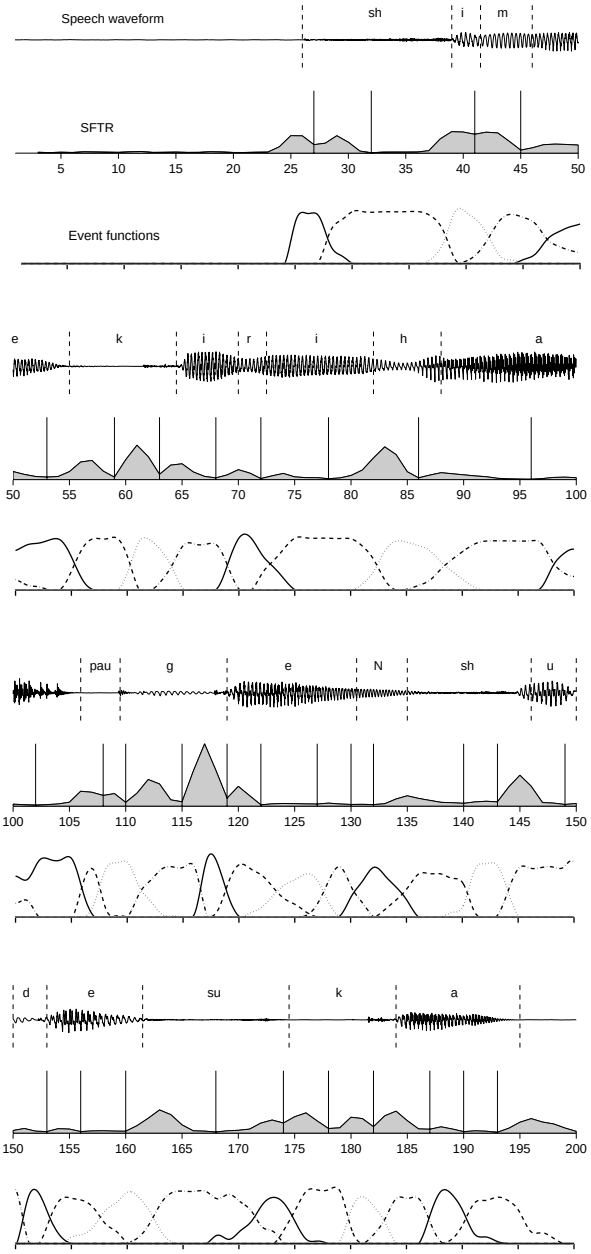

Figure 2. Plot of SFTR and the final event functions for the utterance "shimekiri ha geNshu desu ka". S^{2}BEL-TD analysis has been performed on the utterance on a segmental basis. The speech waveform is also shown together with the phonetic transcription for reference. Broken lines in the speech plot show the phoneme boundaries, while the solid lines in the SFTR plot show the spectrally stable frame locations, i.e. local minima of SFTR.

4. PERFORMANCE EVALUATION

An objective evaluation was performed based on the spectral distortion (SD) between the original spectral parameters, $\mathbf{y}(n)$, and the spectral parameters reconstructed from the TD model, $\hat{\mathbf{y}}(n)$. SD evaluation was performed on a set of five utterances in the ATR Japanese Speech Database. The average spectral distortion, and the percentage number of frames lying in the error bands of SD < 2 dB, 2 dB $\leq$ SD < 4 dB and SD ≥ 4 dB, were found to be 1.5 dB, 78%, 20% and 2%, respectively. The results signify a good approximation of the spectral parameters by the S^{2}BEL-TD model.

5. S^{2}BEL-TD OF SPEECH EXCITATION

We employ the S^{2}BEL-TD technique to describe the temporal characteristics of the speech excitation parameters, i.e gain, pitch and voicing. Here, the same event functions evaluated for the spectral parameters are used to describe the temporal evolution of the gain, pitch and voicing parameters also. We believe that the speech production is 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 a excitation parameter, i.e. gain, pitch or voicing. Then we approximate $b(n)$ by $\hat{b}(n)$, the reconstructed excitation parameter for the n^{th} frame, as follows in terms of excitation targets, b_k's, and the event functions, $\phi_k(n)$'s.

$$\hat{b}(n) = \sum_{k=1}^{K} b_k \phi_k(n), \qquad 1 \leq n \leq N \tag{20}$$

In Eq. (20), the event functions, $\phi_k(n)$'s, are known and therefore we determine the excitation targets, b_k's, by minimizing the squared error between the original excitation parameters and the reconstructed excitation parameters as follows.

$$E_b = \sum_{n=1}^{N} \left(b(n) - \sum_{k=1}^{K} b_k \phi_k(n) \right)^2 \tag{21}$$

$$\sum_{k=1}^{K} b_k \sum_{n=1}^{N} \phi_k(n)\phi_r(n) = \sum_{n=1}^{N} b(n)\phi_r(n), \qquad 1 \leq r \leq K \tag{22}$$

Eq. (22) 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 determine the reconstructed binary voicing parameters and binary voicing targets, from the non-binary results of Eq. (20) and Eq. (22), respectively. Fig. 3 shows an illustration of the gain contour approximation. Simulation results show that RMS gain-error, RMS pitch-error and percentage voicing error to be about 2.5 dB, 2.5 Hz and 4%, respectively. The low reconstruction error justifies a good approximation of the excitation parameters, by the S^{2}BEL-TD model.

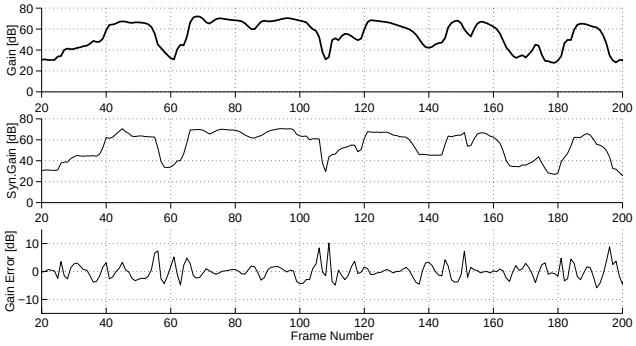

Figure 3. Original and reconstructed gain contours, and the gain-error for the same utterance as in Fig. 2. The RMS gain-error is 2.5 dB.

6. CONCLUSION

In this paper we have presented 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 robust eventually overcoming the instability problem of the Atal's method. The latter gives a greater degree of freedom to the event functions to evolve through iterations, compared to the more constrained quadratic weighting function of the Atal's method. Also, we believe that the S^{2}BEL algorithm which makes no use of SVD algorithm and the redundant calculation of event functions, is a significant improvement in terms of computational time 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, which would be useful in phonemic-level parametric manipulation of speech.

7. REFERENCES

[1] B.S. Atal, "Efficient coding of LPC parameters by temporal decomposition", *Proc. ICASSP '83*, pp. 81-84, 1983.

[2] S.M. Marcus & R.A.J.M. Van-Lieshout, "Temporal decomposition of speech", *IPO Annual Progress Report 19*, pp. 26-31, 1984.

[3] A.M.L. Van Dijk-Kappers & S.M. Marcus, "Temporal decomposition of speech", *Speech Communications*, Vol. 8, No. 2, pp. 125-135, 1989.

[4] A.C.R. Nandasena, "A new approach to temporal decomposition of speech and its application to low-bit-rate speech coding", *Master's Thesis*, Japan Advanced Institute of Science & Technology, 1997.